Bayesian Score Calibration for Approximate Models

Joshua J. Bon[*] JOSHUA.BON@ADELAIDE.EDU.AU
School of Mathematical Sciences,
Adelaide University

David J. Warne DAVID.WARNE@QUT.EDU.AU
School of Mathematical Sciences & Centre for Data Science,
Queensland University of Technology
ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems

David J. Nott STANDJ@NUS.EDU.SG
Department of Statistics and Data Science,
National University of Singapore

Christopher Drovandi C.DROVANDI@QUT.EDU.AU
School of Mathematical Sciences & Centre for Data Science,
Queensland University of Technology
ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems

Editor: Edo Airoldi

Abstract

Scientists continue to develop increasingly complex mechanistic models to reflect their knowledge more realistically. Statistical inference using these models can be challenging since the corresponding likelihood function is often intractable and model simulation may be computationally burdensome. Fortunately, in many of these situations it is possible to adopt a surrogate model or approximate likelihood function. It may be convenient to conduct Bayesian inference directly with a surrogate, but this can result in a posterior with poor uncertainty quantification. In this paper, we propose a new method for adjusting approximate posterior samples to reduce bias and improve posterior coverage properties. We do this by optimizing a transformation of the approximate posterior, the result of which maximizes a scoring rule. Our approach requires only a (fixed) small number of complex model simulations and is numerically stable. We develop supporting theory for our method and demonstrate beneficial corrections to approximate posteriors across several examples of increasing complexity.

Keywords: likelihood-free inference, simulation-based inference, scoring rules, posterior correction, surrogate model

1. Introduction

Scientists and practitioners desire greater realism and complexity in their models, but this can complicate likelihood-based inference. If the proposed model is sufficiently complex, computation of the likelihood can be intractable. In this setting, if model simulation is feasible, then approximate Bayesian inference can proceed via likelihood-free methods (Sisson et al., 2018). However, most likelihood-free methods require a large number of model

[*] Thanks to Ming Xu, Aad van der Vaart, and anonymous referees for their helpful comments.

simulations. It is common for likelihood-free inference methods to require hundreds of thousands of model simulations or more. Thus, if model simulation is also computationally intensive, it is difficult to conduct inference via likelihood-free methods.

Often it is feasible to propose an approximate but more computationally tractable "surrogate" version of the model of interest. However, the resulting approximate Bayesian inferences can be biased (relative to the true posterior), and produce distributions with poor uncertainty quantification that do not have correct coverage properties (see for example, Xing et al., 2019; Warne et al., 2022a).

In this paper, we propose a novel Bayesian procedure for approximate inference. Related literature will be reviewed in Section 2.7. Our approach begins with an approximate model and transforms the resulting approximate posterior to reduce bias and more accurately quantify uncertainty. Only a small number (i.e., hundreds) of model simulations from the target model are required, whilst standard Bayesian inference is only conducted using the approximate model. Furthermore, these computations are trivial to parallelize, which can reduce the time cost further by an order of magnitude or more, depending on available computing resources. Importantly, our procedure does not require any evaluations of the likelihood function of the complex model of interest.

The approximate posterior can be formed on the basis of a surrogate model or approximate likelihood function. For example, a surrogate model may be a deterministic version of a complex stochastic model (e.g., Warne et al., 2022a) and an example of a surrogate likelihood is the Whittle likelihood for time series models (Whittle, 1953). Furthermore, our framework permits the application of approximate Bayesian inference algorithms on the surrogate model. For example, the Laplace approximation, variational approximations, and likelihood-free inference methods that require only a small number of model simulations (e.g., Gutmann and Corander, 2016). Hence, computational approximations and surrogate models can be used, and corrected, simultaneously.

Figure 1 displays a graphical overview of Bayesian score calibration, which we describe in detail throughout the paper. We develop a new theoretical framework to support our method that is also applicable to some related methods. In particular, Theorem 2 generalizes the underlying theory justifying recent simulation-based inference methods (e.g., Pacchiardi and Dutta, 2022). Moreover, we provide practical calibration diagnostics that assess the quality of the adjusted approximate posteriors and alert users to success or failure of the procedure.

The rest of this paper is organized as follows. In Section 2 we provide background, present our approximate model calibration method, and develop theoretical justifications. Our new method is demonstrated on some examples of increasing complexity in Section 3 with further examples deferred to the supplementary materials S.1–S.3. Section 4 contains a concluding discussion of limitations, as well as ongoing and future research directions.

2. Methods

Bayesian inference updates the prior distribution of unknown parameters $\theta \sim \Pi$ with information from observed data $y \sim P(\cdot \mid \theta)$ having some dependence on θ. We take Π as the prior probability measure with probability density (or mass) function π, and $P(\cdot \mid \theta)$ as the data-generating process with probability density (or mass) function $p(\cdot \mid \theta)$. We will use the

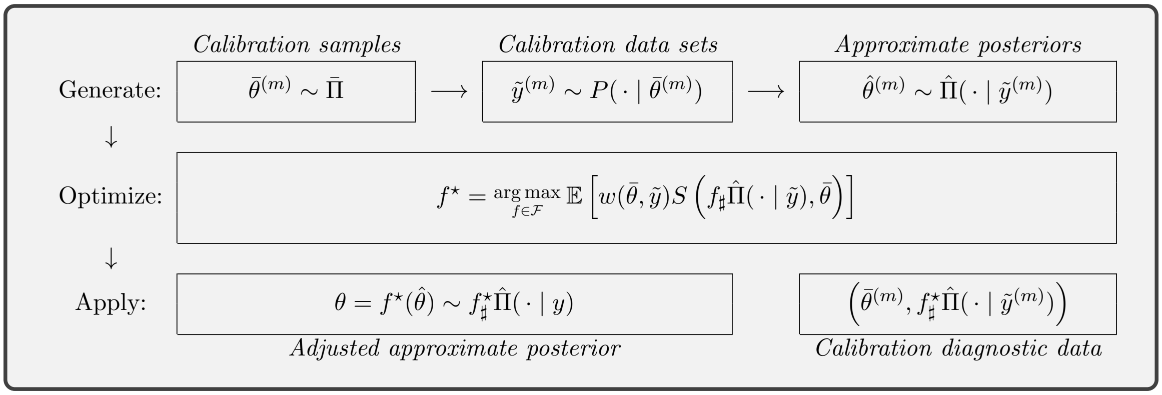

Calibration samples
Calibration data sets
Approximate posteriors
Generate:
$\bar{\theta}^{(m)} \sim \bar{\Pi}$
$\tilde{y}^{(m)} \sim P(\,\cdot \mid \bar{\theta}^{(m)})$
$\hat{\theta}^{(m)} \sim \hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})$
Optimize:
$f^\star = \underset{f \in \mathcal{F}}{\arg\max}\, \mathbb{E}\left[w(\bar{\theta}, \tilde{y}) S\left(f_\sharp \hat{\Pi}(\,\cdot \mid \tilde{y}), \bar{\theta}\right)\right]$
Apply:
$\theta = f^\star(\hat{\theta}) \sim f^\star_\sharp \hat{\Pi}(\,\cdot \mid y)$
$\left(\bar{\theta}^{(m)}, f^\star_\sharp \hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})\right)$
Adjusted approximate posterior
Calibration diagnostic data

Figure 1: Graphical overview of Bayesian score calibration. Firstly, the importance distribution $\bar{\Pi}$ and data-generating process $P(\,\cdot \mid \theta)$ simulate parameter-data pairs $(\bar{\theta}^{(m)}, \tilde{y}^{(m)})$ for $m \in \{1, \ldots, M\}$. Each simulated data set, $\tilde{y}^{(m)}$, defines a new approximate posterior, $\hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})$, which we approximate with Monte Carlo samples. Secondly, we use a strictly proper scoring rule S to find the best transformation of the approximate posterior, defining the pushforward distribution $f_\sharp \hat{\Pi}(\,\cdot \mid \tilde{y})$, with respect to true data-generating parameter $\bar{\theta}$, averaged over $\bar{\Pi}$, with weights $w(\bar{\theta}, \tilde{y})$. The optimization objective function is approximated using Monte Carlo with pre-computed samples from the generation step. Finally, the optimal function, $f^\star$, is used to generate samples from the adjusted approximate posterior with the observed data, y, and produce data for diagnostic summaries.

terms distribution and law of a random variable with reference to an underlying probability measure. The function $p(y \mid \cdot\,)$ is the likelihood for fixed data y and varying θ. The prior distribution Π is defined on measurable space (Θ, ϑ), whilst $P(\,\cdot \mid \theta)$ is defined on $(\mathsf{Y}, \mathcal{Y})$ for fixed $\theta \in \Theta$, where ϑ (resp. $\mathcal{Y}$) is a σ-algebra on Θ (resp. Y).

Bayes theorem determines that the posterior distribution, incorporating information from the data, has density (mass) function $\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)}$, where $p(y) = \int p(y \mid \theta)\Pi(\mathrm{d}\theta)$ for $\theta \in \Theta$. Here, for fixed $y \in \mathsf{Y}$, $Z = p(y)$ is the posterior normalizing constant. For varying $y \in \mathsf{Y}$, $p(y)$ is the density (mass) function of the marginal distribution of the data P, defined on $(\mathsf{Y}, \mathcal{Y})$.

Posterior inference typically requires approximate methods as the normalizing constant Z is unavailable in a closed form. There are two broad families of approximations in general use: (i) sampling methods, including Markov chain Monte Carlo (MCMC, Brooks et al., 2011) and sequential Monte Carlo (SMC, Chopin and Papaspiliopoulos, 2020); and (ii) optimization-based methods including variational inference or Laplace approximations (Bishop, 2006). All standard implementations of these methods rely on pointwise evaluation of the likelihood function, or some unbiased estimate.

When the likelihood is infeasible or computationally expensive to evaluate pointwise we may wish to choose a surrogate posterior which approximates the original in some sense. We consider a surrogate with distribution $\hat{\Pi}(\,\cdot \mid y)$ on (Θ, ϑ) and with density (mass) function $\hat{\pi}(\,\cdot \mid y)$. Such a surrogate can arise from an approximation to the original model, likelihood or posterior. In the case of a surrogate model or approximate likelihood, this is equivalent to $\hat{\pi}(\theta \mid y) \propto \hat{p}(y \mid \theta)\pi(\theta)$ for an approximate likelihood $\hat{p}(y \mid \theta)$.

We design methods to calibrate the approximate posterior when the true likelihood, $p(y \mid \theta)$, cannot be evaluated but we can sample from the data-generating process $P(\,\cdot \mid \theta)$. To facilitate this calibration we need a method to compare approximate distributions to the true posterior distribution we are interested in. To this end, the next section introduces scoring rules and expected scores.

Before proceeding, we define some general notation. We will continue to use a hat to denote objects related to approximate posteriors and use $\tilde{y}$ to denote simulated data. The expectation of f with respect to probability distribution Q is written as $\mathbb{E}_{\theta\sim Q}[f(\theta)]$ or $Q(f)$. The degenerate probability measure at x is denoted by δ_x. If Q_1 and Q_2 are measures where Q_1 is dominated by Q_2 we write $Q_1 \ll Q_2$. The Euclidean norm is denoted by $\|\cdot\|_2$.

2.1 Bayesian Score Calibration

Let $S : \mathcal{P} \times \Theta \to \mathbb{R} \cup \{-\infty, \infty\}$ be a scoring rule for the class of probability distributions $\mathcal{P}$ where $U, V \in \mathcal{P}$ are defined on the measurable space (Θ, ϑ). A scoring rule compares a (single) observation θ to the probabilistic prediction U by evaluating $S(U, \theta)$. The expected score under V is defined as $S(U, V) = \mathbb{E}_{\theta\sim V}[S(U, \theta)]$. A scoring rule S is strictly proper in $\mathcal{P}$ if $S(V, V) \geq S(U, V)$ for all $U, V \in \mathcal{P}$ and equality holds if and only if $U = V$. We refer to Gneiting and Raftery (2007) for a review of scoring rules in statistics. We use the expected score $S(U, V)$ to define a discrepancy between an adjusted approximate posterior U and the true posterior $V(\cdot) = \Pi(\,\cdot \mid y)$.

To motivate our use of scoring rules, consider the variational problem

$$\max_{U \in \mathcal{P}} \mathbb{E}_{\theta\sim\Pi(\cdot\mid y)}\left[S(U, \theta)\right], \tag{1}$$

with optimal distribution $U^\star$. If S is strictly proper and the class of distributions $\mathcal{P}$ is rich enough, i.e., $\Pi(\,\cdot \mid y) \in \mathcal{P}$ for fixed data y, then we recover the posterior as the optimal distribution uniquely, that is $U^\star(\cdot) = \Pi(\,\cdot \mid y)$. Unfortunately, the expectation in (1) is intractable due to its expression with $\Pi(\,\cdot \mid y)$. To circumvent this, we instead consider averaging the objective function over some distribution Q on $(\mathsf{Y}, \mathcal{Y})$, leading to the new optimization problem

$$\max_{K \in \mathcal{K}} \mathbb{E}_{\tilde{y} \sim Q} \mathbb{E}_{\theta \sim \Pi(\cdot \mid \tilde{y})} \left[S(K(\,\cdot \mid \tilde{y}), \theta) \right], \tag{2}$$

where $K(\,\cdot \mid \tilde{y})$ is now a kernel defined for $\tilde{y} \in \mathsf{Y}$, and $\mathcal{K}$ is a family of Markov kernels. If $\mathcal{K}$ is sufficiently rich the optimal kernel at y, $K^\star(\,\cdot \mid y)$, will be the posterior $\Pi(\,\cdot \mid y)$.

Definition 1 (Sufficiently rich kernel family) *Let $\mathcal{K}$ be a family of Markov kernels, $\mathcal{P}$ be a class of probability measures, Q be a probability measure on $(\mathsf{Y}, \mathcal{Y})$, and $\Pi(\,\cdot \mid \tilde{y})$ be the true posterior at $\tilde{y}$. We say $\mathcal{K}$ is sufficiently rich with respect to $(Q, \mathcal{P})$ if for all $K \in \mathcal{K}$, $K(\,\cdot \mid \tilde{y}) \in \mathcal{P}$ almost surely and there exists $K \in \mathcal{K}$ such that $K(\,\cdot \mid \tilde{y}) = \Pi(\,\cdot \mid \tilde{y})$ almost surely, where Q is the law of $\tilde{y}$.*

The maximization problem in (2) is significantly more difficult than that of (1) as it involves learning the form of a Markov kernel dependent on any data generated by Q, rather than a probability distribution (i.e., with data set fixed). However, unlike (1), the new problem (2) can be translated into a tractable optimization as described in Theorem 2.

Theorem 2 *Consider a strictly proper scoring rule S relative to the class of distributions $\mathcal{P}$, and importance distribution $\bar{\Pi}$ with Radon–Nikodym derivative $r = \mathrm{d}\Pi/\mathrm{d}\bar{\Pi}$ where Π is the prior. Let $v : \mathsf{Y} \to [0, \infty)$ and define Q by change of measure $Q(\mathrm{d}\tilde{y}) = P(\mathrm{d}\tilde{y}) v(\tilde{y}) / P(v)$ such that the normalising constant $P(v) \in (0, \infty)$, where P is the marginal distribution of the data. Assume the true posterior $\Pi(\,\cdot \mid \tilde{y}) \in \mathcal{P}$ almost surely, where Q is the law of $\tilde{y}$. Let the optimal Markov kernel $K^\star$ be*

$$K^\star \equiv \arg\max_{K \in \mathcal{K}} \mathbb{E}_{\theta \sim \bar{\Pi}} \mathbb{E}_{\tilde{y} \sim P(\cdot \mid \theta)} \left[w(\theta, \tilde{y}) S(K(\,\cdot \mid \tilde{y}), \theta) \right], \quad w(\theta, \tilde{y}) = r(\theta) v(\tilde{y}). \tag{3}$$

If the family of kernels $\mathcal{K}$ is sufficiently rich with respect to $(Q, \mathcal{P})$ then $K^\star(\,\cdot \mid \tilde{y}) = \Pi(\,\cdot \mid \tilde{y})$ almost surely.

A proof for Theorem 2 is provided in Appendix B.1.

Remark 3 *We can interpret the optimal $K^\star$ in Theorem 2 as recovering the true posterior for $\tilde{y} \sim Q$, that is $K^\star(\,\cdot \mid \tilde{y}) = \Pi(\,\cdot \mid \tilde{y})$ for any $\tilde{y}$ in the support of P (smallest set with probability one) such that $v(\tilde{y}) > 0$.*

Remark 4 *A special case of Theorem 2 is stated by Pacchiardi and Dutta (2022) where $Q = P$, the marginal distribution of the data. Considering $Q \neq P$ is crucial for establishing our subsequent results. Further, Lueckmann et al. (2017) consider the case where the scoring rule is the log-probability score and $v(\tilde{y}) = k(\tilde{y}, y)$, a kernel measuring the discrepancy between the simulated and observed data.*

Remark 5 *The objective function in (3) can also be seen as an amortized variational optimization problem. However, we take the perspective of targeting a fixed data set. We expect the function v to be very useful in this setting, but do not explore this further here.*

Theorem 2 changes the order of expectation by noting the joint distribution of $(\theta, \tilde{y})$ can be represented by the marginal distribution of $\tilde{y}$ and conditional distribution of θ given $\tilde{y}$ or vice versa. It also uses an importance distribution, $\bar{\Pi}$, instead of the prior, Π. As simulators for the importance distribution (or prior) and data-generating process are assumed to be available, it is possible to estimate the objective function of (3) using Monte Carlo approximations.

The weighting function w is an importance sampling correction but also includes an additional component, v. The function v describes a change of measure for the marginal distribution P and represents the flexibility in Q such that the optimization problems in (2) and (3) remain equivalent. Overall then, Theorem 2 tells us we have the freedom to choose the importance distribution $\bar{\Pi}$ and change of measure v, in principle, without affecting the optimization.

In practice it may be difficult to verify the conditions of Theorem 2 that depend on the choice of scoring rule, kernel family, importance distribution, and function v. We discuss these concerns, our choices, and practical consequences in the context of intractable posteriors in Sections 2.3–2.5. We also provide a diagnostic tool to monitor violations of these conditions and failures in the Monte Carlo approximation of the objective function in Section 2.6.

2.2 Bayesian Score Calibration Algorithm

The pseudo-code for Bayesian score calibration is detailed in Algorithm 1, where we use a Monte Carlo approximation of the optimization objective in (5) and the subsequent sections describe implementation details and justifications. We assume that the vector of parameters θ has support on $\mathbb{R}^d$. Should only a subset of parameters in θ need correcting, Steps 5 and 6 can proceed using only this subset. Steps 5 and 6 can also be performed element-wise if correcting the joint distribution is unnecessary. When $\theta \in \Theta \subset \mathbb{R}^d$ (a strict subset) we use an invertible transformation to map θ to $\mathbb{R}^d$ in our examples in Section 3. In this case, some care needs to be taken to ensure the weights are calculated correctly in Step 3. After performing the adjustment we can transform back to the original space.

2.3 The Energy Score

Thus far, we have considered a generic strictly proper scoring rule, S, as the propriety of our method does not rely on a specific scoring rule. For the remainder of the paper we will focus on the so-called energy score (Section 4.3, Gneiting and Raftery, 2007) defined as

$$S(U, \theta) = \frac{1}{2}\mathbb{E}_{u,u' \sim U}\|u - u'\|_2^{\beta} - \mathbb{E}_{u \sim U}\|u - \theta\|_2^{\beta},$$

for distribution U on (Θ, ϑ), $\theta \in \Theta$, and fixed $\beta \in (0, 2)$. Note that u and u' are independent realizations from U. We find that $\beta = 1$ gives good empirical performance and note that the energy score at this value is a multivariate generalization of the continuous ranked probability score (Gneiting and Raftery, 2007). The energy score is a strictly proper scoring rule for the class of Borel probability measures on $\Theta = \mathbb{R}^d$ where $\mathbb{E}_{u \sim U}\|u\|_2^{\beta}$ is finite. Hence, using $\beta = 1$ assumes the posterior has finite mean.

Algorithm 1 Bayesian score calibration using approximate models

Inputs: Number of calibration data sets M, number of Monte Carlo samples N, importance distribution $\bar{\Pi}$, approximate posterior model $\hat{\Pi}$, scoring rule S, transformation function family $\mathcal{F}$, observed data set y, clipping level $\alpha \in [0,1]$. Optional: stabilizing function v (otherwise unit valued).

Outputs: Estimated optimal transformation function f^* and samples from the adjusted approximate posterior based on observed data set y.

```
1. For $m \in \{1, \ldots, M\}$
    (i) Generate calibration samples, $\bar{\theta}^{(m)} \sim \bar{\Pi}$.
    (ii) Generate calibration data sets, $\tilde{y}^{(m)} \sim P(\,\cdot \mid \bar{\theta}^{(m)})$.
    (iii) Calculate weights, $w^{(m)} = \dfrac{\pi(\bar{\theta}^{(m)})}{\bar{\pi}(\bar{\theta}^{(m)})} v(\tilde{y}^{(m)})$ if $\alpha < 1$, else $w^{(m)} = 1$ if $\alpha = 1$.
    (iv) Calculate (or sample from) the approximate posteriors $\hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})$.
         If sampling, then $\hat{\theta}_i^{(m)} \sim \hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})$ for $i \in \{1, \ldots, N\}$.
2. Clip weights, $w_{\text{clip}}^{(m)} = \min\{w^{(m)}, q_{1-\alpha}\}$ for $m \in \{1, \ldots, M\}$, where $q_{1-\alpha}$ is the
   $100(1-\alpha)\%$ empirical quantile of the weights.
3. Solve the optimization, $f^\star = \arg\max_{f \in \mathcal{F}} \sum_{m=1}^{M} w_{\text{clip}}^{(m)} S(f_\sharp \hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)}), \bar{\theta}^{(m)})$.
   If using the energy score, use $\{\hat{\theta}_i^{(m)}\}_{i=1}^{N}$ from Step (iv) to estimate $S$ as per (4).
4. Generate approximate adjusted samples from pushforward $f_\sharp^\star \hat{\Pi}(\,\cdot \mid y)$ by calculating
   $f^\star(\theta)$ where $\theta \sim \hat{\Pi}(\,\cdot \mid y)$ for the desired number of samples.
```

The energy score is appealing as it can be approximated using Monte Carlo methods. Assuming we can generate N samples from U, as in our case, then it is possible to construct an approximation to this scoring rule as

$$\hat{S}(U, \theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2} \left\| u_i - u_{k_i} \right\|_2^{\beta} - \left\| u_i - \theta \right\|_2^{\beta} \right), \tag{4}$$

where $u_i \sim U$ for $i \in \{1, \ldots, N\}$ and k is a random variable, uniformly distributed over permutation vectors of length N.

If samples from U can be generated exactly, then the approximation will be consistent if $\mathbb{E}_{u \sim U} \|u\|_2^{\beta} < \infty$, while inexact Monte Carlo samples from U (generated by SMC or MCMC for example) will lead to a consistent estimator under the same condition.

2.4 Approximate Posterior Transformations

Bayesian score calibration requires a family of kernels to optimize over, $\mathcal{K}$. We consider the family defined by conditional deterministic transformations of random variables drawn from approximate posterior distributions, that is, a pushforward measure[1] of the approximate posterior conditional on the data. This choice of kernel family allows efficient inference in our current context of expensive or intractable likelihoods, when only Monte Carlo samples from the approximate posterior are available.

We use a class of pushforward kernels, analogous to pushforward measures but with additional parameters. Consider a probability measure ν on (Θ, ϑ). We write the pushforward measure of g on ν as $g_{\sharp}\nu$ when g is a measurable function on (Θ, ϑ). Now consider the Markov kernel M from $(\mathsf{Y}, \mathcal{Y})$ to (Θ, ϑ) and function $f : \mathsf{Y} \times \Theta \to \Theta$ such that $f_y(\cdot) = f(y, \cdot)$ is a measurable function on (Θ, ϑ) for each $y \in \mathsf{Y}$. We define the pushforward kernel $f_{\sharp}M$ by stating the pushforward kernel emits (i) a conditional probability measure $f_{\sharp}M(\,\cdot \mid y) = M(f_y^{-1}(\cdot) \mid y)$ for fixed $y \in \mathsf{Y}$ and (ii) a function $f_{\sharp}M(B \mid \cdot\,) = M(f_y^{-1}(B) \mid \cdot\,)$ for fixed $B \in \vartheta$. The dependence of the function f on y allows information from the data to inform the transformation.

The family of kernels we consider can be described as $\mathcal{K} = \{f_{\sharp}\hat{\Pi} : f \in \mathcal{F}\}$ where $\mathcal{F}$ is some family of functions. Under such a family of kernels, we now express our idealized optimization problem as

$$f^{\star} = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{\theta \sim \bar{\Pi}} \mathbb{E}_{\tilde{y} \sim P(\cdot|\theta)} \left[w(\theta, \tilde{y}) S(f_{\sharp}\hat{\Pi}(\,\cdot \mid \tilde{y}), \theta) \right]. \tag{5}$$

One appeal of this family of pushforward (approximate) posteriors is that we only need to sample approximate draws from $\hat{\Pi}(\,\cdot \mid \tilde{y})$ once for each $\tilde{y}$, after which samples from $f_{\sharp}\hat{\Pi}(\,\cdot \mid \tilde{y})$ can be generated by applying the deterministic transformation to the set of approximate draws.

If the approximate model is computationally inexpensive to fit, then generating samples with the pushforward will also be inexpensive. This cost is also predetermined, since we fix the number of calibration data sets and the approximate posteriors only need to be learned (or sampled from) once. Moreover, once the transformation $f^{\star}$ is found, samples from the

[1] Also known as "transformations of measures" (see for example, Billingsley, 1995, p. 185).

adjusted approximate posterior can be generated by drawing from the approximate model with observed data y and applying $f^\star$.

2.4.1 Kernel Richness from Approximate Posterior Transformations

To recover the true posterior, Theorem 2 requires that the family of kernels is sufficiently rich. A transformation family $\mathcal{F}$ will be sufficiently rich if there exists $f \in \mathcal{F}$ such that $f_\sharp \hat{\Pi}(\cdot \mid \tilde{y}) = \Pi(\cdot \mid \tilde{y})$ almost surely for $\tilde{y} \sim Q$. Therefore, the richness will depend on the class $\mathcal{F}$ and the approximate posterior $\hat{\Pi}(\cdot \mid \tilde{y})$. In applications it may be difficult to specify practical families that meet this criterion. Specifically, in our context of expensive model simulators we judge a practical family as one that is parametric with relatively few parameters—thus requiring fewer calibration data sets to be simulated (and samples from each approximate posterior).

We can also consider the richness of certain approximate posterior transformation families asymptotically. Consider realizations of the target posterior $\theta_{\tilde{y},n} \sim \Pi(\cdot \mid \tilde{y}_{1:n})$ and approximate posterior $\hat{\theta}_{\tilde{y},n} \sim \hat{\Pi}(\cdot \mid \tilde{y}_{1:n})$ for some $\tilde{y}_{1:n} \sim P(\cdot \mid \bar{\theta})$, such that

$$\sqrt{n}(\theta_{\tilde{y},n} - \mu_{\bar{\theta}}) \to \mathrm{N}(0, \Sigma_{\bar{\theta}}), \quad \sqrt{n}(\hat{\theta}_{\tilde{y},n} - \hat{\mu}_{\bar{\theta}}) \to \mathrm{N}(0, \hat{\Sigma}_{\bar{\theta}}),$$

as $n \to \infty$ in distribution for some $\mu_{\bar{\theta}}, \hat{\mu}_{\bar{\theta}} \in \mathbb{R}^d$ and fixed $\bar{\theta} \in \mathbb{R}^d$. Choosing the transformation $f_{\bar{\theta}}(\theta) = L_{\bar{\theta}}[\theta - \hat{\mu}_{\bar{\theta}}] + \hat{\mu}_{\bar{\theta}} + b_{\bar{\theta}}$ ensures that

$$\sqrt{n}(f_{\bar{\theta}}(\hat{\theta}_{\tilde{y},n}) - \hat{\mu}_{\bar{\theta}} - b_{\bar{\theta}}) \to \mathrm{N}(0, L_{\bar{\theta}} \hat{\Sigma}_{\bar{\theta}} L_{\bar{\theta}}^\top),$$

for some $b_{\bar{\theta}} \in \mathbb{R}^d$ and $L_{\bar{\theta}} \in \mathbb{R}^{d \times d}$. To recover the true posterior asymptotically with our method (by ensuring sufficient richness) we require $b^\star_{\bar{\theta}} = \mu_{\bar{\theta}} - \hat{\mu}_{\bar{\theta}}$ and $L^\star_{\bar{\theta}} = \Sigma_{\bar{\theta}}^{1/2} \hat{\Sigma}_{\bar{\theta}}^{-1/2}$ with $\bar{\theta}$ varying. As such, $\mathcal{F}$ must contain the function $f_{\bar{\theta}}(\theta) = L^\star_{\bar{\theta}}(\theta - \hat{\mu}_{\bar{\theta}}) + \hat{\mu}_{\bar{\theta}} + b^\star_{\bar{\theta}}$ where $\hat{\mu}_{\bar{\theta}}$, $b^\star_{\bar{\theta}}$, and $L^\star_{\bar{\theta}}$ vary in $\bar{\theta}$. Practically speaking, $\hat{\mu}_{\bar{\theta}}$ can be estimated from the approximate posterior and $\bar{\theta}$ can be estimated from the simulated data $\tilde{y}$. Hence both are conditional on $\tilde{y}$, yielding the approximate posterior mean $\hat{\mu}_{\tilde{y}}$ and estimator $\theta^*_{\tilde{y}}$ respectively. In this case, the class

$$\mathcal{F} = \{f : f_{\tilde{y}}(\theta) = L(\theta^*_{\tilde{y}})[\theta - \hat{\mu}_{\tilde{y}}] + \hat{\mu}_{\tilde{y}} + b(\theta^*_{\tilde{y}}), b \in \mathcal{B}, L \in \mathcal{L}\}, \tag{6}$$

would define an (asymptotically) sufficiently rich family of transformations if $\hat{\mu}_{\tilde{y}_{1:n}} \to \hat{\mu}_{\bar{\theta}}$ and $\theta^*_{\tilde{y}_{1:n}} \to \bar{\theta}$ as $n \to \infty$, and $\bar{\theta} \mapsto b^\star_{\bar{\theta}} \in \mathcal{B}$ and $\bar{\theta} \mapsto L^\star_{\bar{\theta}} \in \mathcal{L}$. Thus simplifying an (asymptotically) sufficiently rich class to affine functions in θ, where L and b are only functions of a consistent estimator of $\bar{\theta}$.

2.4.2 Relative Moment-Correcting Transformation

For this paper, we choose to use a simple transformation that corrects the location and scale of the approximate posterior under each simulated data set by the same relative amount. Assuming each approximate posterior only needs a fixed location-scale correction is a strong assumption, slightly stronger than the form described in the previous section, but empirically we find this to be a pragmatic and effective choice. Moreover, we show how to monitor and detect violations of this assumption in Section 2.6 and justify its use asymptotically. Using more flexible transformation families would allow for the correction of poorer approximate posterior distributions, but also require more draws from the (potentially expensive) data-generating process to estimate the transformation. We leave the exploration of more flexible transformation families for future work.

We are motivated to consider moment-matching transformations (see Warne et al., 2022a; Lei and Bickel, 2011; Sun et al., 2016, for example) due to the class of asymptotically sufficiently richness of kernels derived in (6). However, instead of matching moments between two random variables, we correct multiple random variables by the same relative amounts. Let the mean and covariance of a particular approximate posterior, $\hat{\Pi}(\,\cdot \mid \tilde{y})$, be $\hat{\mu}(\tilde{y})$ and $\hat{\Sigma}(\tilde{y})$ for some data set $\tilde{y}$ and $\Theta \subseteq \mathbb{R}^d$. We denote the relative change in location and covariance by $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ respectively, where A is the decomposition of a positive definite matrix B, that is $AA^\top = B$. The transformation is applied to realizations from the approximate posterior, $\theta \sim \hat{\Pi}(\,\cdot \mid \tilde{y})$, as

$$f(\tilde{y}, \theta) = A[\theta - \hat{\mu}(\tilde{y})] + \hat{\mu}(\tilde{y}) + b, \tag{7}$$

for $\theta \in \Theta$. The mean and covariance of the adjusted approximate posterior, $f_\sharp\hat{\Pi}(\,\cdot \mid \tilde{y})$, is $\hat{\mu}_f(\tilde{y}) = \hat{\mu}(\tilde{y}) + b$ and $\hat{\Sigma}_f(\tilde{y}) = A\hat{\Sigma}(\tilde{y})A^\top$, respectively. We consider the Eigen decomposition of B to parameterize A. In particular, we take $A = VD^{1/2}$ such that $VV^\top = I$ and D is a strictly positive diagonal matrix.

Unlike the family in (6), the relative moment-correcting transformations we consider restrict b and A to be equal for all data sets $\tilde{y}$. To justify this, we make the following observation. If we use a function v such that $v(\tilde{y}_{1:n}) \to \delta_{y_{1:n}}(\tilde{y}_{1:n})$ as $n \to \infty$, then our optimization (3) will simplify to the original problem (1) asymptotically. In this regime, b and A will transform the approximate posterior of a single data set, $y_{1:n}$, correcting its mean and variance. Therefore, we can use v to focus our calibration on the data at hand with the aim of making the optimization less sensitive to the insufficiency of $\mathcal{K}$, and asymptotically sufficiently rich. Our default choice of weights, which we discuss in Section 2.5, imply such a property for v.

2.5 Choice of Weighting Function

There are two components of the weighting function w that can be chosen to refine our estimate of the optimal approximate posterior through (3). The first is the importance distribution $\bar{\Pi}$ which we use to concentrate the samples of θ around likely values of the posterior distribution conditional on the observed data y. The second choice is the change of measure function v, or *stabilizing function*, which we use to stabilize the weighting function w after choosing $\bar{\Pi}$. Using an importance distribution $\bar{\Pi}$ has been considered previously (for example in Lueckmann et al., 2017; Pacchiardi and Dutta, 2022) but with the inclusion of a generic v in Theorem 2, the importance weights can be stabilized in various ways. We discuss idealized and practical weight functions further in Appendix D.

We explore weight truncation, or clipping (Ionides, 2008), to ensure finite variance of the weights. In general, clipping the empirical weights is achieved by

$$w^{(m)}_{\text{clip}} = \min\{w^{(m)}, q_{1-\alpha}\}, \quad m \in \{1, \dots, M\}, \tag{8}$$

where the truncation value $q_{1-\alpha}$ is the $100(1-\alpha)\%$ empirical quantile based on weights $\{w^{(m)}\}_{m=1}^M$. Letting $\alpha \in [0, 1]$ depend on M such that $\alpha \to 0$ as $M \to \infty$ is sufficient for asymptotic consistency. However, full clipping i.e., $\alpha = 1$ or equivalently *unit weights*, will not satisfy this. Instead, we establish asymptotic consistency (in the size of the data set) for unit weights next, and demonstrate good empirical performance in Section 3.

2.5.1 Unit Weighting Function

In this section we will consider the effect of approximating the weight function $w(\theta, \tilde{y})$ with unit weights, i.e., $\hat{w} = 1$. This can also be viewed as clipping with $\alpha = 1$ in (8). We consider weights of the form $w(\theta, \tilde{y}_{1:n}) = v(\tilde{y}_{1:n})\pi(\theta)/\bar{\pi}(\theta)$ where the number of observations $n \to \infty$ and $r(\theta) = \pi(\theta)/\bar{\pi}(\theta)$. The results of this section are possible due to the stability function, v, which is free to be chosen without affecting the validity of the method, as established by Theorem 2. We first consider the consistency of the unit weights, when a consistent estimator θ^* exists.

Theorem 6 *Let $g(x) = \bar{\pi}(x)/\pi(x)$ for $x \in \Theta$. If there exists an estimator $\tilde{\theta}^*_n \equiv \theta^*(\tilde{y}_{1:n})$ such that $\tilde{\theta}^*_n \overset{a.s.}{\to} z$ as $n \to \infty$ when $\tilde{y}_i \overset{iid}{\sim} P(\cdot \mid z)$ for $z \in \Theta$, and g is positive and continuous at z then the error when using $\hat{w} = 1$ satisfies*

$$\hat{w} - w(\theta, \tilde{y}_{1:n}) \overset{a.s.}{\to} 0,$$

as $n \to \infty$ with choice of stabilizing function $v(\tilde{y}_{1:n}) = g(\tilde{\theta}^*_n)$.

Theorem 6 establishes a strong consistency result; unit weights are a large sample approximation to the theoretically correct weights for some choice of stabilizing function v. A corresponding weak consistency result follows analogously. Crucially, a consistent estimator is not required to implement our method in practice. The existence of such an estimator simply ensures that unit weights are a valid choice asymptotically. From a practical standpoint, unit weights effectively remove the importance sampling component of (3), ensuring our approach is numerically stable, and practically appealing. A proof of Theorem 6 is provided in Appendix B.2, whilst a central limit theorem for the unit weight approximation appears in Appendix B.3.

2.5.2 Importance Distribution

The importance distribution $\bar{\Pi}$ can be chosen to focus the calibration on regions of Θ. For this paper we use the approximate posterior $\hat{\Pi}(\cdot \mid y)$ to choose $\bar{\Pi}$. Specifically, we use the scale transformation

$$\bar{\Pi}(\mathrm{d}\theta) \propto \hat{\pi}(D^{-1}(\theta - \hat{\mu}) + \hat{\mu} \mid y)\mathrm{d}\theta, \tag{9}$$

where $\hat{\pi}(\cdot \mid y)$ is the density of the approximate posterior $\hat{\Pi}(\cdot \mid y)$, D is a positive-definite diagonal matrix used to inflate the variance of the approximate posterior, and $\hat{\mu}$ is the estimated mean of the approximate posterior. We can draw samples from $\bar{\Pi}$ using the transformation $D(\theta' - \hat{\mu}) + \hat{\mu}$ when $\theta' \sim \hat{\Pi}(\cdot \mid y)$.

With unit weights and importance distribution (9), the stabilizing function from Theorem 6 will have the form

$$v(\tilde{y}_{1:n}) = \hat{p}(y_{1:n} \mid D^{-1}(\tilde{\theta}^*_n - \hat{\mu}) + \hat{\mu})c(\tilde{\theta}^*_n),$$
$$\text{where } c(\tilde{\theta}^*_n) \propto \frac{\pi(D^{-1}(\tilde{\theta}^*_n - \hat{\mu}) + \hat{\mu})}{\pi(\tilde{\theta}^*_n)},$$

which will behave as $v(\tilde{y}_{1:n}) \to \delta_{\hat{\theta}_0}(\tilde{\theta}_n^*)$ for $n \to \infty$ where $\hat{\theta}_0$ is the maximum likelihood estimator (MLE) from $\hat{p}(y_{1:n} \mid \cdot\,)$ as $n \to \infty$. Asymptotically, this would reduce the support of the importance distribution Q to the manifold $\mathsf{M}_n = \{\tilde{y}_{1:n} \in \mathsf{Y}^n : \tilde{\theta}_n^* = \hat{\theta}_0\}$ and will depend on the sufficient statistics for the true posterior. On the manifold M_n, each $\tilde{y}_{1:n}$ will have the same limiting distribution (if it exists) as the consistent estimator $\tilde{\theta}_n^*$ for each $\tilde{y}_{1:n}$ is constrained to be equal. This justifies our use of constant b and A to define our approximate posterior transformation family (when using unit weights) as there is only one true and one approximate posterior to correct for in this regime. Hence, transformations defined by (7) are asymptotically sufficiently rich under some mild conditions when using unit weights.

2.6 Calibration Diagnostic

To assist using Bayesian score calibration in practice, we suggest a performance diagnostic to warn users if the adjusted approximate posterior is unsuitable for inference. The diagnostic detects when the learned transformation does not adequately correct the approximate posteriors from the calibration data sets. The diagnostic can be computed with trivial expense as it requires no additional simulations.

To elaborate, whilst executing Algorithm 1, we have access to the true data-generating parameter value $\bar{\theta}^{(m)}$ for each adjusted approximate posterior, $f^\star_\sharp\hat{\Pi}(\,\cdot \mid \tilde{y}^{(m)})$. Therefore, we can measure how these adjusted approximate posteriors perform (on average) relative to this true value. Various metrics could be used for this task, but we find the empirical coverage probabilities for varying nominal levels of coverage to be suitable.

Specifically, if $\mathrm{Cr}(U, \rho)$ is a $(100 \times \rho)\%$ credible interval (or highest probability region) for distribution U then we calculate the achieved coverage (AC) by estimating

$$\mathrm{AC}(\rho) = \mathsf{P}\left[\bar{\theta} \in \mathrm{Cr}(f^\star_\sharp\hat{\Pi}(\,\cdot \mid \tilde{y}), \rho)\right], \quad \text{where } \tilde{y} \sim P(\,\cdot \mid \bar{\theta}). \tag{10}$$

for a sequence of $\rho \in (0, 1)$, using pairs of $(\bar{\theta}^{(m)}, \{f^\star(\hat{\theta}_i^{(m)})\}_{i=1}^N)$ for $m \in \{1, \dots, M\}$ generated by Algorithm 1. We refer to ρ as the target coverage. Whilst this diagnostic can be calculated on the joint distribution of the posterior, for simplicity we will use the marginal version of (10) resulting in a diagnostic for each parameter in the posterior. We forgo a multivariate diagnostic as the marginal version requires less user input (only the type of credible intervals to use) and multivariate versions are far more difficult to compute. For our experiments we use credible intervals with end-points determined by symmetric tail-probabilities and plot the miscoverage for an interval of target coverage levels $\rho \in [0.1, 0.95]$. The miscoverage is $\mathrm{MC}(\rho) = \mathrm{AC}(\rho) - \rho$, where positive values indicate over-coverage, while negative values indicate under-coverage.

The miscoverage diagnostic will be sensitive to failures of the method in the high probability regions of the importance distribution, $\bar{\Pi}$. If one wishes to test areas outside this region or in specific areas, new pairs of transformed approximate posteriors and their data-generating value could be produced at the cost of additional computation. With respect to the importance distribution, this diagnostic will help to detect if the quality of the approximate distribution is insufficient, if the transformation family is too limited, if the weights have too high variance or if the optimization procedure otherwise fails (for example due to an insufficient number of calibration data sets).

2.7 Related Research

Lee et al. (2019) and Xing et al. (2019) develop similar calibration procedures to ours but to estimate the true coverage of approximate credible sets as a diagnostic tool or means to adjust their posterior. We adjust the approximate posterior samples directly based on their distribution rather than just correcting coverage. Relatedly, Menéndez et al. (2014) correct confidence intervals from approximate inference for bias and nominal coverage in the frequentist sense, and Rodrigues et al. (2018) calibrate the entire approximate posterior based on similar arguments.

Xing et al. (2020) develop a method to transform the marginal distributions of an approximate posterior without expensive likelihood evaluation. They estimate a distortion map, which, theoretically, transports the approximate posterior to the exact posterior (marginally). Since the true distortion map is unavailable, Xing et al. (2020) learn the distortion map using simulated data sets, their associated approximate posteriors, and the true value of the parameter used to generate the data set (similar to our approach). They fit a beta regression model to the training data, which consists of approximate CDF values as the response and the data sets (or summary statistics thereof) as the features. Xing et al. (2020) learn the parameters of the beta distribution using neural networks. Their approach ensures that the approximate posterior transformed with the estimated map reduces the Kullback–Leibler divergence to the true posterior. However, in their examples, Xing et al. (2020) use $\mathcal{O}(10^6)$ simulations from the model of interest, in order to have a sufficiently large sample to train the neural network. Another reason for the large number of model simulations is they only retain a small proportion of simulated data sets from the prior predictive distribution that are closest to the observed data, in an effort to obtain a more accurate neural network localized around the observed data. Our method only requires generating $\mathcal{O}(10^2)$ data sets from the target model, and thus may be more suited to models where it is moderately or highly computationally costly to simulate. Further, in their examples, Xing et al. (2020) require fitting $\mathcal{O}(10^4)$ to $\mathcal{O}(10^5)$ approximate posteriors, whereas we only require $\mathcal{O}(10^2)$. Thus our approach has a substantially reduced computational cost.

Rodrigues et al. (2018) develop a calibration method based on the coverage property that was previously used in Prangle et al. (2014) as a diagnostic tool for approximate Bayesian computation (ABC, Beaumont et al., 2002; Sisson et al., 2018). Even though a key focus of Rodrigues et al. (2018) is to adjust ABC approximations, the method can be used to recalibrate inferences from an approximate model. Like Xing et al. (2020), Rodrigues et al. (2018) require a much larger number of model simulations and approximate posterior calculations compared to our approach. Furthermore, Rodrigues et al. (2018) require evaluating the CDF of posterior approximations at parameter values used to simulate from the target model. Thus, if the surrogate model is not sufficiently accurate, the CDF may be numerically 0 or 1, and the corresponding recalibrated sample will not be finite. The method of Xing et al. (2020) may also suffer from similar numerical issues. Our approach using the energy score is numerically stable.

Vandeskog et al. (2024) develop a post-processing method for posterior samples to correct composite and otherwise misspecified likelihoods that have been used for computational convenience. Their method uses a linear transformation to correct the asymptotic variance of the model at the estimated mode (or suitable point estimate). They show that their method can greatly improve the low coverage resulting from the initial misspecification. Their adjustment requires an analytical form for the true likelihood with first and second order derivatives. Our method does not require an analytical form for the true likelihood, nor calculable derivatives. Moreover, we derive an adjustment in the finite-sample regime.

A related area is delayed acceptance MCMC (e.g., Sherlock et al., 2017) or SMC (e.g., Bon et al., 2021). In delayed acceptance methods, a proposal parameter is first screened through a Metropolis-Hastings (MH) step that depends only on the likelihood for the surrogate model. If the proposal passes this step, it progresses to the next MH stage that depends on the likelihood of the expensive model, otherwise the proposal can be rejected quickly without probing the expensive likelihood. Although exact Bayesian inference can be generated with delayed acceptance methods, they require a substantial number of expensive likelihood computations, which limits the speed-ups that can be achieved. Our approach, although approximate, does not require any expensive likelihood calculations, and thus is more suited to complex models with highly computational expensive or completely intractable likelihoods. The idea of delayed acceptance is generalized with multifidelity methods in which a continuation probability function is optimized based on the receiver operating characteristic curve with the approximate model treated as a classifier for the expensive model (Prescott and Baker, 2020; Prescott et al., 2024; Warne et al., 2022b).

In other related work, Warne et al. (2022a) consider two approaches, preconditioning and moment-matching methods, that exploit approximate models in an SMC setting for ABC. The preconditioning approach applies a two-stage mutation and importance resampling step that uses an approximate model to construct a more efficient proposal distribution that reduces the number of expensive stochastic simulations required. The moment-matching approach transforms particles from an approximate SMC sampler to increase particle numbers and statistical efficiency of an SMC sampler using the expensive model. Of these two methods, the moment-matching SMC approach is demonstrated to be particularly effective in practice. As a result, we use the moment-matching transformation to inform the moment-correcting transformation used in this work.

As for frequentist-based uncertainty quantification, Warne et al. (2024) develop a method for valid frequentist coverage for intractable likelihoods with generalized likelihood profiles, whilst Müller (2013) obtain valid frequentist properties in misspecified models based on sandwich covariance matrix adjustments. Frazier et al. (2023) use a similar adjustment to correct for a misspecified Bayesian synthetic likelihood approach to likelihood-free inference. A deliberate misspecification of the covariance matrix in the synthetic likelihood is used to speed-up computation. Then a post-processing step compensates for the misspecification in the approximate posterior. Their approach does not have the goal of approximating the posterior distribution for the correctly specified model, and the post-processing step performs only a covariance adjustment without any adjustment of the mean.

Related work has considered learning the conditional density of the posterior as a neural network (Papamakarios and Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019). In these methods, the conditional density estimate is updated sequentially using samples from the current approximate posterior. Sequential neural posterior estimation, as it is called, is built upon by Papamakarios et al. (2019) with a focus on learning a neural approximation to the likelihood, rather than the full posterior. We expect the theoretical framework we propose to be useful in these contexts for developing methods to stabilize the importance weights, or understanding existing attempts at this (e.g., Deistler et al., 2022).

Pacchiardi and Dutta (2022) explore similar concepts and relate these to generative adversarial networks. Our work has related theoretical foundations, in that we use expectations over the marginal probability of the data to circumvent the intractability of the posterior, though we have a more general formulation. Moreover, we also focus on the case where our learned posterior is a correction of an approximate posterior we have access to samples from. As in Pacchiardi and Dutta (2022) we are also concerned with scoring rules, in particular, the energy score. We refer to an extended review of machine learning approaches for likelihood-free inference surmised by Cranmer et al. (2020) for further reading.

3. Examples

This section contains a number of empirical examples demonstrating Bayesian score calibration. A worked example to elucidate the elements of our framework is provided in Appendix A for a Bayesian logistic regression. We defer a further three empirical examples to the supplementary materials S.1–S.3 for brevity.

In the following examples we use $\beta = 1$ for the tuning parameter of the energy score, $M = 100$ or $M = 200$ as the number of calibration data sets and use the approximate posterior with scale inflated by a factor of 2 as the importance distribution, unless otherwise stated. The examples in Section 3.1 are tractable and inexpensive to run so that we can assess the performance of the calibration method against the true posterior on repeated independent data sets. The examples in Sections 3.2 and 3.3 have intractable likelihoods, so we do not perform an exact inference comparison but do compute coverage comparisons (since the truth is known) to investigate the results. A `julia` (Bezanson et al., 2017) package implementing our methods and reproducing our examples is available online, see Appendix C for details.

For examples where we validate the performance of our model calibration procedure using new independent data sets we compare the adjusted (approximate) posterior with the original approximate posteriors, and true posteriors using the average (over the M independent data sets); mean square error (MSE), bias of the posterior mean, posterior standard deviation, and the coverage rate of the nominal 90% credible intervals. For some marginal posterior samples $\{\theta_j\}_{j=1}^{J}$, we approximate the MSE using $\widehat{\text{MSE}} = \frac{1}{J}\sum_{j=1}^{J}(\theta_j - \theta)^2$, where θ is the known scalar parameter value (i.e., a particular component of the full parameter vector that we are adjusting).

3.1 Ornstein–Uhlenbeck Process

We first consider the Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930). The OU process, $\{X_t\}_{t\geq 0}$ for $X_t \in \mathbb{R}$, is a mean-reverting stochastic process that is governed by the Itô stochastic differential equation (SDE)

$$\mathrm{d}X_t = \gamma(\mu - X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{11}$$

with mean μ and volatility of the process denoted by σ. The rate at which X_t reverts to the mean is γ, whilst W_t is a standard Wiener process.

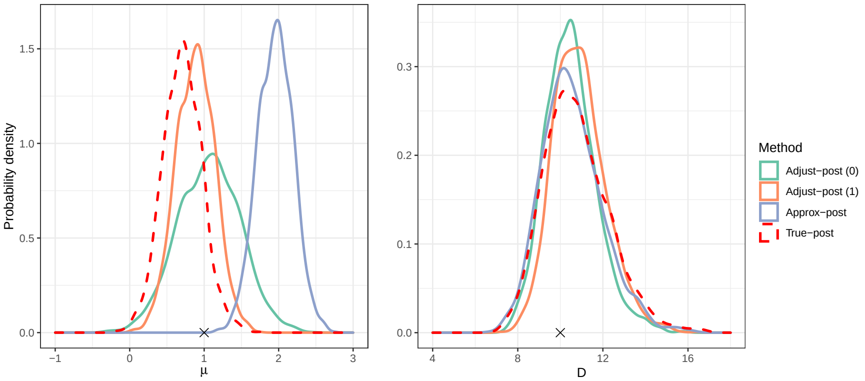


Figure 2: Univariate densities estimates of approximations to the OU Process model posterior distribution from a single simulation. The original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping are shown with solid lines. The true posterior (True-post) is shown with a dashed line. The true generating parameter value is indicated with a cross (×).

Given an initial condition, $X_0 = x_0$ at $t = 0$, we can obtain the distribution of the state at future time T through the solution to the forward Kolmogorov equation (FKE) for (11). For the OU process, the FKE is tractable with the solution

$$X_T \sim \mathrm{N}\left(\mu + (x_0 - \mu)e^{-\gamma T}, \frac{\sigma^2}{2\gamma}(1 - e^{-2\gamma T})\right). \tag{12}$$

However, we note that analytical results are not available for most SDE models and one must rely on numerical methods such as Euler-Maruyama schemes (Maruyama, 1955).

The next sections illustrate our method on two OU processes, the processes are one- and two-dimensional respectively. In the first example we approximate the likelihood using the limiting distribution of the OU process, whilst in the second we approximate the posterior distribution using variational inference. The first example uses independent transformations for each parameter and the second uses a multivariate transformation.

3.1.1 Univariate OU Process with Limiting Distribution Approximation

Take X_T as defined in (12) as the true model (and corresponding likelihood) in this example. For the observed data we take 100 independent realizations simulated from the above model with $x_0 = 10$, $\mu = 1$, $\gamma = 2$, $T = 1$ and $\sigma^2 = 20$. We assume that x_0 and γ are known and we attempt to infer μ and $D = \sigma^2/2$. We use independent priors where $\mu \sim \mathrm{N}(0, 10^2)$ and $D \sim \mathrm{Exp}(1/10)$ (parameterized by the rate). We sample and perform our adjustment over the space of $\log D$, but report results in the original space of D. The limiting distribution $T \to \infty$ is the approximate model, $X_\infty \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{2\gamma}\right)$, from which we define the approximate likelihood. Clearly there will be a bias in the estimation of μ.

For this example we sample from the approximate and true posteriors using the `Turing.jl` library (Ge et al., 2018) in `Julia` (Bezanson et al., 2017). We use the default No-U-Turn Hamiltonian Monte Carlo algorithm (Hoffman and Gelman, 2014). For simplicity, we set the stabilizing function $v(\tilde{y}) = 1$.

	μ				D			
Method	MSE	Bias	SD	AC[1]	MSE	Bias	SD	AC[1]
Approx-post	1.54	1.21	0.22	0	4.73	0.18	1.46	85
Adjust-post (0)	0.12	0.15	0.20	64	4.83	0.28	1.24	72
Adjust-post (0.5)	0.12	0.15	0.23	81	5.08	0.41	1.42	81
Adjust-post (1)	0.12	0.15	0.23	82	5.13	0.42	1.45	83
True-post	0.12	-0.01	0.26	94	5.00	0.37	1.48	85

Table 1: Average results for each parameter over 100 independent data sets for the univariate OU process example. The posteriors compared are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping, and the true posterior (True-post). [1]Achieved coverage with 90% target.

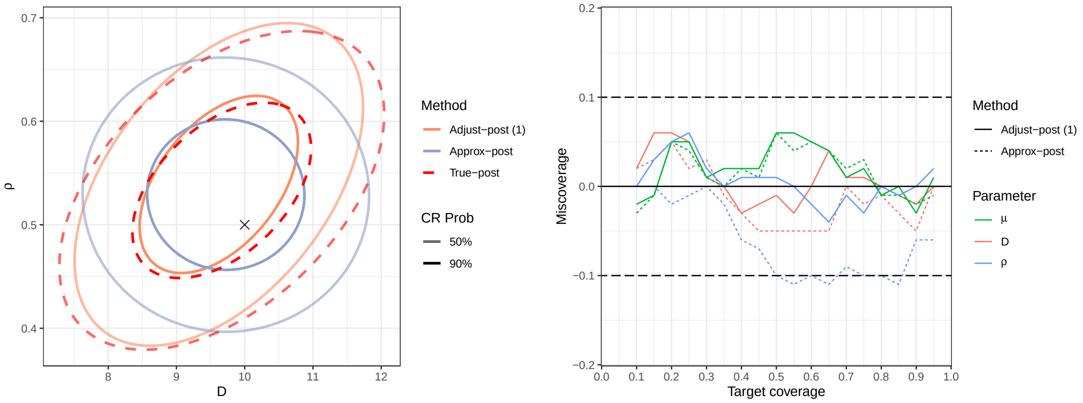
Results for the mean squared error (MSE), bias, standard deviation (SD), and achieved coverage (AC, 90%) are presented in Table 1. It is clear that the approximate posterior performs poorly for μ. Despite this, the adjustment method is still able to produce results that are similar to the true posterior on average. The approximate method already produces accurate inferences similar to the true posterior for D so that the adjustment is negligible. As an example, the posterior results based on running the adjustment process on a single data set are shown in Figure 2.

3.1.2 Bivariate OU Process with Variational Approximation

We can define a bivariate OU process by considering $\{Y_t\}_{t\geq 0}$ for $Y_t \in \mathbb{R}^2$, such that the components are $Y_{t,1} = X_{t,1}$ and $Y_{t,2} = \rho X_{t,1} + (1-\rho)X_{t,2}$ where $X_{t,1}$ and $X_{t,2}$ are independent OU processes, conditional on shared parameters (μ, γ, σ), and governed by (11). The additional parameter $\rho \in [0,1]$ measures the correlation between $Y_{t,1}$ and $Y_{t,2}$.

Again, we consider the true model for $X_{t,1}$ and $X_{t,2}$ to be defined by (12), therefore $(Y_{t,1}, Y_{t,2})$ have joint distribution that is bivariate Gaussian with correlation ρ. For the observed data we take 100 independent realizations simulated from the above model with $x_{0,1} = x_{0,2} = 5$, $\mu = 1$, $\gamma = 2$, $T = 1$, $\sigma^2 = 20$, and $\rho = 0.5$. We assume that $x_{0,1}$, $x_{0,2}$, and γ are known and we attempt to infer μ, $D = \sigma^2/2$, and ρ. We use independent priors where $\mu \sim \mathrm{N}(0, 10^2)$, $D \sim \mathrm{Exp}(1/10)$, and $\rho \sim \mathrm{U}(0,1)$. We use automatic differentiation variational inference (Kucukelbir et al., 2017) for the approximate model with a mean-field approximation as the variational family as implemented in `Turing.jl`. We expect the correction from our method will need to introduce correlation in the posterior due to the independence inherited from the mean-field approximation.

The variational approximation of the bivariate OU posterior estimates the mean and variance well in this example. The univariate bias, MSE, and coverage metrics are very similar for the approximate, adjusted ($\alpha = 1$) and true posteriors (Table 3 in Appendix E). However, the adjusted posterior with $\alpha = 0$ is poor due to high variance of the weights, perhaps due to the increased dimension of this example. This could be corrected using an appropriate stabilizing function, but we leave this for future research.



(a) Bivariate density contours of ρ and D

(b) Calibration diagnostic

Figure 3: Posterior summaries of the bivariate OU Process model from a single simulation. Plot (a) shows 50% and 90% credible region probability (CR Prob) contours from a Gaussian approximation to the bivariate density of ρ and D. The original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping are shown with solid lines. The true posterior (True-post) is shown with a dashed line. The true generating parameter value is indicated with a cross ($\times$). Plot (b) is a calibration diagnostic showing the marginal miscoverage for all parameters (see Section 2.6) for $\alpha = 1$ with ± 0.1 deviation from parity shown with a dotted line.

Despite the good univariate properties of the variational approximate posterior, the choice of a mean-field approximation cannot recover the correlation between the parameters, in particular, ρ and D. Figure 3(a) shows an example of the independence between ρ and D in the approximate posterior and how the adjusted posterior corrects this (from one data set). To investigate the method's ability to recover the correlation structure we monitor the empirical correlation between ρ and D over 100 independent trials. Adjusting the approximate posterior ($\alpha = 1$) increased the mean correlation from 0.00 to 0.31, a major improvement compared to 0.41 for the true posterior. Further correlation summaries are provided in Table 2.

Figure 3(b) illustrates the calibration diagnostic from Section 2.6 for the same simulation as Figure 3(a). We can see the learned transformation leads to a well-calibrated adjusted posterior as the miscoverage is close to the target coverage for the range considered, and improves the coverage compared to the original approximate posterior of ρ.

3.2 Lotka-Volterra Model with Kalman Filter Approximation

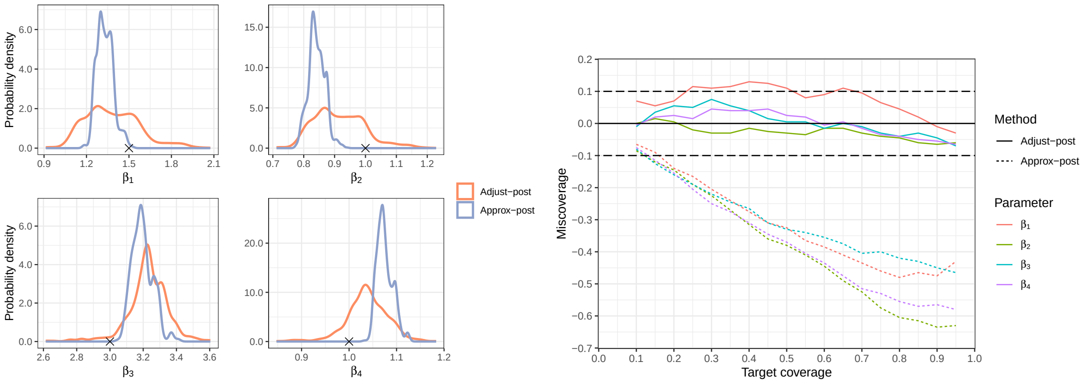
We consider the Lotka-Volterra, or predator-prey, dynamics governed by a stochastic differential equation (SDE). Let $\{(X_t, Y_t)\}_{t\geq 0}$ be a continuous time stochastic process defined by the SDE

$$\mathrm{d}X_t = (\beta_1 X_t - \beta_2 X_t Y_t)\mathrm{d}t + \sigma_1 \mathrm{d}B_t^1, \quad \mathrm{d}Y_t = (\beta_4 X_t Y_t - \beta_3 Y_t)\mathrm{d}t + \sigma_2 \mathrm{d}B_t^2,$$

where $\{B_t^k\}_{t\geq 0}$ are independent Brownian noise processes for $k \in \{1, 2\}$. For this example, we assume the pairs (x_t, y_t) are observed without error at times $t \in \{0, 0.2, 0.4, \ldots, 6\}$, for a total of $n = 31$ observations. We use initial values $X_0 = Y_0 = 1$ and simulate the observations with true parameter values $\beta_1 = 1.5, \beta_2 = \beta_4 = 1.0, \beta_3 = 3.0$, and $\sigma_1 = \sigma_2 = 0.1$, using the SOSRI solver (Rackauckas and Nie, 2020).

We define our target posterior with priors $\beta_i \sim \mathcal{U}(0.1, 4)$ iid for $i \in \{1, 2, 3, 4\}$ and $\sigma_j \sim \mathcal{U}(0.01, 0.25)$ iid for $j \in \{1, 2\}$. We use a continuous-discrete Extended Kalman Filter (EKF, Jazwinski, 1970) as the approximate likelihood and draw samples using the No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2014) for Hamiltonian Monte Carlo (HMC, Neal, 2011) algorithm from `Turing.jl`. For the prediction step, the EKF requires numerical integration of the moment equations and we use a Euler-Maruyama scheme with step-size $\Delta t = 0.05$ (see Frogerais et al., 2011, for an overview). The EKF also requires specification of the observation noise, which we choose to be $(x_t, y_t) \sim \mathcal{N}((X_t, Y_t), \tau^2 I)$, where the prior for the standard deviation is the positive truncated normal distribution, $\tau \sim \mathcal{N}^+(0, 0.05^2)$, chosen to favor the low noise regime we are simulating data from.

We use $M = 200$ calibration samples, unit weights (i.e., $\alpha = 1$), and draw 1000 samples for the approximate posteriors. In this example, the adjusted posterior does significantly better than the approximate posterior. Figure 4(b) shows the marginal approximate posteriors have extreme under-coverage which is corrected by our method. In Figure 4(a) we observe strong bias in parameters β_2 and β_4 which is removed, whilst the variance of β_1 is inflated to ensure coverage. Whilst the marginal distribution of β_3 appears mostly unchanged, the adjustment does ensure probability mass at the true value, where the approximate posterior had none.



(a) Posterior distribution comparison

(b) Calibration diagnostic

Figure 4: Comparison of original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) for Lotka-Volterra example with EKF likelihood. Plot (a) shows the estimated marginal posterior densities, with true generating parameter value indicated with a cross (×). Plot (b) is a calibration diagnostic showing the marginal miscoverage for all parameters with ±0.1 deviation from parity shown with a dotted line.

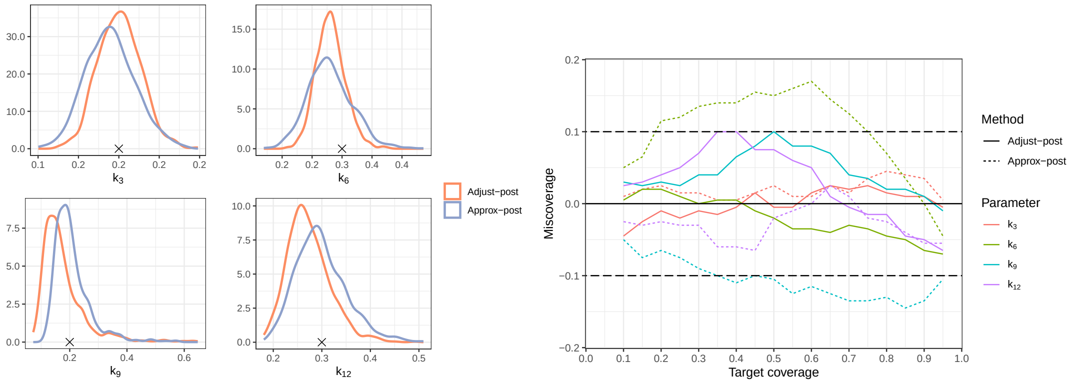
3.3 Stochastic Chemical Kinetic Model

Finally, we demonstrate our method on a difficult biological model with intractable likelihood and parameter non-identifiability. The chemical reaction network we consider is a two-step Mitogen Activated Protein Kinase (MAPK) enzymatic cascade (Dhananjaneyulu et al., 2012). Such reaction networks are often used as components of larger systems to model cell signaling processes (Brown et al., 2004; Oda et al., 2005).

The two-step MAPK model is a stochastic chemical kinetic model for proteins X and Y, activated (phosphorylated) proteins X^a and Y^a, enzyme E, and phosphatase molecules P_1 and P_2. The two-step MAPK reaction network governs the phosphorylation and dephosphorylation of X and Y by coupled Michaelis–Menten components that can be expressed as

$$\begin{array}{lll} X + E \xrightarrow{k_1} [XE], & [XE] \xrightarrow{k_2} X + E, & [XE] \xrightarrow{k_3} X^a + E, \\ X^a + P_1 \xrightarrow{k_4} [X^aP_1], & [X^aP_1] \xrightarrow{k_5} X^a + P_1, & [X^aP_1] \xrightarrow{k_6} X + P_1, \\ X^a + Y \xrightarrow{k_7} [X^aY], & [X^aY] \xrightarrow{k_8} X^a + Y, & [X^aY] \xrightarrow{k_9} X^a + Y^a, \\ Y^a + P_2 \xrightarrow{k_{10}} [Y^aP_2], & [Y^aP_2] \xrightarrow{k_{11}} Y^a + P_2, & [Y^aP_2] \xrightarrow{k_{12}} Y + P_2. \end{array} \tag{13}$$

Equation (13) describes the processes of X activating (to X^a) by compounding with E, and deactivating similarly by P_1, whilst Y (to Y^a) can be activated by compounding with X^a and deactivated similarly by P_2. Hence, there is a cascading effect, or two steps, where X must be activated in order for Y to be activated. The network also allows for the compounds to form without an eventual activation or deactivation taking place. The parameters $k_1, k_2, \dots, k_{12}$ dictate the rate of the reactions in the system and are the parameters of interest.



(a) Posterior distribution comparison

(b) Calibration diagnostic

Figure 5: Comparison of original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) for reaction network example with EKF likelihood. Plot (a) shows the estimated marginal posterior densities, with true generating parameter value indicated with a cross (×). Plot (b) is a calibration diagnostic showing the marginal miscoverage for all parameters with ±0.1 deviation from parity shown with a dotted line.

Following Warne et al. (2022b) we assume that only the activated proteins are observable with additive normal noise $(x_t^a, y_t^a) \sim \mathcal{N}((X_t^a, Y_t^a), \sigma^2 I)$, with observations taken at $t \in \{4, 8, \ldots, 200\}$, known initial counts $E = 94$, $X = 757$, $Y = 567$, $P_1 = P_2 = 32$, and remaining counts zero at $t = 0$. The partial observation of proteins leads to parameter unidentifiability, hence we fix $k_1 = k_4 = k_{10} = 10^{-3}$, $k_7 = 10^{-4}$, whilst assigning the following priors $k_i \sim \mathcal{U}(0, 10^{-3})$ iid for $i \in \{2, 5, 11\}$, $k_j \sim \mathcal{U}(0, 1)$ iid for $j \in \{3, 6, 9, 12\}$, and $k_8 \sim \mathcal{U}(0, 10^{-4})$. Using the aforementioned initial conditions and noise $\sigma = 1$ we simulate the observations using the Gillespie's direct method (Gillespie, 1992), with parameters $k_1 = k_4 = k_{10} = 10^{-3}$, $k_2 = k_1/120$, $k_3 = 0.18$, $k_5 = k_4/22$, $k_6 = k_{12} = 0.3$, $k_7 = 10^{-4}$, $k_8 = k_7/110$, $k_9 = 0.2$, and $k_{11} = k_{10}/22$.

The approximate likelihood is defined by two layers of approximations in this example. Firstly, we make a diffusion approximation to the discrete-state continuous-time chemical reaction network, resulting in an SDE. Then, as in Section 3.2, we use the EKF to approximate the SDE's likelihood. We take $\Delta t = 1$ for the intermediary integration times between observations in the EKF, and sample 1000 draws using NUTS-HMC for each approximate posterior. We use $M = 200$ calibration samples, unit weights (i.e., $\alpha = 1$), and scale the variance of the approximate posterior (with the observed data) by a factor of 1.5 to define $\bar{\Pi}$. We correct the marginal posterior distribution of (k_3, k_6, k_9, k_{12}) as the remaining parameters are unidentifiable from the data.

From Figure 5(a) we can see the correction to the approximate posterior is modest compared to previous examples, indicating the EKF may provide a reasonable approximate likelihood for this model and data set. However, Figure 5(b) provides a strong motivation for calculating and using the adjusted posterior as the over-coverage of k_6 and under-coverage of k_9 have been corrected to within ±0.1 for all target coverage values. Despite the unknown properties of the EKF approximation to the MAPK model (and other reaction networks), the calibration diagnostic provides reassurance that this approximation can be reasonable for this data set after the score calibration correction is made.

4. Discussion

In this paper we have presented a new approach for modifying posterior samples based on an approximate model or likelihood to improve inferences with respect to some complex target model. Our approach does not require any likelihood evaluations of the target model, and only a small number of target model simulations and approximate posterior computations, which are easily parallelizable. Our approach is particularly suited to applications where the likelihood of the target model is completely intractable, or if the surrogate likelihood is several orders of magnitude faster to evaluate than the target likelihood. We also demonstrated in Section 3.3 that several layers of approximation could be corrected for.

We focused on correcting inferences from an approximate model, but our approach can also be applied when the inference algorithm is approximate. For example, we could use our approach to adjust inferences from likelihood-free algorithms or to correct the bias in short MCMC runs. We plan to investigate this in future research.

We propose a straightforward clipping method when we wish to guarantee finite weights in our importance sampling step. However, a more sophisticated approach to clipping is possible, namely Pareto smoothed importance sampling (PSIS, Vehtari et al., 2024). PSIS can be used instead of clipping or unit weights if desired, but we leave investigation of this strategy for future work.

In general, our method can be used with any proper scoring rule. We concentrated on the energy score because of the ease with which it can be estimated using transformed samples from the approximate distribution. The logarithmic score could also be used by calculating a kernel density estimate from the adjusted posteriors. We found that using $\beta = 1$ for the energy score provided good results in our experiments, but it may be of interest to try $\beta \neq 1$ in future work.

Using scoring rules that are not strictly proper would alter the objective of Theorem 2 to recovering certain properties of the target posterior, rather than the entire target posterior itself. Such a choice of scoring rule would simplify the class of sufficiently rich kernels, both asymptotically and non-asymptotically, at the expense of not attempting to recover the full target posterior approximately. This trade-off represents an interesting research direction we are pursuing.

In ongoing work we are considering more flexible transformations when warranted by deficiencies in the approximate posterior. This will be particularly important when the direction of the bias in the approximate model changes in different regions of the parameter space. However, learning more flexible transformations will likely necessitate more calibration samples and data sets, and increased optimization time. We note that our method with the energy score does not require invertible or differentiable transformations, making it quite flexible compared to most transformation-based inference algorithms which require differentiability.

Another limitation of our approach is that we do not expect our current method to necessarily generate useful corrections when the approximate posterior is very poor. An example where this occurs, to some extent, in a Lotka-Volterra model is provided in

Supplement S.3. A poor approximation would likely lead to a family of kernels that is not sufficiently rich. If the correcting transformation is learned in a region of the parameter space far away from the true posterior mass, the calibration data sets are likely to be far away from the observed data, and the transformation may not successfully calibrate the approximate posterior conditional on the observed data. We did provide arguments for when we expect moment-correcting transformations to be asymptotically sufficiently rich and a practical method for detecting insufficiency. However, as already alluded to, a more flexible transformation may be able to calibrate more successfully across a wider set of parameter values (e.g., the prior or inflated version of the approximate posterior). We also note that in this paper we assume that the complex target model is correctly specified, and an interesting future direction would consider the case where the target model itself is possibly misspecified.

Acknowledgments

JJB was supported by a First Byte Grant from the Centre for Data Science (Queensland University of Technology) and the European Union under the GA 101071601, through the 2023-2029 ERC Synergy grant OCEAN. DJW is supported by an ARC Discovery Early Career Researcher Award (DE250100396). DJN was supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 2 (MOE-T2EP20123-0009). CD was supported by an Australian Research Council Future Fellowship (FT210100260).

Appendix A. Illustration with Logistic Regression

In this section, we illustrate Bayesian score calibration on a simple worked example to elucidate the framework. Consider binary data $y_i \in \{0,1\}$ and fixed regression variables $X_i \in \mathbb{R}^p$ for $i \in \{1,\dots,n\}$. Suppose we wish to fit a logistic regression with a default prior Π for parameter vector $\theta \in \mathbb{R}^p$ (e.g., Gelman et al., 2008) related to the data by $\mathbb{P}(y_i = 1 \mid \theta) = \frac{1}{1+\exp(-\theta^\top X_i)}$. Such a regression would typically involve some type of Monte Carlo sampling algorithm to draw approximate samples from the posterior distribution. Instead, assume that we are interested in using a fast approximation of the posterior distribution that is subsequently calibrated using our method.

Suppose our approximate posterior is generated using the Laplace approximation to a Bayesian logistic regression with flat priors. For a given data set $\tilde{y} \in \{0,1\}^n$ this approximation is a multivariate normal distribution $\hat{\Pi}(\cdot \mid \tilde{y}) = \mathcal{N}(\hat{\theta}(\tilde{y}), \hat{\Omega}(\tilde{y})^{-1})$ with approximate mean, $\hat{\theta}(\tilde{y})$ the logistic regression MLE, and approximate precision, $\hat{\Omega}(\tilde{y})$ the Hessian matrix of the negative log-likelihood at the MLE. Both the mean and precision are functions of the simulated data $\tilde{y}$ with fixed covariate matrix $X^\top = [X_1 \cdots X_n]$.

We calibrate the Laplace approximation with a correction from the relative moment-correcting transformation family, described in Section 2.4.2 and denoted here by $\mathcal{F}_\mathrm{m}$. This implies that our family of kernels $\mathcal{K}_\mathrm{m}$ is described by transformations, $f \in \mathcal{F}_\mathrm{m}$, applied to the Laplace approximation for data $\tilde{y}$. As such, the best calibrated posterior kernel is selected from

$$\mathcal{K}_\mathrm{m} = \{f_\sharp \hat{\Pi} : f \in \mathcal{F}_\mathrm{m}\} = \{\mathcal{N}(\hat{\theta}(\cdot) + b, A\hat{\Omega}(\cdot)^{-1}A^\top) : b \in \mathbb{R}^p, A \in \mathcal{A}\}.$$

The space $\mathcal{A} \subset \mathbb{R}^{p\times p}$ is specified in Section 2.4.2 and ensures the optimal A is unique.

Several further choices are required to instantiate the Bayesian score calibration framework. We use the energy scoring rule with $\beta = 1$, unit stabilizing function, and importance distribution, $\bar{\Pi} = \mathcal{N}(\hat{\theta}(y), 2\hat{\Omega}(y)^{-1})$, based on the Laplace approximation of the true data, with variance rescaled by 2. With this specification, the idealized objective function is

$$\mathbb{E}_{\theta\sim\bar{\Pi}}\mathbb{E}_{\tilde{y}\sim P(\cdot|\theta)}\left\{r(\theta)\left[\frac{1}{2}\mathbb{E}_{u,u'\sim\hat{\Pi}(\cdot|\tilde{y})}\|f(u)-f(u')\|_2 - \mathbb{E}_{u\sim\hat{\Pi}(\cdot|\tilde{y})}\|f(u)-\theta\|_2\right]\right\}, \tag{14}$$

where $r(\theta)$ is the importance weight of prior Π to importance distribution $\bar{\Pi}$, and $\tilde{y} \sim P(\cdot \mid \theta)$ is the data-generating process implied by the logistic regression. That is, $\mathbb{P}(\tilde{y}_i = 1 \mid \theta) = \frac{1}{1+\exp(-\theta^\top X_i)}$ and $\mathbb{P}(\tilde{y}_i = 0 \mid \theta) = 1 - \mathbb{P}(\tilde{y}_i = 1 \mid \theta)$ with X_i fixed by assumption. Maximizing (14) for $f \in \mathcal{F}_{\mathrm{m}}$ yields a correction for the Laplace approximation that is better calibrated to the true posterior distribution (on average, according to the energy score). Finally, using the observed data y and optimal parameters $(b^\star, A^\star)$ from (14), the calibrated approximate posterior is $\mathcal{N}(\hat{\theta}(y) + b^\star, A^\star\hat{\Omega}(y)^{-1}A^{\star\top})$.

To run the Bayesian score calibration algorithm in practice, we construct a Monte Carlo approximation to the objective function (14) and find the maximizer $f \in \mathcal{F}_{\mathrm{m}}$ by optimizing over $b \in \mathbb{R}^p$ and $A \in \mathcal{A}$.

Appendix B. Additional Theorems and Proofs

B.1 Proof of Theorem 2

Proof Denote the objective function in (3) as $E(K)$. Using the Radon-Nikodym derivative $\mathrm{d}\Pi/\mathrm{d}\bar{\Pi}$ we can rewrite this as

$$E(K) = \mathbb{E}_{\theta\sim\Pi}\mathbb{E}_{\tilde{y}\sim P(\,\cdot|\theta)}\left[v(\tilde{y})S(K(\,\cdot\mid\tilde{y}),\theta)\right] = \int \Pi(\mathrm{d}\theta)P(\mathrm{d}\tilde{y}\mid\theta)v(\tilde{y})S(K(\,\cdot\mid\tilde{y}),\theta),$$

then substituting $\Pi(\mathrm{d}\theta)P(\mathrm{d}\tilde{y}\mid\theta) = P(\mathrm{d}\tilde{y})\Pi(\mathrm{d}\theta\mid\tilde{y})$ we find that

$$E(K) = Z_v\mathbb{E}_{\tilde{y}\sim Q}\mathbb{E}_{\theta\sim\Pi(\,\cdot|\tilde{y})}\left[S(K(\,\cdot\mid\tilde{y}),\theta)\right],$$

where $Q(\mathrm{d}\tilde{y}) = P(\mathrm{d}\tilde{y})v(\tilde{y})/Z_v$, with $Z_v = P(v) \in (0,\infty)$ by assumption.

Now for maximizing $E(K)$, we can ignore the constant Z_v, and find that

$$\begin{aligned}K^\star = \underset{K\in\mathcal{K}}{\arg\max}\, E(K) &= \underset{K\in\mathcal{K}}{\arg\max}\, \mathbb{E}_{\tilde{y}\sim Q}\mathbb{E}_{\theta\sim\Pi(\,\cdot|\tilde{y})}\left[S(K(\,\cdot\mid\tilde{y}),\theta)\right]\\ &= \underset{K\in\mathcal{K}}{\arg\max}\, \mathbb{E}_{\tilde{y}\sim Q}\left[S(K(\,\cdot\mid\tilde{y}),\Pi(\,\cdot\mid\tilde{y}))\right].\end{aligned}$$

We note that $S(K(\,\cdot\mid\tilde{y}),\Pi(\,\cdot\mid\tilde{y}))$ is maximized if and only if $K(\,\cdot\mid\tilde{y}) = \Pi(\,\cdot\mid\tilde{y})$ and $K(\,\cdot\mid\tilde{y}),\Pi(\,\cdot\mid\tilde{y}) \in \mathcal{P}$ for fixed $\tilde{y} \in \mathsf{Y}$ since S is a strictly proper scoring rule relative to $\mathcal{P}$. Then under expectation with respect to Q, if $\mathcal{K}$ is sufficiently rich, the optima $K^\star$ must satisfy $K^\star(\,\cdot\mid\tilde{y}) = \Pi(\,\cdot\mid\tilde{y})$ almost surely. ■

B.2 Proof of Theorem 6

Proof Let $g_z(x) = \frac{\pi(z)}{\bar{\pi}(z)}\frac{\bar{\pi}(x)}{\pi(x)}$ and consider $v(\tilde{y}_{1:n}) = g(\theta_n^*)$. Therefore $w(z, \tilde{y}_{1:n}) = g_z(\theta_n^*)$. Applying the continuous mapping theorem with function $h(x) = g_z(z) - g_z(x) = 1 - g_z(x)$ yields the result. ■

B.3 A Central Limit Theorem for Unit Weights

Theorem 7 *Let $g(x) = \bar{\pi}(x)/\pi(x)$ for $x \in \Theta$. If there exists an estimator $\theta_n^* \equiv \theta^*(\tilde{y}_{1:n})$ such that $\sqrt{n}(\theta_n^* - z) \xrightarrow{d} N(0, \Sigma_z)$ as $n \to \infty$ when $\tilde{y}_i \overset{iid}{\sim} P(\cdot \mid z)$ for $z \in \Theta$, $g(\theta_n^*) \leq h(\tilde{y}_{1:n})$ a.s. for some integrable function h, $g > 0$ a.e., and $\nabla g \neq 0$ a.e., then the error from approximating the weights with $\hat{w} = 1$ as the size of the data $y_{1:n}$ grows satisfies*

$$\sqrt{n}(\hat{w} - w(\theta, \tilde{y}_{1:n})) \xrightarrow{d} U,$$
$$(U \mid \theta) \sim N\left(0, \Sigma'_\theta\right), \theta \sim \bar{\Pi},$$

as $n \to \infty$ with choice of stabilizing function $v(\tilde{y}_{1:n}) = g(\theta_n^*)$, where

$$\Sigma'_\theta = \nabla \log g(\theta)^\top \Sigma_\theta \nabla \log g(\theta).$$

Moreover, $\mathbb{E}(U) = 0$ and the unit weights $\hat{w}$ therefore have asymptotic distribution with variance equal to $\text{var}(U) = \mathbb{E}_{\theta\sim\bar{\Pi}}(\Sigma'_\theta)$.

Proof Take $\theta = z$ fixed and let $g_z(x) = \frac{\pi(z)}{\bar{\pi}(z)}\frac{\bar{\pi}(x)}{\pi(x)}$ for $x \in \Theta$. Consider $v(\tilde{y}_{1:n}) = g(\theta_n^*)$ and therefore $w(z, \tilde{y}) = g_z(\theta_n^*)$. Using the delta method we can deduce that

$$U_n(z) \equiv \sqrt{n}(g_z(\theta_n^*) - g_z(z)) \xrightarrow{d} U(z), \qquad \text{where } U(z) \sim \mathrm{N}\left(0, \nabla g_z(z)^\top \Sigma_z \nabla g_z(z)\right),$$

noting that $g_z(\theta_n^*) = w(z, \tilde{y})$ and $g_z(z) = 1 = \hat{w}$. Now let $\theta \sim \bar{\Pi}$ on measurable space (Θ, ϑ). For all $A \in \vartheta$, consider

$$\begin{aligned}\lim_{n\to\infty} \mathsf{P}(U_n(\theta) \in A) &= \lim_{n\to\infty} \mathbb{E}_{z\sim\bar{\Pi}}\mathsf{P}(U_n(\theta) \in A \mid \theta = z)\\ &= \mathbb{E}_{z\sim\bar{\Pi}}\mathsf{P}(U(\theta) \in A \mid \theta = z)\\ &= \mathsf{P}(U(\theta) \in A),\end{aligned}$$

by the law of total probability and noting that dominated convergence theorem holds since $0 < g(\theta_n^*) \leq h(y_{1:n})$ implies that $|U_n(z)|$ is also dominated. Therefore $U_n(\theta) \xrightarrow{d} U(\theta)$ as $n \to \infty$ where $U(\theta) \sim \mathbb{E}_{z\sim\bar{\Pi}}U(z)$, i.e., a continuous mixture of Gaussian distributions. Using the continuous mapping theorem we can also state that $U_n \xrightarrow{d} U$ where $U_n \equiv -U_n(\theta)$ and $U \equiv -U(\theta)$ then noting that $\Sigma'_z \equiv \nabla g_z(z)^\top \Sigma_z \nabla g_z(z) = \nabla \log g(z)^\top \Sigma_z \nabla \log g(z)$ gives the limiting distribution result. Moreover, we can see $\mathbb{E}(U) = 0$ by the law of total expectation and $\text{var}(U) = \mathbb{E}_{z\sim\bar{\Pi}}(\Sigma'_z)$ by the law of total variance. ■

Remark 8 *If one wishes to estimate the asymptotic variance of the weight approximation, we have freedom to choose the estimator* θ^**. If possible we should choose the estimator that results in the smallest asymptotic variance* $\mathrm{var}(U)$ *or smallest conditional variance* Σ_θ *if equivalent or more convenient.*

Remark 9 *If* $\max_{x\in\Theta} g(x) = m < \infty$ *then* $g(\theta_n^*) \leq m$ *and the dominating condition holds. This indicates that using a distribution* $\bar{\Pi}$ *with lighter tails than* Π *is appropriate. Such a statement is surprising as this disagrees with well-established importance sampling guidelines. Moreover, if* $\bar{\pi}(\theta) = \hat{\pi}(\theta \mid y_{1:n})$ *then* $g(\theta)$ *is the approximate likelihood (ignoring the normalizing constant) and a sufficient condition for the domination is that the approximate likelihood* $g(\theta) = \hat{p}(y_{1:n} \mid \theta)$ *is bounded. This is the case for any approximate likelihood for which a maximum likelihood estimate exists.*

Remark 10 *The dominating condition can also be enforced by only considering bounded* Θ*. As such the estimator* θ_n^* *and hence* $g(\theta_n^*)$ *will typically be bounded. This is the approach taken by Deistler et al. (2022) for sequential neural posterior estimation but no asymptotic justification is given. Our results may be useful in this case, and more generally for this area, but we leave exploration for future research.*

Appendix C. Package and Code Acknowledgments

A `julia` package with code that can be applied to any approximate model and data-generating process can be found at `https://github.com/bonStats/BayesScoreCal.jl`, our examples are contained in `https://github.com/bonStats/BayesScoreCalExamples.jl`.

Our implementation relies heavily on `Optim.jl` (Mogensen and Riseth, 2018) and our examples make use of `DifferentialEquations.jl` (Rackauckas and Nie, 2017), `Distributions.jl` (Besançon et al., 2021), and `Turing.jl` (Ge et al., 2018).

Appendix D. Idealized Versus Practical Weighting Functions

An optimal stabilizing function would necessitate that $w(\theta, \tilde{y}) = C$, for some constant C, though such a function need not exist. However, considering the properties of a theoretical optimal stabilizing function is useful for our asymptotic results in Section 2.5.1. If there were a deterministic function g perfectly predicting θ from $\tilde{y}$, i.e., $g(\tilde{y}) = \theta$ if $\tilde{y} \sim P(\cdot \mid \theta)$, then $v(\tilde{y}) = \bar{\pi}[g(\tilde{y})]/\pi[g(\tilde{y})]$ would be the optimal stabilizing function.

In the absence of such a g, we could approximate the stabilizing function by

$$v(\tilde{y}) = \frac{\bar{\pi}[\theta^\star(\tilde{y})]}{\pi[\theta^\star(\tilde{y})]}, \quad \theta^\star(\tilde{y}) = \arg\max_{\vartheta\in\Theta} p(\tilde{y} \mid \vartheta)\bar{\pi}(\vartheta), \tag{15}$$

where $\theta^\star(\tilde{y})$ is the maximum *a posteriori* (MAP) estimate of θ given $\tilde{y}$. The maximum likelihood estimate could also be used. In the case that $\theta^\star(\tilde{y}) \approx \theta$ we can deduce that $w(\theta, \tilde{y}) \approx C$, though deviations from this may be quite detrimental to the variance of the weights. Unfortunately we do not have access to the likelihood $p(\tilde{y} \mid \cdot)$, so $\theta^\star$ is intractable. The approximate likelihood $\hat{p}$ is a practical replacement for p but will be likely to further increase the variance of the weights.

As for the importance distribution, a natural way to concentrate θ about likely values of the posterior given y is to use $\bar{\Pi}(\cdot) = \hat{\Pi}(\cdot \mid y)$. This generates data sets $\tilde{y}$ such that they are consistent with y according to the approximate posterior. The idealized setting, with no Monte Carlo error and an accurate MAP using the approximate likelihood, therefore uses

$$v(\tilde{y}) = \frac{\bar{\pi}[\theta^{\diamond}(\tilde{y})]}{\pi[\theta^{\diamond}(\tilde{y})]}, \quad \theta^{\diamond}(\tilde{y}) = \arg\max_{\vartheta \in \Theta} \hat{p}(\tilde{y} \mid \vartheta)\bar{\pi}(\vartheta),$$

with $\bar{\Pi}(\cdot) = \hat{\Pi}(\cdot \mid y)$. If $\theta^{\diamond}(\tilde{y})$ is a biased estimator of θ, we could estimate this bias and correct for it to ensure $w(\theta, \tilde{y}) \approx C$.

In some cases choosing $\bar{\Pi}(\cdot) = \hat{\Pi}(\cdot \mid y)$ may be adequate, but it depends crucially on the tail behavior of the ratio $\pi(\theta)/\bar{\pi}(\theta)$ and how well the stabilizing function performs. It may be pertinent to artificially increase the variance of the $\bar{\Pi}$ by transformation or consider an approximation to the chosen distribution with heavier tails.

The simple countermeasure we consider is to truncate or clip the weights as discussed in Section 2.5.

Appendix E. Additional Results from Examples

This section contains additional results from the examples in Section 3.

	Correlation	
Method	Mean	SD
Approx-post	0.00	0.02
Adjust-post (0)	0.25	0.38
Adjust-post (0.5)	0.29	0.12
Adjust-post (1)	0.31	0.12
True-post	0.41	0.06

Table 2: Summary of empirical correlation between parameter samples of ρ and D over 100 independent data sets for the bivariate OU process example. The posteriors compared are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping, and the true posterior (True-post).

Method	MSE	Bias	SD	AC[1]
μ				
Approx-post	0.10	0.00	0.22	92
Adjust-post (0)	0.11	0.00	0.18	73
Adjust-post (0.5)	0.10	0.01	0.22	90
Adjust-post (1)	0.10	0.01	0.22	89
True-post	0.10	0.00	0.22	90
D				
Approx-post	2.51	0.11	1.01	84
Adjust-post (0)	2.99	0.17	0.96	72
Adjust-post (0.5)	2.73	0.11	1.08	83
Adjust-post (1)	2.73	0.12	1.08	83
True-post	2.84	0.15	1.15	87
ρ				
Approx-post	0.01	−0.02	0.07	84
Adjust-post (0)	0.01	−0.01	0.06	73
Adjust-post (0.5)	0.01	−0.01	0.07	85
Adjust-post (1)	0.01	−0.01	0.07	86
True-post	0.01	−0.02	0.08	85

Table 3: Average results for each parameter over 100 independent data sets for the bivariate OU process example. The posteriors compared are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping. [1]Achieved coverage with 90% target.

References

O. Barndorff-Nielsen and G. Schou. On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3(4):408–419, 1973.

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

M. Besançon, T. Papamarkou, D. Anthoff, A. Arslan, S. Byrne, D. Lin, and J. Pearson. Distributions.jl: Definition and modeling of probability distributions in the JuliaStats ecosystem. *Journal of Statistical Software*, 98(16):1–30, 2021.

A. Beskos, A. Jasra, N. Kantas, and A. Thiery. On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146, 2016.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

P. Billingsley. *Probability and Measure.* John Wiley & Sons, Inc, New York, NY, 3rd edition, 1995.

C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, New York, NY, 2006.

J. J. Bon, A. Lee, and C. Drovandi. Accelerating sequential Monte Carlo with surrogate likelihoods. *Statistics and Computing*, 31(5):1–26, 2021.

P. Bortot, S. G. Coles, and S. A. Sisson. Inference for stereological extremes. *Journal of the American Statistical Association*, 102(477):84–92, 2007.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo.* CRC Press, Boca Raton, FL, 2011.

K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, and R. A. Cerione. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Physical Biology*, 1(3):184, 2004.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo.* Springer, Cham, Switzerland, 2020.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

M. Deistler, P. J. Goncalves, and J. H. Macke. Truncated proposals for scalable and hassle-free simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 35, pages 23135–23149, 2022.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

V. Dhananjaneyulu, V. N. Sagar P, G. Kumar, and G. A. Viswanathan. Noise propagation in two-step series MAPK cascade. *PloS One*, 7(5):e35958, 2012.

D. T. Frazier, D. J. Nott, C. Drovandi, and R. Kohn. Bayesian inference using synthetic likelihood: Asymptotics and adjustments. *Journal of the American Statistical Association*, 118(544):2821–2832, 2023.

P. Frogerais, J.-J. Bellanger, and L. Senhadji. Various ways to compute the continuous-discrete extended Kalman filter. *IEEE Transactions on Automatic Control*, 57(4):1000–1004, 2011.

H. Ge, K. Xu, and Z. Ghahramani. Turing: A language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1682–1690. PMLR, 2018.

A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2 (4):1360 – 1383, 2008.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

D. Greenberg, M. Nonnenmacher, and J. Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, volume 97, pages 2404–2414. PMLR, 2019.

M. U. Gutmann and J. Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125):1–47, 2016.

M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, New York, 1970.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

J. E. Lee, G. K. Nicholls, and R. J. Ryder. Calibration procedures for approximate Bayesian credible sets. *Bayesian Analysis*, 14(4):1245–1269, 2019.

J. Lei and P. Bickel. A moment matching ensemble filter for nonlinear non-Gaussian data assimilation. *Monthly Weather Review*, 139(12):3964–3973, 2011.

J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

G. Maruyama. Continuous Markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4:48–90, 1955.

P. Menéndez, Y. Fan, P. H. Garthwaite, and S. A. Sisson. Simultaneous adjustment of bias and coverage probabilities for confidence intervals. *Computational Statistics & Data Analysis*, 70:35–44, 2014.

P. K. Mogensen and A. N. Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):1–3, 2018.

U. K. Müller. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, New York, NY, 2011.

K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology*, 1(1):1–17, 2005.

L. Pacchiardi and R. Dutta. Likelihood-free inference with generative neural networks via scoring rule minimization. *arXiv preprint arXiv:2205.15784*, 2022.

G. Papamakarios and I. Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, volume 89, pages 837–848. PMLR, 2019.

D. Prangle, M. G. Blum, G. Popovic, and S. A. Sisson. Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, 56(4):309–329, 2014.

T. P. Prescott and R. E. Baker. Multifidelity approximate Bayesian computation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):114–138, 2020.

T. P. Prescott, D. J. Warne, and R. E. Baker. Efficient multifidelity likelihood-free Bayesian inference with adaptive computational resource allocation. *Journal of Computational Physics*, 496:112577, 2024.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

C. Rackauckas and Q. Nie. DifferentialEquations.jl – a performant and feature-rich ecosystem for solving differential equations in julia. *The Journal of Open Research Software*, 5(1), 2017.

C. Rackauckas and Q. Nie. Stability-optimized high order methods and stiffness detection for pathwise stiff stochastic differential equations. In *IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–8. IEEE, 2020.

G. S. Rodrigues, D. Prangle, and S. A. Sisson. Recalibration: A post-processing method for approximate Bayesian computation. *Computational Statistics & Data Analysis*, 126: 53–66, 2018.

R. Salomone, M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran. Spectral subsampling MCMC for stationary time series. In *International Conference on Machine Learning*, pages 8449–8458. PMLR, 2020.

C. Sherlock, A. Golightly, and D. A. Henderson. Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *Journal of Computational and Graphical Statistics*, 26 (2):434–444, 2017.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, Boca Raton, FL, 2018.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. *AAAI Conference on Artificial Intelligence*, 30(1), 2016.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of Brownian motion. *Physical Review*, 36:823–841, 1930.

S. M. Vandeskog, S. Martino, and R. Huser. An efficient workflow for modelling high-dimensional spatial extremes. *Statistics and Computing*, 34(4):137, 2024.

A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

D. J. Warne, R. E. Baker, and M. J. Simpson. Rapid Bayesian inference for expensive stochastic models. *Journal of Computational and Graphical Statistics*, 31(2):512–528, 2022a.

D. J. Warne, T. P. Prescott, R. E. Baker, and M. J. Simpson. Multifidelity multilevel Monte Carlo to accelerate approximate Bayesian parameter inference for partially observed stochastic processes. *Journal of Computational Physics*, 469:111543, 2022b.

D. J. Warne, O. J. Maclaren, E. J. Carr, M. J. Simpson, and C. Drovandi. Generalised likelihood profiles for models with intractable likelihoods. *Statistics and Computing*, 34 (1):50, 2024.

P. Whittle. Estimation and information in stationary time series. *Arkiv för Matematik*, 2(5): 423–434, 1953.

H. Xing, G. Nicholls, and J. E. Lee. Calibrated approximate Bayesian inference. In *International Conference on Machine Learning*, volume 97, pages 6912–6920. PMLR, 2019.

H. Xing, G. Nicholls, and J. E. Lee. Distortion estimates for approximate Bayesian inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 1208–1217. PMLR, 2020.

Supplementary Materials

The supplementary materials contain three additional examples of the Bayesian score calibration method.

S.1 Conjugate Gaussian model

We consider a toy conjugate Gaussian example. Here the data y are $n = 10$ independent samples from a $\mathrm{N}(\mu, \sigma^2)$ distribution with $\sigma^2 = 1$ assumed known and μ unknown. Assuming a Gaussian prior $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, the posterior is $(\mu|y) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\mu_{\text{post}} = \frac{1}{\sigma_0^{-2} + n\sigma^{-2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n y_i}{\sigma^2} \right) \text{ and } \sigma^2_{\text{post}} = \frac{1}{\sigma_0^{-2} + n\sigma^{-2}}.$$

We assume this is the target model. For the approximate model, we introduce random error into the posterior mean and standard deviation

$$\mu_{\text{approx}} = \frac{\mu_{\text{post}} - \mu_{\text{error}}}{\sigma_{\text{error}}} \text{ and } \sigma_{\text{approx}} = \frac{\sigma_{\text{post}}}{\sigma_{\text{error}}}, \tag{16}$$

where $\mu_{\text{error}} \sim \mathrm{N}(0.5, 0.025^2)$ and $\sigma_{\text{error}} \sim \mathrm{FN}(1.5, 0.025^2)$ and FN denotes the folded-normal distribution. During our simulation the approximate posterior distribution is calculated for each dataset according to (16). The perturbation is random but remains fixed for each dataset. For the stabilizing function we use $v(\tilde{y}) = 1$ for simplicity.

We coded this simulation in `R` (R Core Team, 2021) using exact sampling for the true and approximate posteriors. Results based on 100 independent datasets generated from the model with true value $\mu = 1$, and prior parameters $\mu_0 = 0, \sigma_0^2 = 4^2$ are shown in Table 4. We truncate the weights from the model at quantiles from the empirical weight distribution. We test truncating the weights for $\alpha = 0$ (no clipping), $0.25, 0.5, 0.9$, and 1 (uniform weights). It can be seen that the adjusted approximation (for all α) is a marked improvement over the initial approximation, which is heavily biased and has poor coverage. The estimated posterior distributions based on a single dataset is shown in Figure 6 as an example. We can see that, for this example dataset, the adjusted approximate posteriors are a much better approximation to the true posterior.

S.2 Fractional ARIMA model

Let $\{X_t\}_{t=1}^n$ be a zero-mean equally spaced time series with stationary covariance function $\kappa(\tau, \theta) = \mathbb{E}(X_t X_{t-\tau})$ where θ is a vector of model parameters. Here we consider an autoregressive fractionally integrated moving average model (ARFIMA) model for $\{X_t\}_{t=1}^n$, described by the polynomial lag operator equation as

$$\phi(L)(1 - L)^d X_t = \vartheta(L)\epsilon_t,$$

where $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$, L is the lag operator, $\phi(z) = 1 - \sum_{i=1}^p \phi_i z_i^i$ and $\vartheta(z) = 1 + \sum_{i=1}^q \vartheta_i z_i^i$. We denote the observed realized time series as $y = (y_1, y_2, \dots, y_n)^\top$ where n is the number of observations. Here we consider an ARFIMA(p, d, q) model where $p = 2$, $q = 1$, and the observed data is simulated with the true parameter $\theta = (\phi_1, \phi_2, \vartheta_1, d)^\top = (0.45, 0.1, -0.4, 0.4)^\top$

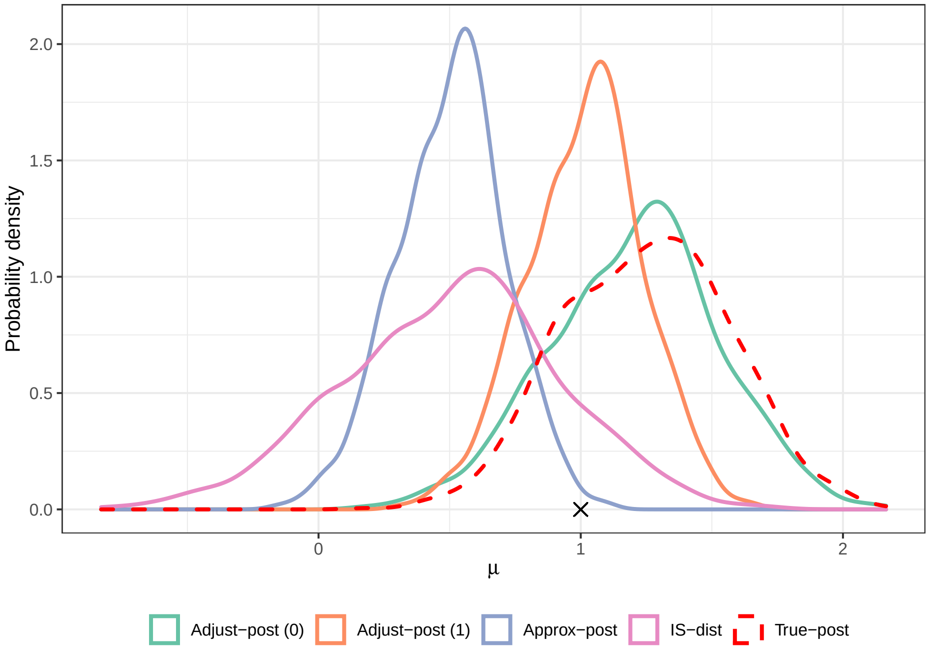


Figure 6: Conjugate Gaussian model univariate densities estimates of approximations to the posterior distribution for a single dataset. The original approximate posterior (Approx-post), importance (IS-dist), and adjusted posteriors (Adjust-post) with (α) clipping are shown with solid lines. The true posterior (True-post) is shown with a dashed line. The true generating parameter value is indicated with a cross ($\times$).

Method	MSE	Bias	SD	AC[1]
Approx-post	0.48	-0.64	0.21	0.24
Adjust-post (0)	0.21	-0.16	0.31	0.99
Adjust-post (0.25)	0.15	-0.18	0.26	0.98
Adjust-post (0.5)	0.15	-0.18	0.25	0.98
Adjust-post (0.9)	0.14	-0.18	0.25	0.98
Adjust-post (1)	0.14	-0.18	0.25	0.98
True-post	0.16	0.02	0.31	1.00

Table 4: Average results over 100 independent observed datasets for the Gaussian example. The posteriors compared are the original approximate posterior (Approx-post), adjusted posteriors (Adjust-post) with α clipping, and the true posterior (True-post). [1]Achieved coverage with 90% target.

with $n = 15,000$. As in Bon et al. (2021), we impose stationarity conditions by transforming the polynomial coefficients of $\phi(z)$ and $\vartheta(z)$ to partial autocorrelations (Barndorff-Nielsen and Schou, 1973) taking values on $[-1, 1]^p$ and $[-1, 1]^q$ respectively, to which we assign a uniform prior. We apply an inverse hyperbolic tangent transform to map the ARMA parameters to the real line to facilitate posterior sampling. The fractional parameter d has bounds $(-0.5, 0.5)$, we sample over the transformed parameter $\tilde{d} = \tanh^{-1}(2d)$, and assume that $\tilde{d} \sim \mathrm{N}(0, 1)$ *a priori.*

The likelihood function of the ARFIMA model for large n is computationally intensive. As in, for example, Salomone et al. (2020) and Bon et al. (2021), we use the Whittle likelihood (Whittle, 1953) as the approximate likelihood to form the approximate posterior. Transforming both the data and the covariance function to the frequency domain enables us to construct the Whittle likelihood with these elements rather than using the time domain as inputs. The Fourier transform of the model's covariance function, or the spectral density $f_\theta(\omega)$, is

$$f_\theta(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \kappa(\tau, \theta) \exp(-i\omega\tau),$$

where the angular frequency $\omega \in (-\pi, \pi]$. Whereas the discrete Fourier transform (DFT) of the time series data is defined as

$$J(\omega_k) = \frac{1}{\sqrt{2\pi}} \sum_{t=1}^{n} X_t \exp(-i\omega_k t), \quad \omega_k = \frac{2\pi k}{n},$$

using the Fourier frequencies $\{\omega_k : k = -\lceil n/2 \rceil + 1, \ldots, \lfloor n/2 \rfloor\}$. Using the DFT we can calculate the periodogram, which is an estimate of the spectral density based on the data:

$$\mathcal{I}(\omega_k) = \frac{|J(\omega_k)|^2}{n}.$$

Then the Whittle log-likelihood (Whittle, 1953) can be defined as

$$\ell_{\text{whittle}}(\theta) = -\sum_{k=-\lceil n/2 \rceil+1}^{\lfloor n/2 \rfloor} \left(\log f_\theta(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_\theta(\omega_k)} \right).$$

In practice the summation over the Fourier frequencies, ω_k, need only be evaluated on around half of the values due to symmetry about $\omega_0 = 0$ and since $f_\theta(\omega_0) = 0$ for centred data.

The periodogram can be calculated in $\mathcal{O}(n \log n)$ time, and only needs to be calculated once per dataset. After dispersing this cost, the cost of each subsequent likelihood evaluation is $\mathcal{O}(n)$, compared to the usual likelihood cost for time series (with dense precision matrix) which is $\mathcal{O}(n^2)$.

Since this example is more computationally intensive, we do not repeat the whole process 100 times in our simulation study. Instead, we fix the observed data and base the repeated dataset results on the 100 calibration datasets (generated from the 100 calibration parameter values) that are produced in a single run of the process. However, we do not validate based on the datasets used in the calibration step, but rather generate 100 fresh datasets from the

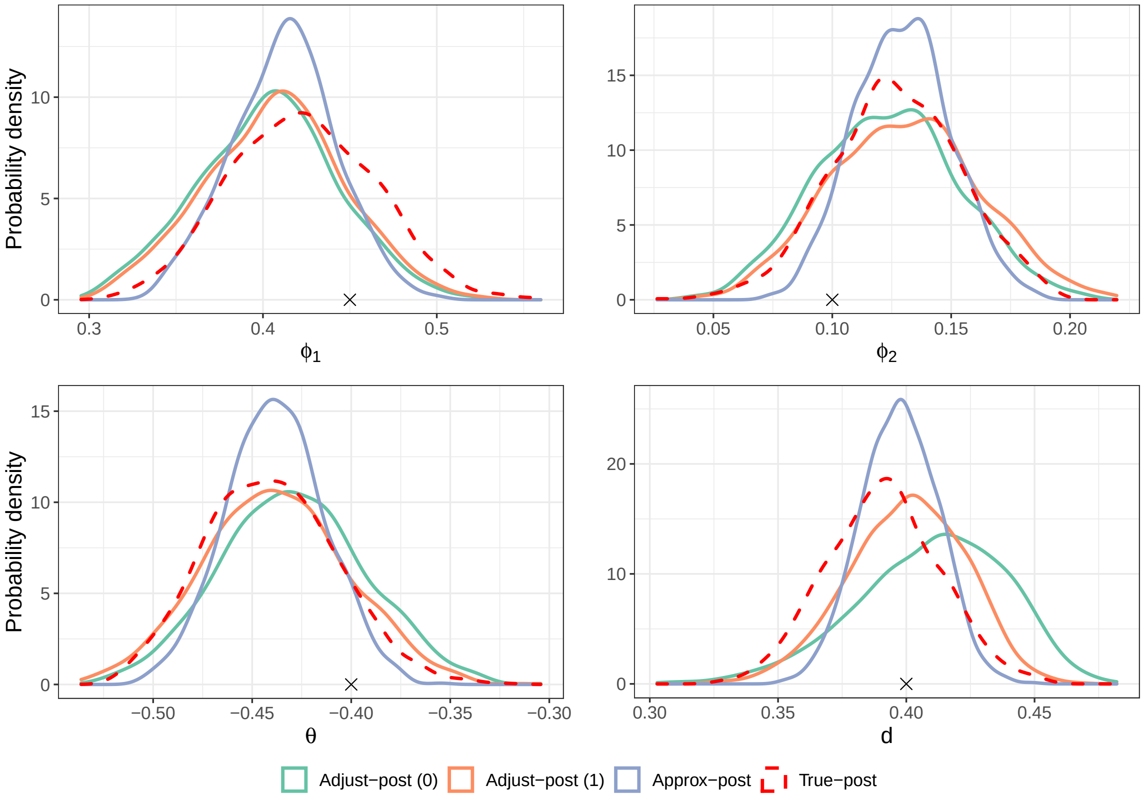


Figure 7: Estimated univariate posterior distributions for the Whittle likelihood example. Distributions shown are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping. The true posterior (True-post) is shown with a dashed line. The true generating parameter value is indicated with a cross (×).

calibration parameter values. As such, we do not provide a comparison to the true posterior in Table 5 as it is fixed and has a high computational cost to sample from. We consider univariate moment-correcting transformations, i.e. A in (8) is diagonal, since the covariance structure is well approximated by the Whittle likelihood in this example. We also choose the stabilizing function to be $v(\tilde{y}) = 1$ for simplicity.

To generate samples from the approximate and true posterior distributions we use a sequential Monte Carlo sampler (Del Moral et al., 2006). In particular, we use likelihood annealing with adaptive temperatures (Jasra et al., 2011; Beskos et al., 2016) and a Metropolis-Hastings mutation kernel with a multivariate Gaussian proposal. The covariance matrix is learned adaptively as in Chopin (2002). The simulation is coded in R (R Core Team, 2021).

The repeated run results for the parameters are shown in Table 5. It is evident that the Whittle approximation performs well in terms of estimating the location of the posterior, but the estimated posterior standard deviation is slightly too small, which leads to some undercoverage. The adjusted posteriors inflate the variance and obtain more accurate coverage of the calibration parameters. An example adjustment for the true dataset is shown in Figure 7, which shows that the adjustment inflates the approximate posterior variance. We also compute the calibration diagnostic to confirm the method is performing appropriately on the original data. Figure 8 shows that the achieved coverage is close to the target coverage, across a range of targets, hence the method is performing well.

Method	MSE	Bias	SD	AC[1]
ϕ_1				
Approx-post	0.004	0.003	0.031	67
Adjust-post (0)	0.005	−0.007	0.041	83
Adjust-post (0.5)	0.004	−0.003	0.041	83
Adjust-post (1)	0.005	−0.003	0.041	83
ϕ_2				
Approx-post	0.002	−0.001	0.021	74
Adjust-post (0)	0.002	−0.007	0.031	86
Adjust-post (0.5)	0.002	0.001	0.031	89
Adjust-post (1)	0.002	0.000	0.032	89
ϑ_1				
Approx-post	0.002	0.000	0.025	73
Adjust-post (0)	0.003	0.009	0.037	87
Adjust-post (0.5)	0.003	0.000	0.036	88
Adjust-post (1)	0.003	0.000	0.037	88
d				
Approx-post	0.001	−0.004	0.017	74
Adjust-post (0)	0.002	0.008	0.032	96
Adjust-post (0.5)	0.001	0.000	0.024	89
Adjust-post (1)	0.001	−0.001	0.025	90

Table 5: Average results for each parameter over 100 independent calibration datasets (fixed observation dataset) for the Whittle example. The posteriors compared are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with (α) clipping. [1]Achieved coverage with 90% target.

S.3 Lotka-Volterra model with ABC-like posterior

We also interested in testing how well our method can perform in a situation where there are no guarantees about the approximate model being employed. We test an alternative approximate posterior for the Lotka-Volterra SDE in Section 3.2 once[2] appearing in the `Turing.jl`

[2] See `https://github.com/TuringLang/Turing.jl/issues/2216` for discussion, the SDE example has now been removed from the tutorial but is available here `https://web.archive.org/web/20210419133415/https://turing.ml/dev/tutorials/10-bayesiandiffeq/`.

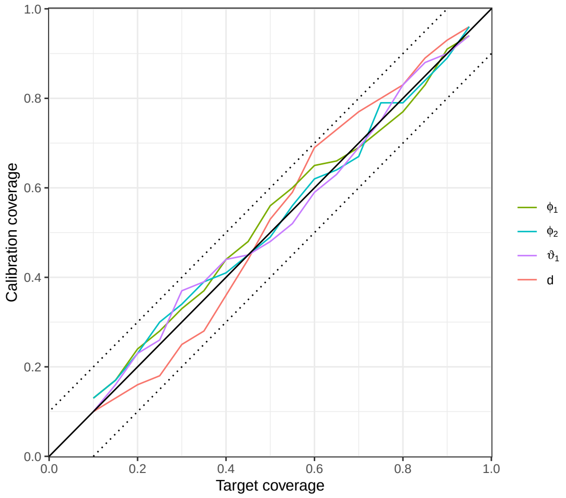


Figure 8: Calibration checks for all parameters in the Whittle likelihood example for $\alpha = 1$ with ± 0.1 deviation from parity shown with a dotted line.

tutorials (Ge et al., 2018). It used an unreferenced method for inference on the parameters of an SDE similar to an ABC posterior but with unknown kernel bandwidth. Despite its unknown inferential properties, we can correct the approximation using Bayesian score calibration and assess the correction using the calibration diagnostics. For the approximate model we use the noisy quasi-likelihood

$$l(\beta_{1:4}, \tau \mid x_{1:n}, y_{1:n}) = \tau^{2n} \exp\left(-\frac{\tau^2}{2}\sum_{i=1}^{n}\left[(x_i' - x_i)^2 + (y_i' - y_i)^2\right]\right),$$

where $\{(x_i', y_i')\}_{i=1}^n$ are simulated conditional on the $\beta_{1:4}$ using a rough approximation to the SDE (14). In particular, we use the Euler-Maruyama method with $\Delta t = 0.01$. For priors we use $\beta_i \overset{\text{iid}}{\sim} \text{U}(0.1, 5)$ for $i \in \{1, 2, 3, 4\}$ and $\tau \sim \text{Gamma}(2, 3)$. The quasi-likelihood is reminiscent of an approximate likelihood used in ABC. In particular, a Gaussian kernel is used to compare observed and simulated data, where τ plays a similar role to the tolerance in ABC. Usually in ABC, the tolerance is chosen to be small and fixed, but here we take it as random. However, as shown in Bortot et al. (2007), for example, using a random tolerance can enhance mixing of the MCMC chain. The resulting approximate posterior therefore has two levels of approximation; (i) an ABC-like posterior which uses (ii) a coarse approximate simulator rather than the true data generating process (which would require relatively more computation). For the calibration procedure we use $M = 200$ calibration datasets and use the unit weight approximation. We draw samples from the approximate posterior using the NUTS Hamiltonian Monte Carlo sampler with 0.25 target acceptance rate since the gradient is noisy. Finally, we use a multivariate moment-correcting transformation with dimension $d = 4$ and add a squared penalty term to all parameters of the lower-diagonal scaling matrix L. We shrink the diagonal elements of L to one, and off-diagonals elements to zero with rate $\lambda = 0.05$.

The results show that the adjusted posterior in this example does significantly better than the approximate posterior. Figure 9 shows the marginal posteriors have much

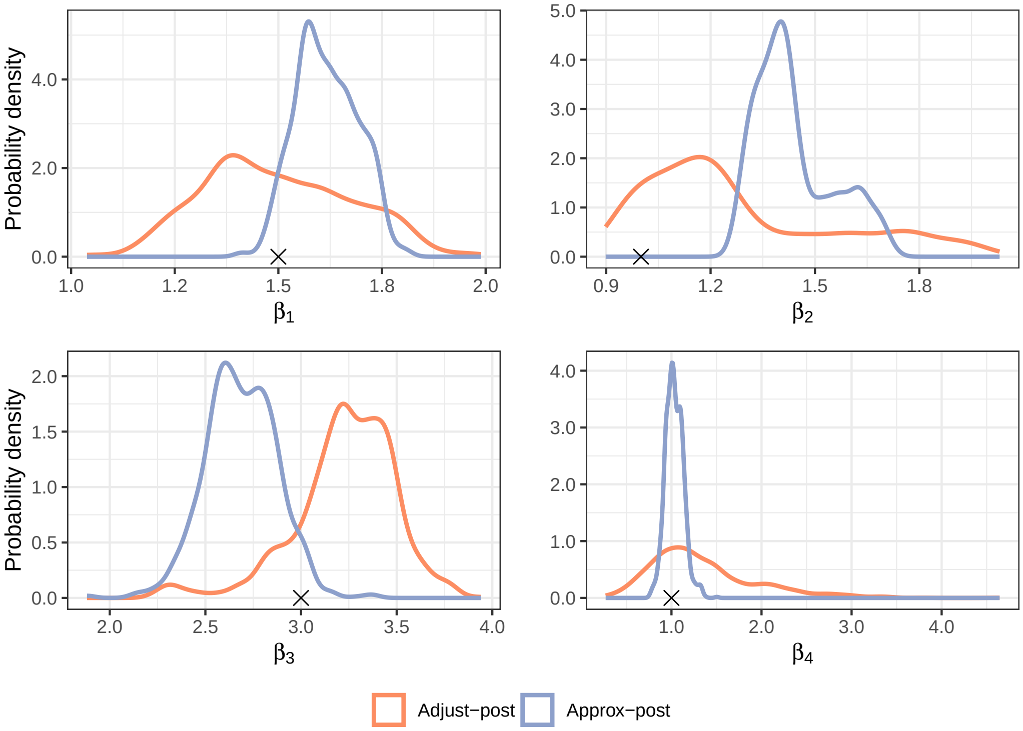


Figure 9: Estimated marginal posterior distributions for the Lotka-Volterra example. Distributions shown are the original approximate posterior (Approx-post) and adjusted posteriors (Adjust-post) with $\alpha = 1$ clipping. The true generating parameter value is indicated with a cross ($\times$).

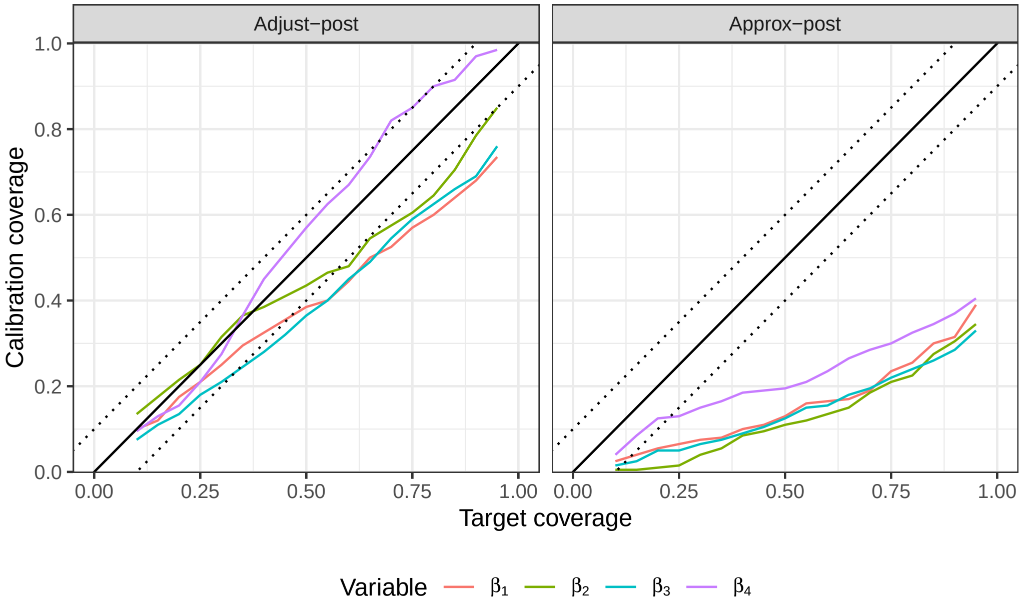


Figure 10: Calibration checks for all parameters in the Lotka-Volterra (ABC-like posterior) example for $\alpha = 1$ with ± 0.1 deviation from parity shown with a dotted line.

better coverage of the true parameter values and generally increase the variance of the approximate posterior to reflect that the approximation is too precise. Significant bias in parameter β_2 is also corrected. The target 90% coverage estimate from the calibration diagnostic was $(0.32, 0.31, 0.29, 0.37)$ for the approximate posterior, and $(0.68, 0.79, 0.69, 0.97)$ for the adjusted posterior for $(\beta_1, \beta_2, \beta_3, \beta_4)$ respectively.

Despite overall positive results, the calibration diagnostic shows that caution is required when using this adjusted approximate posterior. Though we see the adjusted posterior has significantly better achieved coverage than the approximate posterior in Figure 10, it still lags the nominal target coverage for parameters β_1 and β_3. Hence a conservative approach should be taken to constructing credible regions for those parameters. The calibration diagnostic is a useful byproduct of our method, alerting users to potential problems.

We also note a potential over-correction in the bias of β_3 seen in Figure 9. Investigating this further we find that the calibration samples that were simulated were in the range of $\beta_3 \in [1.5, 2.9]$ which excluded the true parameter value. Hence the transformation learned may have been skewed to this region. Increasing the scale of the importance distribution that generates the calibration samples may help this. Using a more sophisticated transformation or performing the calibration sequentially, as in Pacchiardi and Dutta (2022), may also be possible in future work. The ranges covered by the importance distribution for the other parameters were $\beta_1 \in [1.5, 2.4], \beta_2 \in [0.9, 2.7]$ and $\beta_4 \in [0.6, 1.1]$. A different importance distribution should be used if one suspects these ranges do not adequately cover the likely values of the true posterior.