Calibrated Approximate Bayesian Inference

Hanwen Xing [1] **Geoff K. Nicholls** [1] **Jeong Eun Lee** [2]

Abstract

We give a computational framework for estimating the bias in coverage resulting from making approximations in Bayesian inference. Coverage is the probability credible sets cover prior parameter values. We show how to estimate the coverage an approximation scheme achieves when the ideal but intractable observation model and the prior can be simulated, but have been replaced, in the Monte Carlo, with approximations. Coverage estimation procedures given in Lee et al. (2018) work on simple problems, but do not scale well, as those authors note. For example, Lee et al. (2018) calibrate a completely collapsed MCMC algorithm for partition structure in a Dirichlet process model for random effects in a hierarchical model and a small data set, but they note it fails when the model is applied to clustering on a larger dataset. By exploiting the symmetry of the coverage error under permutation of low level group labels and smoothing with Bayesian Additive Regression Trees, we show that the original approximate inference had poor coverage for these data and should not be trusted.

1. Introduction

Bayesian credible sets with stated nominal coverage are a fundamental way to communicate statistical uncertainty. However, we usually report approximate credible sets with uncalibrated coverage as some approximation is inevitable for large data sets and complex models. Approximation comes in many forms. In MCMC samples are only asymptotically distributed according to the posterior. The precision parameter is the run length. Approximate Bayesian Computation (ABC, Pritchard et al. (1999)) typically has two kinds of precision parameters, a distance threshold and a Monte Carlo sample. There are also "fixed" approximations with no precision parameters, in which a likelihood evaluation is replaced by an approximation which cannot be improved by varying a control parameter. Pseudo-likelihood (Besag, 1975) and variational inference (Jordan et al., 1999; Hoffman et al., 2013) often lead to a fixed approximation.

A number of methods have been developed to check the approximation is acceptable. Recent new generic diagnostic tools given in Talts et al. (2018) and Yao et al. (2018) are related to earlier work in Prangle et al. (2014) and exploit an idea, developed in Geweke (2004); Cook et al. (2006) as a MCMC convergence diagnostic, and going back to Monahan & Boos (1992). In early related work, Menendez et al. (2014) gives procedures for correcting credible sets under conditions stronger than those required here and Rodrigues et al. (2018) recalibrates ABC samples.

We consider an approximate Bayesian credible set with given nominal level α. What coverage does the credible set actually achieve? Wherever approximate Bayesian inference reports a credible set, an associated coverage measure should be given. We do not build a new credible set with improved coverage, although this is easy in our AIS method, because we would then have to estimate the coverage of that corrected credible set.

Let $\pi(\phi)$ be the prior for $\phi \in \Omega$, let $p(y|\phi)$ be the observation model (the likelihood) for data $y \in \mathcal{Y}$ and let $\pi(\phi|y) \propto \pi(\phi)p(y|\phi)$ be the posterior for ϕ given data y. Let $\tilde{\pi}(\theta)$ and $\tilde{p}(y|\theta)$ be the approximate prior and likelihood for parameter $\theta \in \Omega$ with approximate posterior $\tilde{\pi}(\theta|y) \propto \tilde{\pi}(\theta)\tilde{p}(y|\theta)$. This paper is motivated by problems where we cannot in practice sample $\pi(\phi|y)$ using any known Monte Carlo method. We assume a tractable approximation $\tilde{\pi}(\theta|y)$ is available, and we assume it is possible to sample $\phi \sim \pi(\cdot)$ and $y' \sim p(\cdot|\phi)$ (just as in ABC).

The estimated credible set is computed for a posterior distribution $\tilde{\pi}(\theta|y)$ which approximates the exact posterior $\pi(\phi|y)$. The exact level α credible set C_y for the exact posterior $\pi(\phi|y)$ satisfies

$$\alpha = \int_{\Omega} \mathbb{1}_{\phi \in C_y} \pi(\phi|y) d\phi.$$

This set C_y has perfect Bayes coverage in the sense that, if $\phi \sim \pi(\cdot)$ is a draw from the prior, and $y \sim p(\cdot|\phi)$ is a draw

[1] Department of Statistics, University of Oxford, UK [2] Department of Statistics, University of Auckland, New Zealand. Correspondence to: Hanwen Xing <hanwen.xing@stx.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

from the observation model, then $\Pr(\phi \in C_Y | Y = y) = \alpha$. The credible set covers the true parameter ϕ with probability α if nature drew ϕ from the prior, and the data y really was generated using the observation model we are using. This is the definition of Bayesian coverage, not an assumption.

In practice we compute a credible set $\tilde{C}_y$ using the approximate posterior $\tilde{\pi}(\theta|y)$. This is a set $\tilde{C}_y$ satisfying

$$\alpha = \int_{\Omega} \mathbb{1}_{\theta \in \tilde{C}_y} \tilde{\pi}(\theta|y) d\theta.$$

This will not in general have the right coverage for the exact posterior. If $\phi \sim \pi(\cdot)$, and $y \sim p(\cdot|\phi)$, and we let $b(y) = \Pr(\phi \in \tilde{C}_Y | Y = y)$, then

$$b(y) = \int_{\Omega} \mathbb{1}_{\phi \in \tilde{C}_y} \pi(\phi|y) d\phi \tag{1}$$

is the operational coverage $\tilde{C}_Y$ achieves for ϕ a draw from the exact posterior. This is not equal α in general. The coverage bias $b(y) - \alpha$ can vary markedly over data space. Cook et al. (2006) observe that coverage may be estimated, but the quantity they mention is equivalent to $\int_{\mathcal{Y}} p(y)b(y)dy$, an average over data space which may differ a great deal from $b(y)$, as we see in the example in Section 4.

In practice we may not be able to compute an exact credible set, even after making the approximation leading to $\tilde{\pi}(\theta|y)$. In this case we would typically simulate $\theta_j \sim \tilde{\pi}(\cdot|y)$ for $j = 1, \dots, J$, set $\underline{\theta} = (\theta_1, \dots, \theta_J)$ and compute an estimate, $\hat{C}_y(\underline{\theta})$, for $\tilde{C}_y$ based on J samples. There is an additional Monte Carlo error in the coverage and so, with ϕ, y and θ distributed as prior, observation model and approximate posterior, we let

$$c(y) = \Pr(\phi \in \hat{C}_Y(\underline{\theta}) | Y = y),$$

denote the realised coverage allowing for Monte Carlo error. We give algorithms for estimation of $b(y_{obs})$ and $c(y_{obs})$. We discuss the function $c(y)$ by default, as estimation of $b(y)$ is a simpler special case.

The joint distribution of ϕ, y and $\underline{\theta}$ in the generative model is given by

$$m(\phi, y, \theta) = \pi(\phi) p(y|\phi) \tilde{\pi}(\underline{\theta}|y). \tag{2}$$

The conditional distribution of ϕ, θ given y is

$$m(\phi, \theta | y) = \pi(\phi|y) \tilde{\pi}(\underline{\theta}|y),$$

writing $\tilde{\pi}(\underline{\theta}|y)$ for the joint distribution of $\underline{\theta} \in \Omega^J$ (an abuse of notation). Now $c(y)$ is an expectation over m,

$$c(y) = \int_{\Omega^J} \int_{\Omega} \mathbb{1}_{\phi \in \hat{C}_Y(\underline{\theta})} m(\phi, \underline{\theta}|y) d\phi d\theta.$$

The coverage is a posterior expectation in the exact distribution. Coverage estimation resembles the original problem of estimating credible sets from the exact posterior, which we have said we cannot do! However we can sample $\pi(\phi)$ and the observation model $p(y|\phi)$, and it proves easier to estimate $c(y)$ than some general expectation in the posterior.

If it were possible to simulate $\phi \sim \pi(\cdot|y)$ then estimation of $b(y)$ and $c(y)$ would be straightforward. For $i = 1, \dots, M$, we simulate $\phi_{(i)} \sim \pi(\cdot|y)$ and $\theta_{i,j} \sim \tilde{\pi}(\cdot|y)$ for $j = 1, \dots, J$. We then construct an α level approximate credible set $\hat{C}_{(i)}$ based on $\theta_{i,1}, \dots, \theta_{i,J}$, and define

$$c_i = \mathbb{1}(\phi_i \in \hat{C}_i).$$

Now $\hat{c}(y) = \frac{1}{M}\sum_{i=1}^{M} c_i$ is an unbiased and consistent estimator for $c(y)$. See Algorithm 1. If we replace $\hat{C}_i$ by $\tilde{C}_i$ then the procedure estimates $b(y)$. *We assume that Algorithm 1 cannot be implemented.* In the examples we give in Section 4 we implement Algorithm 1 to assist understanding. This will not in general be possible.

Algorithm 1 Estimation of operational coverage $c(y)$

```
Input: Observed data $y$; Exact posterior distribution $\pi(\phi|y)$; Approximate posterior distribution $\tilde{\pi}(\theta|y)$; Number of samples $J$ from the approximate posterior; Number of samples $M$ from the generative model.
for $i$ in $1, \dots, M$ do
    Simulate $\phi_i \sim \pi(\phi|y)$ and $\underline{\theta}_{(i)} = \{\theta_{i,1}, \dots, \theta_{i,J}\}$ where $\theta_{i,j} \sim \tilde{\pi}(\theta|y)$ for $j = 1, \dots, J$
    Compute the approximate credible set $\hat{C}_i$ based on $\underline{\theta}_{(i)}$. Set $c_i = \mathbb{1}(\phi_i \in \hat{C}_i)$
end for
Return: Estimated coverage $\hat{c}(y) = \frac{1}{M}\sum_{i=1}^{M} c_i$
```

The paper is structured as follows. In Section 2 we discuss previous work. We then outline two algorithms for the estimation problem in Section 3. In Section 4 we apply the algorithms to calibrate a pseudo-likelihood approximation. In Section 5, we calibrate the approximate posterior of a random partition in a hierarchical model with a Dirichlet-Process prior on the distribution of random effects. We finish with a brief discussion.

2. Relation to Previous Work

Lee et al. (2018) introduce the idea of estimating coverage probabilities as a validation procedure, and give two proof-of-concept estimators for $c(y)$ when simulation from the exact posterior $\pi(\theta|y)$ is not possible. Our own algorithms are a qualitative improvement, as those earlier methods completely fail on the examples we give in this paper.

Lee et al. (2018) give an importance sampling procedure which targets an approximation to $c(y)$. Let $\delta(x, y)$ be a generic distance function on the data space $\mathcal{Y}$, $\rho > 0$ a tolerance level and let $\Delta(y) = \{y' : \delta(y, y') \le \rho\}$ be a closed ball in $\mathcal{Y}$ centered at y. Let

$$d(y) = \Pr(\phi \in \hat{C}_Y(\underline{\theta}) | Y \in \Delta(y))$$

be an ABC-style approximation to $c(y)$. In terms of the density $m(\phi, y', \theta)$ in Equation 2,

$$d(y) = \int_{\Omega^J}\int_{\Omega}\int_{\Delta(y)} \frac{I_{\phi\in\hat{C}(\underline{\theta})}}{z(y,\rho)} m(\phi, y', \underline{\theta}) d\phi d\underline{\theta} dy' \quad (3)$$

with $z(y, \rho) = \Pr(Y \in \Delta(y))$ a normalising constant. Lee et al. (2018) estimate $d(y_{obs})$ with $\hat{d} = \sum_{i=1}^{M} w(\phi_i, y_i, \underline{\theta}_i)$ using M samples $(\phi_i, y_i, \underline{\theta}_i), i = 1, \ldots, M$ from the importance sampling distribution

$$\tilde{m}(\phi, y, \underline{\theta}) \propto \tilde{\pi}(\phi|y_{obs}) p(y|\phi) \tilde{\pi}(\underline{\theta}|y) \mathbb{1}_{y\in\Delta(y_{obs})},$$

and weights $w(\phi, y, \underline{\theta}) \propto m(\phi, y, \underline{\theta})/\tilde{m}(\phi, y, \underline{\theta})$. They use $\hat{d}$ as an estimate of $c(y_{obs})$. This approach has two drawbacks: $\hat{d} \to d$ as $M \to \infty$, however $d \ne c$ in general so the method is asymptotically biased (in M) unless we additionally take ρ to zero. This would be impractical as we simulate no data $y \in \Delta(y_{obs})$. Secondly, as the authors observe, this estimator can be unstable due to high weight variance. Our Annealed Importance Sampling (AIS) estimate is also asymptotically biased, but AIS is a much more powerful tool, and the bias can be made very small. Also AIS gives a much higher Effective Sample Size (ESS) at similar cost. We give examples in the Supplementary Material illustrating this point for a simple normal example. The issue is qualitative. When ESS estimates become small, they cannot be trusted (for example, a poorly estimated ESS may not increase when the sample size increases).

We also take advantage of a simple but important simplification not exploited by Lee et al. (2018). We replace $I_{\phi\in\hat{C}_Y(\underline{\theta})}$ in Equation 3 with $I_{\phi\in\hat{C}_{y_{obs}}(\underline{\theta})}$. As $Y \to y_{obs}$, the distribution of ϕ changes in a complex way, but the limit of $\hat{C}_Y(\underline{\theta})$ as $Y \to y_{obs}$ is computed from the approximate posterior, so we simply substitute the limiting value. This avoids the need to simulate $\underline{\theta}$ for each simulated y-value, a big timesaver in some settings. For comparison with the IS method in Lee et al. (2018), we take their Ising Model example and scale it up by a factor of 25. On this larger problem we find AIS calibration far out-performs Importance Sampling, yielding an ESS some 10 times larger in a comparable run time. In the Supplementary Material we show the ESS for IS falls off to small values as the Ising image size increases.

Lee et al. (2018) suggest an alternative regression procedure for estimating $c(y)$. If we simulate $(\phi_i, y_i, \underline{\theta}_{(i)}) \sim m$ iid for $i = 1, \ldots, M$ then, at each y, we have $(\phi_i, \underline{\theta}_{(i)} | y_i) \sim \pi(\phi_i|y_i)\tilde{\pi}(\underline{\theta}_{(i)}|y_i)$. It follows that if we compute $\hat{C}_i = \hat{C}(\underline{\theta}_{(i)})$ and a "response" $c_i = \mathbb{1}_{\phi_i\in\hat{C}_i}$ then the pairs c_i, y_i are measurements of a Bernoulli process in which $c_i \sim \text{Bernoulli}(c(y_i)), i = 1, \ldots, M$. Lee et al. (2018) suggest using a Generalised Additive Model (GAM) for logistic regression in order to estimate $c(y)$. Those authors take as regression covariates the components of a p-dimensional summary statistic $s(y) = (s_1(y), \ldots, s_p(y))$ with $p \ll \dim(\mathcal{Y})$. This works when $s(y)$ is sufficient, as is the case in our Ising model example in Section 3. However the choice of summary statistics is not in general clear and may bias results. However this is the sort of problem, variable selection, in which random forest regression and Bayesian Additive Regression Trees are very effective. For comparison with previous work, we take an example where the methods of Lee et al. (2018) fail, and show that estimation via BART gives reproducible results.

We exploit a symmetry of the approximation that will hold more widely: the approximation error is invariant under permutation of labels of levels of categorical variables. Let $y \in R^n$ be a data vector and let $\mathcal{L} = \{\ell_1, \ldots, \ell_N\}$ be the levels of a variable $x \in \mathcal{L}^n$. Suppose our data are simply y, x. Let $T(x) = \{T_1, \ldots, T_N\}$ be a collection of subsets of $\{1, \ldots, n\}$ partitioning the indices of x into groups of observations in the same level, so that $i \in T_j \Leftrightarrow x_i = \ell_j$ for each $i = 1, \ldots, n$ and each $j = 1, \ldots, N$. Let $\mathcal{P}_R = \{\sigma \in \mathcal{P} : T(x_\sigma) = T(x)\}$ be the set of permutations corresponding to level-relabeling. If the approximation does not distinguish levels, we have $c(y) = c(y_\sigma)$ for each $\sigma \in \mathcal{P}_R$ (permuting $y's$ with $x's$ fixed gives the same data set). This can be extended to multiple categorical variables. In a complete balanced design every level of every covariate co-occurs with every level of every other covariate the same number of times. Levels are then exchangeable within each variable simultaneously. In this setting we can compute $c(y)$ by computing it on one special quadrant $\mathcal{Y}_0$ of $\mathcal{Y}$ and then mapping identical copies of $c(y)$ out over $\mathcal{Y}$ by permutation. We take $\mathcal{Y}_0 = \{y \in \mathcal{Y} : \bar{y}_{T_1} \le \bar{y}_{T_2} \le \ldots \bar{y}_{T_N}\}$ (*ie* ordered on the averages of y's associated with each level, with additional order constraints for each variable if there are multiple categorical variables). We simulate ϕ, y' and $\underline{\theta}$, map y' back into $\mathcal{Y}_0$ and regress over this smaller space where y-values are more dense and regression (using BART) is easier. We map the function $\hat{c}(y)$ back to the quadrant containing y_{obs}.

3. Estimating the Operational Coverage

3.1. A Weighted-Sample Estimate for Coverage

In this section we estimate $c(y_{obs})$ using Annealed Importance Sampling (AIS) (Neal, 2001) to approximately sample the true posterior $\pi(\phi|y_{obs})$. This leverages our ability to draw samples from the approximate posterior $\tilde{\pi}(\phi|y_{obs})$ as a starting point for the AIS iteration.

Let $\{\gamma_j\}_{j=1}^{N_{AIS}}$ and $\{\beta_j\}_{j=1}^{N_{AIS}}$ be increasing sequences with $\gamma_0 = 0$, $\gamma_N = 1$ and $\beta_0 = 0$. Define an initial distribution

$$p_0(\phi, y) = \tilde{\pi}(\phi|y_{obs}) \times p(y|\phi),$$

and, for $j = 1, \dots, N_{AIS}$, intermediate distributions

$$\begin{aligned} p_j(\phi, y) &\propto \pi(\phi)^{\gamma_j} \tilde{\pi}(\phi|y_{obs})^{1-\gamma_j} p(y|\phi) \exp\left(-\beta_j \delta(y, y_{obs})\right) \\ &\propto \pi(\phi) \tilde{p}(y_{obs}|\phi)^{1-\gamma_j} p(y|\phi) \exp\left(-\beta_j \delta(y, y_{obs})\right). \end{aligned}$$

If $\gamma_{N_{AIS}} = 1$, then $p_{N_{AIS}}(\phi, y)$ converges to $\pi(\phi)p(y_{obs}|\phi)$ as $\beta_{N_{AIS}} \to \infty$ and so $p_{N_{AIS}}(\phi) \to \pi(\phi|y_{obs})$, the true posterior. The approximate posterior $\tilde{\pi}(\phi|y_{obs})$ is a useful part of the initial distribution p_0, as it supports ϕ values for which typical synthetic data $y \sim p(y|\phi)$ are relatively close to the observed y_{obs} from the start.

An update scheme for ϕ and y generates transitions from $p_j(\phi, y)$ to $p_{j+1}(\phi, y)$. For each $j = 1, \dots, N_{AIS}$, let

$$Q_j((\phi, y), (\phi', y')) = q_j((\phi, y) \to (\phi', y'))\alpha_j((\phi, y) \to (\phi', y'))$$

be a transition kernel with proposal distribution $q_j(\{\phi_j, y_j\} \to \{\phi', y'\}) = f_j(\phi_j, \phi')p(y'|\phi')$ based on a simple local proposal distribution $f_j(\phi, \phi')$ for ϕ, and an acceptance probability

$$\begin{aligned} \alpha_j &= 1 \wedge \frac{p_j(\phi', y')q_j(\{\phi', y'\} \to \{\phi_j, y_j\})}{p_j(\phi_j, y_j)q_j(\{\phi_j, y_j\} \to \{\phi', y'\})} \\ &= 1 \wedge \frac{\pi(\phi')\tilde{p}(y_{obs}|\phi')^{1-\gamma_j} f_j(\phi', \phi_j)}{\pi(\phi_j)\tilde{p}(y_{obs}|\phi_j)^{1-\gamma_j} f_j(\phi_j, \phi')} \qquad (*) \\ &\qquad \times \frac{\exp(-\beta_j \delta(y', y_{obs}))}{\exp(-\beta_j \delta(y_j, y_{obs}))} \end{aligned}$$

which admits $p_j(\phi, y)$ as an invariant distribution.

Let $d_{N_{AIS}} = E_{p_{N_{AIS}}}(\mathbb{1}(\phi \in \hat{C}_{y_{obs}}))$. Let $\{w_k, \phi_k\}_{k=1}^K$ be a set of weighted samples generated by the AIS algorithm described above. These are AIS-weighted samples from $p_{N_{AIS}}(\phi, y)$, so that

$$\hat{c}(y_{obs}) = \sum_{k=1}^{K} w_i \mathbb{1}(\phi_k \in \hat{C}_{y_{obs}})$$

is a consistent estimate for $d_{N_{AIS}}$ and $d_{N_{AIS}} \to c(y_{obs})$ as $\beta_{N_{AIS}} \to \infty$.

Theorem 1 *$\hat{c}(y_{obs})$ is a consistent estimator of the true realized coverage $c(y_{obs})$ as $K \to \infty$ and $\beta_{N_{AIS}} \to \infty$.*

Proof: Since $\hat{c}(y_{obs})$ is a self-normalised importance sampling estimator for the quantity

$$\Pr_{p_N}(\phi \in \hat{C}_{y_{obs}}) = \int \mathbb{1}(\phi \in \hat{C}_{y_{obs}}) p_N(\phi) d\phi,$$

where $p_N(\phi)$ is the marginal distribution of $p_N(\phi, y)$, we have that as $K \to \infty$,

$$\hat{c}(y_{obs}) \xrightarrow{p} \Pr_{p_N}(\phi \in \hat{C}_{y_{obs}}). \qquad (4)$$

As $\beta_{N_{AIS}} \to \infty$ and $\gamma_{N_{AIS}} = 1$ the target density $p_{N_{AIS}}(\phi)$ converges to the true posterior density $\pi(\phi|y_{obs})$. Hence by Scheffé's theorem, we must have

$$\Pr_{p_{N_{AIS}}}(\phi \in \hat{C}_{y_{obs}}) \longrightarrow c(y_{obs}), \qquad (5)$$

where $c(y) = \Pr(\phi \in \hat{C}_Y(\underline{\theta})|Y = y_{obs})$ is the true posterior probability. Combining (1) and (2), we conclude that $\hat{c}(y_{obs})$ is a consistent estimator for $c(y_{obs})$ □

Algorithm 2 summarizes the procedure. In contrast to the importance sampling approach in Lee et al. (2018), we do not need to simulate $\underline{\theta}_{(i)}$ and compute the approximate credible set for each synthetic data vector $y_i^{(j)}$ in Algorithm 2. This may speed up computation a great deal.

Algorithm 2 AIS Estimation of operational coverage $c(y)$

```
  Input: Observed data $y_{obs}$; Summary statistics $s : \mathcal{Y} \to \mathbb{R}^p$; Number of samples $J$ from the approximate posterior; Number of samples $M$ from the generative model.
  Simulate $\underline{\theta} = \{\theta_1, \dots, \theta_J\}$ where $\theta_1, \dots, \theta_J \sim \tilde{\pi}(\cdot|y_{obs})$; Compute the approximate credible set $\hat{C}_{y_{obs}}(\underline{\theta})$.
  for $i$ in $1, \dots, M$ do
      Sample $(\phi_i^{(1)}, y_i^{(1)}) \sim p_0(\phi, y)$.
      Compute $w_i^{(1)} \propto \frac{p_1(\phi_i^{(1)}, y_i^{(1)})}{p_0(\phi_i^{(1)}, y_i^{(1)})}$.
      for $j$ in $2, \dots, N_{AIS}$ do
          Sample $\phi'_i \sim f_{j-1}(\cdot|\phi_i^{(j-1)})$, $y'_i \sim p(y|\phi'_i)$.
          Set $\phi_i^{(j)} = \phi'_i$, $y_i^{(j)} = y'_i$ with probability $\alpha_j$
  defined in $(*)$ and set $\phi_i^{(j)} = \phi_i^{(j-1)}$, $y_i^{(j)} = y_i^{(j-1)}$
  otherwise.
          Compute $w_i^{(j)} \propto \frac{p_j(\phi_i^{(j)}, y_i^{(j)})}{p_{j-1}(\phi_i^{(j)}, y_i^{(j)})}$
      end for
      Compute $c_i = \mathbb{1}(\phi_i^{(N_{AIS})} \in \hat{C}_{y_{obs}})$
      Compute $w_i = \prod_{t=1}^{T} w_i^{(t)}$
  end for
  Compute $W_i = w_i / \sum_{i=1}^{M} w_i$
  Return: Estimated coverage $\hat{c}(y_{obs}) = \sum_{i=1}^{M} W_i c_i$
```

3.2. A Regression Estimate for Coverage

In Lee et al. (2018), the authors also suggest estimating $c(y)$ via regression. Let $\{\phi_i, y_i\}_{i=1}^M$ be samples from the generative model $\pi(\phi)p(y|\phi)$, let $\hat{C}_{y_i}$ be an approximate credible set for y_i, and $c_i = \mathbb{1}_{\phi_i \in \hat{C}_{y_i}}$. Conditional on y_i,

$$c_i \sim Bernoulli(c(y_i)), \quad c(y_i) = \Pr(\phi_i \in \hat{C}_{Y_i}|Y_i = y_i).$$

Let $s(y_i) \in \mathbb{R}^p$ be a vector of summary statistics of y_i. Lee et al. (2018) use a regression model with response c_i and covariates $s(y_i)$ to learn the map from y to $c(y)$. This is logistic regression, though Lee et al. (2018) suggest a more flexible GAM logistic regression. Raynal et al. (2018) and Marin et al. (2018) observe that, for ABC work, Random Forests allow us to handle a potentially large number of summary statistics (even if some or many of them are poorly informative) without preliminary selection. Inspired by their ideas, we applied a Probit Bayesian Additive Regression Tree (BART) model (Chipman et al., 2010) to estimate $c(y)$ when low-dimensional sufficient statistics are not available.

BART is a sum-of-trees model where each tree is regularized by a prior to be a weak learner. Let $Y \in \{0, 1\}$ be a generic binary output and $x \in \mathbb{R}^p$ be a generic input. In the classification setting we wish to infer an unknown function f such that $\Pr(Y = 1|x) = \Phi(f(x))$, with $\Phi(\cdot)$ the standard normal CDF. The Probit BART model approximates the function $f(x)$ with a sum of N_T trees, that is

$$f(x) \approx h(x) = \sum_{m=1}^{N_T} g_i(x),$$

where each $g_i(x)$ is given by a separate regression tree. A regularizing prior is imposed on the trees to keep individual tree effects small and prevent overfitting. The posterior distribution over trees is sampled using a Bayesian backfitting procedure. Details can be found in Chipman et al. (2010) and Kapelner & Bleich (2016).

Let $s(y)$ be a high dimensional vector of summary statistics for y. We fit a probit BART model with $\tilde{p}_i = \Phi(h(s(y_i)))$ and $c_i \sim \text{Bernoulli}(\tilde{p}_i)$ using $\{c_i, s(y_i)\}_{i=1}^{M}$ as the training data. For $s_{obs} = s(y_{obs})$, we obtain $\pi_B(h(s_{obs})|\{c_i, s(y_i)\}_{i=1}^{M})$, the posterior distribution of $h(s_{obs})$ in the fitted model, and estimate $c(y_{obs})$ using

$$\hat{c}(y_{obs}) = E_{\pi_B}(\Phi(h(s_{obs})|c_1, \ldots, c_N, s_1, \ldots, s_N)),$$

the (sample) posterior mean of $\Phi(h(s_{obs}))$.

We chose BART for two reasons. First, the tree structure is capable of handling potentially high dimensional input $s(y)$. This is crucial when low dimensional sufficient statistics for $y \in \mathcal{Y}$ are unavailable. Also, since BART is Bayesian, we have a natural way to assess the uncertainty of our estimate. We fit Probit BART models using the R package `bartMachine` (Kapelner & Bleich, 2016).

4. 2-D Ising Model

Figure 1 is a 200 by 200 binary image obtained by thresholding a grey-level image of ice floes from Banfield & Raftery (1992). We illustrate our method on the problem of fitting an Ising model with smoothing parameter ϕ and free boundary conditions to these data. The model with free boundary conditions has an intractable likelihood so we approximate it using a solveable model with toroidial boundary conditions. We then calibrate the approximate credible interval for ϕ. Lee et al. (2018) work on a 40 by 40 subset of the image. We apply the Importance Sampler of Lee et al. (2018) to the full problem in the supplement and find it breaks down at around 100 by 100. However AIS works well as we show.

Algorithm 3 Estimation of operational coverage $c(y)$ via regression

```
Input: Observed data $y_{obs}$; Summary statistics $s : \mathcal{Y} \to \mathbb{R}^p$; Number of samples $J$ from the approximate posterior; Number of samples $M$ from the generative model; A regression model $\mathcal{M}$.
for $i$ in $1, \ldots, M$ do
    Simulate $\phi_i \sim \pi(\phi)$, $y_i \sim p(y|\phi_i)$ and $\underline{\theta}_{(i)} = \{\theta_{i,1}, \ldots, \theta_{i,J}\}$ where $\theta_{i,j} \sim \tilde{\pi}(\theta|y)$ for $j = 1, \ldots, J$
    Compute the approximate credible set $\hat{C}_i$ based on $\underline{\theta}_{(i)}$. Set $c_i = \mathbb{1}(\phi_i \in \hat{C}_i)$ and compute the $p$-dimensional summary statistics $s(y_i)$.
end for
Fit the regression model $\mathcal{M} : c \sim h(s(y))$ to learn the relation between coverage $c(i)$ and summary statistics $s(y_i)$ using $\{c_i, s(y_i)\}_{i=1}^{M}$ as training data.
Return: $\hat{c}(y_{obs})$, the fitted value given $s(y_{obs})$ based on the regression model $\mathcal{M}$.
```

The Ising model is a Markov model on a binary lattice. Let $G = (V, E)$ be a graph with edge set E and vertices V. For each $v \in V$, let $y_v \in \{0, 1\}$ be binary data at v and $y = y_{v_{v \in V}}, y \in \{0, 1\}^{|V|}$ the collection of all y_v. Let $< u, v > \in E$ be an edge in G between vertices u, v. Let

$$f(y, E) = \sum_{<u,v> \in E} \mathbb{1}(y_u \neq y_v)$$

be the number of pairs of vertices with disagreeing neighbours on G. In this example, G is a rectangular $N_I \times N_I$ lattice with $N_I = 200$ and free boundary conditions. We denote the graph by $G_F = (E_F, V)$. Interior vertices on G have degree 4, edge vertices have degree 3 and corner vertices have degree 2. We consider also the toroidal boundary condition. The lattice $G_T = (E_T, V)$ is wrapped onto a torus so all vertices in G have degree 4.

Let $\phi > 0$ be a scalar parameter. The likelihood under free boundary conditions is

$$p_F(y|\theta) = Z_F(\theta)^{-1} \exp(-\theta f(y, E_F))$$

where

$$Z_F(\theta) = \sum_{x \in \{0,1\}^{|V|}} \exp(-\theta f(x, E_F))$$

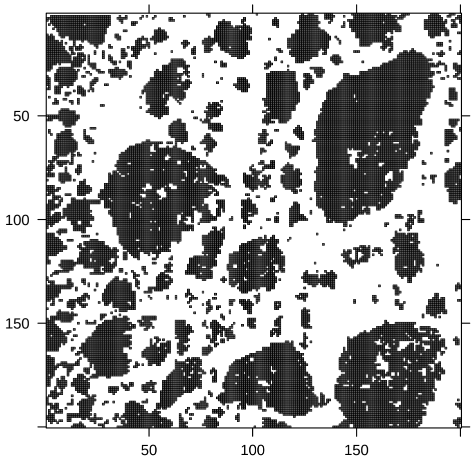

Figure 1. Ice floe image from Banfield & Raftery (1992)

is a normalizing constant. Similarly, let $Z_P(\theta)$ be the normalizing constant under toroidal boundary conditions. When N_I is large, $Z_F(\theta)$ is computationally intractable. However, $Z_P(\theta)$ is given in closed form in Kaufman (1949).

Following Lee et al. (2018) we impose a uniform prior $\pi(\theta) \propto \mathbb{1}(\theta \in (0,2))$ on θ. The posterior is

$$\pi(\theta|y) \propto Z_F(\theta)^{-1}\exp(-\theta f(y, E_F))\mathbb{1}(\theta \in (0,2)).$$

Although $\pi(\theta|y)$ is doubly intractable we can use an exchange algorithm (Murray et al., 2006) to get asymptotically exact inference. Alternatively we can approximate $Z_F(\theta)$ by $Z_P(\theta)$. Let

$$\tilde{\pi}(\theta|y) \propto Z_P(\theta)^{-1}\exp(-\theta f(y, E_F))\mathbb{1}(\theta \in (0,2))$$

denote the approximate posterior. The univariate approximate posterior density $\tilde{\pi}(\theta|y)$ can be evaluated pointwise up to a normalising constant, and the corresponding CDF is readily evaluated. We compute the equal-tail, 95% approximate credible set $\tilde{C}_y$. We find $\tilde{C}_y = (0.899, 0.913)$. We used an exchange algorithm (Murray et al., 2006) and Algorithm 1 to estimate the coverage $c(y_{obs})$ as a check. This Monte Carlo estimate, which we treat as the truth, is $c^{(1)}(y_{obs}) = 0.518$. Clearly we should be rejection this approximation scheme.

We run Algorithm 2 to estimate the operational coverage. We initialise the AIS sampler with $M = 1000$ samples $\{\phi_i, y_i\}_{i=1}^M$ with $\phi_i \sim \tilde{\pi}(\phi|y_{obs})$ and $y_i \sim p_F(y_i|\phi_i)$, the *true* likelihood. We set the number of AIS iterations $N_{AIS} = 60$ with cooling schedule $\beta_j = 1.05^j$ for $j = 1, \ldots, N_{AIS}$, $\gamma_j = 0.02j$ for $j = 1, \ldots 50$ and $\gamma_j = 1$ for $j = 51, \ldots, N_{AIS}$ at the jth iteration. We use the K-S distance $\delta(y_i, y_{obs}) = |G_{y_i} - G_{y_{obs}}|_\infty$, where G_{y_i} the CDF of the approximate posterior $\tilde{\pi}(\phi|y_i)$, as the distance metric. Algorithm 2 gives $\hat{c}(y_{obs}) = 0.545$ with standard error 0.037 and an effective sample size (ESS) equal 180. In Figure 2 we see the estimated operational coverage at iteration $N_{AIS} = 60$ is close to the true value, indicating the effectiveness of AIS here. We repeated the whole experiment 15 times with excellent consistency within uncertainty. The importance sampler (Lee et al., 2018) gives a similar estimate $\hat{c}_{IS}(y_{obs}) = 0.529$ with the threshold $\rho = 0.5$ (varying ρ gave no improvement). However, the ESS is just 22 for a similar runtime. This ESS is likely to be an overestimate. See the Supplementary material for details.

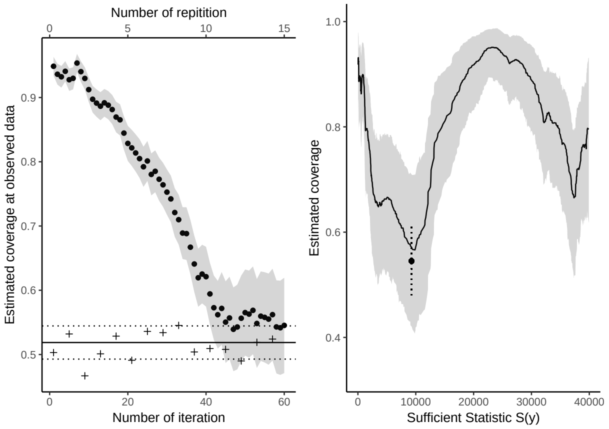

Figure 2. Left: Algorithm 2; the estimated $\hat{c}(y_{obs})$ based on the intermediate distribution p_j at each iteration j of the AIS sampler. The grey ribbon represents the 2σ error bar for the true value. We see that $\hat{c}(y_{obs})$ converges to $c^{(1)}(y_{obs})$ while the standard error of $\hat{c}(y_{obs})$ increases due to decreasing effective sample size. Crosses correspond to final results for 15 repeats of the algorithm (with arbitary x-values). Right: Algorithm 3; the estimated $c(y)$ as a function of the natural sufficient statistics $s(y)$. Vertical dotted segment is the 2σ error bar of $\hat{c}(y_{obs})$ based on Algorithm 2.

In Algorithm 3, we first simulate $M = 1000$ samples $\{\phi_i, y_i\}$ from $\pi(\phi)p_F(y|\phi)$, the joint prior distribution, for $i = 1, \ldots, M$. Simulation of synthetic data from $p_F(y|\phi)$ is straightforward using MCMC. Figure 3 shows a BART estimate with "covariate" $S(y_i) = f(y, E_F)$ and "response" $c_i = \mathbb{1}_{\phi_i \in \hat{C}_{y_i}}$. This gives $\hat{c}(y_{obs}) = 0.565$ with 95% credible set $(0.420, 0.702)$, in good agreement with the AIS estimate. This is an easy problem for the regression approaches as there is a scalar sufficient statistic. It is harder for importance sampling, due to the sharply peaked target. In the supplement we show that BART reproduces Figure 2 (right) using the raw $N_I^2 = 40000$ binary data.

5. Dirichlet Process Random Effect Model

Lee et al. (2018) show that their IS approach works for calibration of a completely collapsed MCMC algorithm for partition structure in a Dirichlet process. However it is easy to find problems on which the IS estimator performs poorly. In this section, we use Algorithm 3 to estimate the realised coverage $c(y_{obs})$ for a dataset and approximation procedure on which the methods of Lee et al. (2018) fail (no sufficient statistic, output ESS close to one, estimated $\hat{c}(y)$ values close to zero or one). We show that credible sets based on the Laplace approximation are unreliable in this example.

Our dataset has the classical format of a complete design, with five categorical variables, including `Treatment` ($N = 12$ levels), and four block variables, `B1` (with $q = 3$ levels) `B2` ($r = 2$ levels), `B3` (two levels) and `B4` (seven levels) so that we have $n = 1008$ observations, $y = (y_1, \ldots, y_n)$. In our example we fit a hierarchical model with known fixed effects `B1` $*$ `B2`. Let X be the $n \times p$ design matrix for the fixed effects ($p = 6$ here).

We take a Dirichlet process prior for the hierarchy of random effects in our hierarchical model. Our aim here is not to develop new models but to illustrate calibration. The model is similar in structure to the model Malsiner-Walli et al. (2018) fit, differing mainly in the choice of partition model, a Chinese Restaurant Process (CRP) in our case. In related work Pauger et al. (2018) cluster variances in the marginal model. We suppose the scientist wants to cluster the treatments into groups with similar interaction effects. Each treatment has a vector of random effects, so the object here is to estimate a partition of the $N = 12$ random effect vectors for treatment. The output of the uncalibrated analysis is an approximate HPD credible set for the unknown partition of treatment effects. We calibrate a credible *set* here, not simply a credible interval, reflecting the ease of application of our methods in more general settings.

Let $\mathcal{A} = \{1, \ldots, N\}$ give the distinct levels of `Treatment` and let A be the $n \times 1$ covariate vector with $\mathrm{A}_i \in \mathcal{A}$ giving the level of `Treatment` in the i'th observation. These are the levels we cluster. Let $S = \{S_1, \ldots, S_K\}$ be a partition of $\mathcal{A}$ and let S be a $n \times 1$ unobserved categorical covariate giving the grouping, so that $\mathrm{S}_i = k$ means $\mathrm{A}_i \in S_k$. The interaction between `B1` and `Treatment` is a random effect so we have a vector of random effects $\eta_j^A \in \mathbb{R}^q$, $\eta_j^A \sim N(0, \Sigma_A)$ iid for $j = 1, \ldots, N$ for the different levels of A and another offset vector of random effects $\eta_k^S \in R^q$, $\eta_k^S \sim N(0, \Sigma_S)$ i.i.d. for $k = 1, \ldots, K$. Let Z be a $n \times q$ matrix of indicators for the levels of `B1`. Denote by x_i, z_i the ith row of X and Z, let $\beta = (\beta_1, \ldots, \beta_p)$ be the vector of fixed effects, and let $\epsilon_i \sim N(0, \sigma^2)$ i.i.d. for $i = 1, \ldots, n$. The observation model is

$$y_i = x_i\beta + z_i\eta^S_{\mathrm{S}_i} + z_i\eta^A_{\mathrm{A}_i} + \epsilon_i, \quad i = 1, \ldots, n$$

with likelihood $p(y|\psi)$ for parameter $\psi = (\beta, \eta^A, \eta^S, \Sigma_A, \Sigma_S, \sigma^2), \psi \in \Omega_S$. The parameter space Ω_S of ψ depends on the partition S, as the dimension of η^S is determined by S.

The partition S is an unknown parameter in the posterior with a Chinese Restaurant Process (CRP) prior

$$\pi(S) = \frac{\alpha^K \Gamma(\alpha) \prod_{k=1}^K \Gamma(|S_k|)}{\Gamma(\alpha + n)}$$

for $S \in \mathcal{P}$, where α is a model parameter, $|S_k|$ is the number of elements in the set S_k and $S \in \mathcal{P}$ is the space of partitions. We took $\alpha = 1$. Our setup is equivalent to a Dirichlet process prior $G_S \sim \mathrm{DP}(\alpha, H)$, $\eta_k^S \sim G_S$ with base distribution $H(\eta_k^S) = N(\eta_k^S; 0, \Sigma_S)$ for $k = 1, \ldots, K$ for the random effects η^S due to partition S. The joint prior $\pi(\psi, S)$ is

$$\pi(\psi, S) = \pi(\eta^S|S, \Sigma_S)\pi(\beta, \eta^A, \Sigma_A, \Sigma_S, \sigma^2)\pi(S)$$

with $\pi(\eta^S|S, \Sigma_S) = \prod_{k=1}^K N(\eta_k^S; 0, \Sigma_S)$ and

$$\pi(\beta, \eta^A, \Sigma_A, \Sigma_S, \sigma^2) = N(\beta; 0, \sigma_\beta^2 I_p) \prod_{j=1}^N N(\eta_j^A; 0, \Sigma_A)$$
$$\times \mathrm{IW}(\Sigma_A; \nu_A, V_A)\mathrm{IW}(\Sigma_S; \nu_S, V_S)\mathrm{IG}(\sigma^2; \alpha_\sigma, \beta_\sigma).$$

Here $\sigma_\beta, \nu_A, V_A, \nu_S, V_S, \alpha_\sigma, \beta_\sigma$ are prior hyperparameters. The joint posterior distribution is then

$$\pi(\psi, S|y) \propto p(y|\psi)\pi(\psi, S). \tag{6}$$

Estimation of S by sampling the joint posterior $\pi(\psi, S|y)$ using MCMC is a variable dimension problem. It is convenient to work with the marginal posterior $\pi(S|y) \propto p(y|S)\pi(S)$, where

$$p(y|S) = \int p(y|\psi)\pi(\psi|S)d\psi \tag{7}$$

is the marginal likelihood. However, $p(y|S)$ is computationally intractable. Suppose we approximate it with a Laplace approximation. How much harm does this do? Let b_S be the Bayesian Information Criterion (BIC) of the model in Equation 7 if the partition is S so that

$$\tilde{p}_{BIC}(y|S) = \exp(-b_S/2)$$

approximates the marginal likelihood $p(y|S)$. This corresponds to a choice of unit information priors on model parameters ψ ((Kass & Raftery, 1995; Raftery, 1999)) and can be seen as part of the approximation we are calibrating. Packages computing the BIC for complex models are available (we use the R-package `lme4`, see Bates et al. (2014)). The approximate posterior for S is

$$\tilde{\pi}(S|y) \propto \tilde{p}_{BIC}(y|S)\pi(S).$$

We sample $\tilde{\pi}(S|y)$ and construct an approximate credible set for S (see Supplementary Material) using standard Metropolis-Hasting MCMC. Can we trust this credible set?

Table 1. Estimate of $c(y_{obs})$ and the corresponding 95% credible interval based on two Probit BART models.

Model	$\hat{c}(y_{obs})$	95% Credible Interval
$M1$	0.262	(0.065,0.549)
$M2$	0.308	(0.124,0.552)
$M2_1$	0.285	(0.067,0.564)
$M2_2$	0.322	(0.086,0.618)
$M2_3$	0.347	(0.095,0.673)
$M2_4$	0.270	(0.056,0.565)

We apply Algorithm 3 to the problem. For $i = 1, \dots, M$ we sample partitions $S^{(i)} \sim \pi(S)$, $\psi^{(i)} \sim \pi(\cdot|S^{(i)}), \psi \in \Omega_S$ and $y^{(i)} \sim p(\cdot|\psi^{(i)}), y^{(i)} \in \mathbb{R}^n$. Low dimensional summary statistics are not available, so we try two sets of high dimensional summary statistics. Covariates `B3` and `B4` do not appear in the model, so e average data with `Treatment`, `B1` and `B2` fixed. Denote by $\bar{y}^{(i)}_{jkl}$ the mean of observations $\{y_{i'} : A_{i'} = j, \mathtt{B2}_{i'} = k, \mathtt{B3}_{i'} = l\}$ and let

$$T(y^{(i)}) = \{\bar{y}^{(i)}_{jkl}\}, j = 1, \dots, N; k = 1, \dots, r; l = 1, \dots, q$$

denote these summary statistics, with $N = 12, r = 2$ and $q = 3$ and dimension $Nrq = 72$. Tree-based BART has no difficulty with summary statistics of this dimension.

As noted at the end of Section 2, level-labels are exchangeable so we can permute the data vectors $y^{(i)}, i = 1, \dots, M$ and map them into a "tighter" subregion of $\mathbb{R}^n$. Regression is easier on the more densely packed $y^{(i)}$-values. Let $\sigma \in \mathcal{P}_R$ be the set of relabeling permutations for which $c(y_\sigma) = c(y)$. In our setting with three categorical covariates `Treatment`, `B1` and `B2` with $N = 12$, $q = 3$ and $r = 2$ levels respectively, and a complete design, the number of "legal" permutations of the collapsed data T is $|\mathcal{P}_R| = N!q!r!$.

We now define a second coarser set of summary statistics. Consider the $N = 12$ treatment levels. For $i = 1, \dots, M$, let $H_N(y^{(i)}) = \{\bar{y}^{(i)}_j\}_{j=1}^N$, where $\bar{y}^{(i)}_j$ is the sample mean of $\{y^{(i)}_{i'} : i' \in 1 : N, \mathtt{A}_{i'} = j\}$. Take the permutation $\sigma \in \mathcal{P}_R$ such that $H_N(\sigma(y^{(i)})) = \{\bar{y}^{(i)}_{(1)}, \dots, \bar{y}^{(i)}_{(N)}\}$ matches the order statistics of $\{\bar{y}^{(i)}_j\}_{j=1}^N$. Let $H_r(y^{(i)})$ and $H_q(y^{(i)})$ give the corresponding sorted averages for `B2` and `B1`. Let

$$H(y^{(i)}) = (H_N(\sigma(y^{(i)})), H_r(y^{(i)}), H_q(y^{(i)}))$$

denote this collection of the $p = 17$ order statistics. We take $H(y)$ as a second set of summary statistics.

We simulate $M = 810$ pairs $\{c^{(i)}, y^{(i)}\}_{i=1}^M$ of training data following Algorithm 3. We fit two probit BART models $M1 : c^{(i)} \sim T(y^{(i)})$ and $M2 : c^{(i)} \sim H(y^{(i)})$ i.e. we fit two models using $T_i = T(y^{(i)})$ and $H_i = H(y^{(i)})$ as summary statistics. The estimated values $\hat{c}(y_{obs})$ at y_{obs} are given in Table 1. Estimates based on the different summary statistics $M1$ and $M2$ agree. In order to further test the robustness of our method, we partition the full synthetic dataset $\{c^{(i)}, y^{(i)}\}_{i=1}^M$ into four equal-size subsets and fit a Probit BART model using formula $M2$ to each subset. Let $M2_1, M2_2, M2_3$ and $M2_4$ label these models. Fitted values at y_{obs} are in line with fitted values determined on the full training set, with wider credible intervals as training sets are smaller. Our estimate of coverage is robust. The estimated coverage $\hat{c}(y_{obs})$ is far lower than the nominal level $\alpha = 0.95$, so the approximate marginal likelihood $\tilde{p}_{BIC}(y|S)$ is a poor approximation here, and the credible set should not be trusted.

6. Conclusion

In this paper we give a computational framework for estimating the bias in coverage due to approximations made in carrying out Bayesian inference. We provide estimators for the calibration problem defined in Lee et al. (2018). We demonstrate their effectiveness by diagnosing poor approximate coverage in two examples. The quality of the approximate coverage may depend on the data, so an approximation may work well for some data sets and not others.

Our assumptions are similar to those of ABC (we can simulate the prior and observation model). A vanilla application of ABC would give a natural though in general inefficient estimator for $b(y)$ in Equation 1. We have help from the approximate posterior $\tilde{\pi}$ and so our AIS method for estimating coverage can be seen as a hybrid of the two approximation schemes. Our BART based regression uses the same simulation stage as ABC, but the regression is used to estimate a probability function over data space, in a manner similar to the model selection procedure in Pudlo et al. (2016).

The two coverage estimators we suggest, AIS and BART, have complimentary strengths. AIS uses local information without variable selection. Regression with BART can more easily exploit global structure (such as the symmetry we found in $c(y)$) and does not require careful specification of summary statistics or related distance measures. Finally we note that what we are offering is a consistency check: a good outcome (ie an estimated coverage close to α) is a necessary but not a sufficient condition for us to trust the original estimated credible set.

References

Banfield, J. D. and Raftery, A. E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16, 1992.

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. lme4:

Linear mixed-effects models using Eigen and S4. *R package version*, 1(7):1–23, 2014.

Besag, J. Statistical analysis of non-lattice data. *The Statistician*, pp. 179–195, 1975.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Cook, S. R., Gelman, A., and Rubin, D. B. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15 (3):675–692, 2006.

Geweke, J. Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467):799–804, 2004.

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

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kapelner, A. and Bleich, J. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(i04), 2016.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Kaufman, B. Crystal statistics. II. Partition function evaluated by spinor analysis. *Physical Review*, 76(8):1232, 1949.

Lee, J. E., Nicholls, G. K., and Ryder, R. J. Calibration procedures for approximate Bayesian credible sets. *arXiv preprint arXiv:1810.06433*, 2018.

Malsiner-Walli, G., Pauger, D., and Wagner, H. Effect fusion using model-based clustering. *Statistical Modelling*, 18 (2):175–196, 2018.

Marin, J.-M., Pudlo, P., Estoup, A., and Robert, C. Likelihood-free model choice. *Handbook of Approximate Bayesian Computation*, pp. 153, 2018.

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

Monahan, J. F. and Boos, D. D. Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278, 1992.

Murray, I., Ghahramani, Z., and MacKay, D. J. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 359–366. AUAI Press, 2006.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Pauger, D., Wagner, H., et al. Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, 2018.

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

Pritchard, J. K., T, S. M., Perez-Lezaun, A., and Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. Reliable ABC model choice via random forests. *Biometrika*, 32(6):859–866, 2016.

Raftery, A. E. Bayes factors and BIC: Comment on A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3):411–427, 1999.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C., and Estoup, A. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 2018.

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

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 5581–5590, 2018.