Bayesian Analysis (2019) **14**, Number 4, pp. 1245–1269

Calibration Procedures for Approximate Bayesian Credible Sets

Jeong Eun Lee*, Geoff K. Nicholls†,§, and Robin J. Ryder‡

Abstract. We develop and apply two calibration procedures for checking the coverage of approximate Bayesian credible sets, including intervals estimated using Monte Carlo methods. The user has an ideal prior and likelihood, but generates a credible set for an approximate posterior based on some approximate prior and likelihood. We estimate the realised posterior coverage achieved by the approximate credible set. This is the coverage of the unknown "true" parameter if the data are a realisation of the user's ideal observation model conditioned on the parameter, and the parameter is a draw from the user's ideal prior. In one approach we estimate the posterior coverage at the data by making a semi-parametric logistic regression of binary coverage outcomes on simulated data against summary statistics evaluated on simulated data. In another we use Importance Sampling from the approximate posterior, windowing simulated data to fall close to the observed data. We illustrate our methods on four examples.

Keywords: Monte Carlo, approximation, calibration, credible intervals.

MSC 2010 subject classifications: 65C05, 68W25, 62F15.

1 Introduction

When we carry out Bayesian inference it is often convenient, even when not strictly necessary, to make approximations. We work with likelihoods and priors which only approximately equal those we would ideally use. Examples of popular approximations include Approximate Bayesian Computation (Pritchard et al., 1999; Marin et al., 2012), pseudo-likelihoods (Besag, 1975), synthetic likelihood (Wood, 2010), Variational Bayes (Jordan et al., 1999), and Expectation Propagation (Minka, 2001). If we use an approximate posterior distribution to get an approximate credible set with nominal level α, we should expect the approximation to distort the coverage, so that the operational coverage of our approximate credible set is not α. In this paper we give a procedure for measuring the operational coverage. We ignore questions of goodness of fit. We are not aiming to calibrate coverage of the true parameter, but to measure the distortion in coverage due to target approximation.

Our approach was inspired by Geweke (2004) and Cook et al. (2006), and later related papers including Fearnhead and Prangle (2012), Prangle et al. (2014), and Yao

*The Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, kate.lee@auckland.ac.nz

†Department of Statistics, 24-29 St Giles, Oxford, OX1 3LG, UK, nicholls@stats.ox.ac.uk

‡CEREMADE, CNRS, Université Paris-Dauphine, PSL University, 75016 Paris, France, ryder@ceremade.dauphine.fr

§Corresponding author.

 https://doi.org/10.1214/19-BA1175

et al. (2018) which exploit an idea set out by Monahan and Boos (1992). Let $\pi(\phi)$ be a prior for a parameter $\phi \in \Omega$, let $p(y|\phi)$ be an observation model for data $y \in \mathcal{Y}$ with

$$\pi(\phi|y) \propto \pi(\phi)p(y|\phi)$$

the posterior for ϕ given y. Let $\tilde{\pi}(\theta)$ and $\tilde{p}(y|\theta)$ be an approximate prior and approximate likelihood for a parameter $\theta \in \Omega$ with

$$\tilde{\pi}(\theta|y) \propto \tilde{\pi}(\theta)\tilde{p}(y|\theta)$$

the approximate posterior for θ given y. Suppose we simulate $\phi \sim \pi(\cdot)$, $y' \sim p(\cdot|\phi)$ and $\theta \sim \tilde{\pi}(\cdot|y')$. The joint conditional distribution of ϕ and θ, $m(\theta, \phi|y')$ say, is

$$m(\theta, \phi|y') = \pi(\phi|y')\tilde{\pi}(\theta|y'), \tag{1.1}$$

so, conditional on y', ϕ and θ are exchangeable if and only if there is no approximation and $\tilde{\pi}(\theta|y') = \pi(\theta|y')$ for all $\theta \in \Omega$. The joint marginal distribution $m(\theta, \phi)$ is

$$m(\theta, \phi) = \int \pi(\phi|y')\tilde{\pi}(\theta|y')p(y')\, dy', \tag{1.2}$$

where $p(y') = \int_\Omega \pi(\phi)p(y'|\phi)\, d\phi$ is the exact marginal likelihood, so ϕ and θ are marginally exchangeable if there is no approximation.

In work to date, Equation (1.2) has been taken as a starting point, as it gives a necessary condition on correctly distributed samples θ, which can be tested to check an approximation is good. For example, for $i = 1, \ldots, M$, simulate $\phi_{(i)} \sim \pi(\cdot)$, $y_{(i)} \sim p(\cdot|\phi_{(i)})$ and $\theta_{(i)} \sim \tilde{\pi}(\cdot|y_{(i)})$; here $y_{(i)} \in \mathcal{Y}$ is a data set, $\phi_{(i)}, \theta_{(i)} \in \Omega$ are parameter vectors, and the realisation $\theta_{(i)} \sim \tilde{\pi}(\cdot|y_{(i)})$ might for example be the last sample in a Markov Chain Monte Carlo (MCMC) run. The parameter vectors $\phi_{(1)}, \ldots, \phi_{(M)}$ and $\theta_{(1)}, \ldots, \theta_{(M)}$ can be compared by a non-parametric test such as a rank test, if the parameters are scalar, and otherwise comparing single components. Under the null, $\tilde{\pi}(\theta|y) = \pi(\theta|y)$, the two sets have the same distribution. If we reject the null, this is evidence for approximation. Geweke (2004) and Cook et al. (2006) use ideas along these lines to check for correct MCMC (implementation, convergence) sampling of $\theta \sim \tilde{\pi}(\cdot|y')$, since in their case simulation of $\theta \sim \pi(\cdot|y')$ is possible using MCMC and they want to check it is working correctly. Yao et al. (2018) and Talts et al. (2018) move from testing MCMC convergence to diagnosing poor approximation of priors and likelihoods.

It is well known that this sort of check for $\tilde{\pi}(\theta|y) \simeq \pi(\theta|y)$, based on testing for exchangeable marginal distributions, can be fooled by an approximation which is far from the posterior. In particular if $0 \leq a \leq 1$ and for all $y' \in \mathcal{Y}$ our approximation is a linear combination of prior and posterior, $\tilde{\pi}(\theta|y') = a\pi(\theta|y') + (1-a)\pi(\theta)$ then

$$\begin{aligned} m(\theta, \phi) &= \int \pi(\phi|y')[a\pi(\theta|y') + (1-a)\pi(\theta)]p(y')\, dy' \\ &= a\int \pi(\phi|y')\pi(\theta|y')p(y')\, dy' + (1-a)\pi(\phi)\pi(\theta) \end{aligned} \tag{1.3}$$

and θ and ϕ are marginally exchangeable. The test passes with $\tilde{\pi}(\theta|y) \neq \pi(\theta|y)$ for any sample size M. The case where $a = 0$ (the approximate posterior is the prior) is discussed in Prangle et al. (2014) for Approximate Bayesian Computation (ABC) where it cannot be ignored, as this sort of error is a real possibility in ABC. They treat this issue by conditioning (ϕ, y') on $y' \in A$ for some set $A \subset \mathcal{Y}$. One of our algorithms (Importance Sampling in Algorithm 3) uses the same idea. We explain the connection between the two approaches below (2.5). The focus shifts in Rodrigues et al. (2018) from testing for good calibration to recalibration of approximate samples. This approach is discussed further in Section 4. In brief, Rodrigues et al. (2018) estimate a recalibration map and use it to map ABC samples onto the data in an ABC regression adjustment. By contrast, we extract the closely related coverage error map at the data, as it gives us the realised coverage we are achieving for an arbitrary nominal or intended coverage. In earlier work, Menendez et al. (2014) give a procedure for correcting a credible interval to give the nominal frequentist coverage for a parameter ϕ where a consistent estimator $\bar{\phi}$ is available. In Yao et al. (2018) and Talts et al. (2018), the approximation framework is unrestricted, however these authors take (1.2) as their starting point. They are interested in identifying how badly and in what ways the approximate distribution $\tilde{\pi}(\theta|y)$ differs from the exact distribution $\pi(\theta|y)$. They expect an approximation error, so there seems little point in testing for one, but characterising and visualising any shift in distribution is still useful.

We assume that the desired output of an analysis is a credible set, and that we have a method for estimating the credible set which involves making an approximation. Is the estimated approximate credible set good, in the following sense? In the original analysis, without an approximation, the credible set is designed to achieve the nominal coverage α for a parameter ϕ with two assumed properties: $\phi \sim \pi(\cdot)$ and $y \sim p(\cdot|\phi)$, that is ϕ is a draw from the prior and the data y inform ϕ through the observation model. We estimate the coverage our approximation actually achieves for a parameter ϕ satisfying these two properties. Coverage is usually taken to mean coverage of the unknown true parameter. Our definition of coverage is equivalent if we assume that the prior $\pi(\phi)$ is the true generative process for the unknown parameter ϕ and the observation model $p(y|\phi)$ is similarly correct. This is a shift from basing a test on (1.2) to basing a bias estimate on (1.1). This definition is appropriate as we focus on measuring approximation bias and not model misspecification error. In Section 2 we introduce regression and importance-sampling (IS) methods for the purpose. Although most of our examples treat credible intervals for a real scalar variable, this is not a restriction. Our simulation-based methods apply to any measurable credible sets, as long as they can be conveniently computed and specified. In our final example we work with a credible set for a random partition with no intrinsic linear order.

The methodology we describe may be computationally costly, since it may involve repeating the approximate inference procedure M times with M large. In some settings this would defeat the purpose of using an approximate scheme, since these are usually chosen to provide rapid answers. There are some mitigating factors. The runs can be processed in parallel, thus decreasing substantially the computation time. Also, for Algorithm 2, once the procedure as been run once, it can be used to evaluate the coverage at any future data set without further calibration simulation. However, although the

parallelism in particular is very helpful, it is also sometimes the case that the approximation we want to use cannot be made asymptotically exact, that is, we have no family of approximations with a "resolution" or "sample size" we can vary to improve accuracy, but just a single "fixed approximation". The Ising model example in Section 5 illustrates this. We replace the intractable partition function for a free boundary condition with the tractable partition function for periodic boundary conditions. We have no practical way to improve this approximation. Where this is the case, any serious analysis must provide some measure of the impact of the approximation on the reliability of results. A measure of the kind we provide, which measures the damage done to coverage, at the data we actually care about, directly addresses the impact of the approximation on a quantity central to the analysis. In all our examples, the likelihood only is approximated. In the notation above the approximate posterior may involve an approximation to the prior, the likelihood, or both. Talts et al. (2018) give an example with an approximate prior. Calibrating a posterior based on an approximate prior $\tilde{\pi}(\theta)$ is a straightforward variant of our approach, so long as we can simulate from the true prior $\phi \sim \pi(\cdot)$.

The remainder of the paper is structured as follows. We state our coverage estimation problem and give two algorithms which solve it in Section 2. We show how they work on a very simple Gaussian model with a tempered likelihood in Section 3. We describe a methodology to correct credible intervals in Section 4, building on Rodrigues et al. (2018). We give three further examples: an Ising model with a pseudo-likelihood in Section 5 illustrates all the methods in a simple setting where we have a sufficient statistic; a mixture model is analysed with a Variational Bayes approximation in Section 6; in Section 7 we calibrate the coverage of a random partition in a Dirichlet-Process model for the distribution of random effects in a hierarchical model. Code generating the results in Sections 5, 6 and 7 is available in the online supplementary material (Lee et al., 2019b).

2 Estimating coverage under an approximation

Let $\tilde{C}_y$ and C_y be level α credible sets for $\tilde{\pi}(\theta|y)$ and $\pi(\theta|y)$ respectively. These could for example be highest posterior density (HPD) sets (as in the examples in Sections 3 and 7) or equal- or lower tailed intervals (as in the examples in Sections 5 and 6). If $\pi(\theta|y) = \tilde{\pi}(\theta|y)$ for all $\theta \in \Omega$ then $\tilde{C}_y = C_y$ is an exact credible set for ϕ when $\phi \sim \pi(\cdot)$ and $y \sim p(\cdot|\phi)$, that is

$$\Pr(\phi \in C_Y | Y = y) = \alpha.$$

In our approximation, we take $\tilde{C}_y$ as an approximate level α credible set for $\pi(\phi|y)$. In this case we refer to α as the *nominal* level. Denote by $b(y)$,

$$b(y) = \Pr(\phi \in \tilde{C}_Y | Y = y), \tag{2.1}$$

the *operational* coverage probability.

We have additional Monte Carlo error if we use an estimate $\hat{C}_y(\theta)$ for $\tilde{C}_y$ computed using samples $\theta = (\theta_1, \ldots, \theta_J)$ simulated so that $\theta_j \sim \tilde{\pi}(\cdot|y)$ for $j = 1, \ldots, J$ (an abuse

of notation as θ had $J = 1$ up to this point). Let $c(y)$ give this second *realised* coverage probability at $Y = y$, averaged over ϕ and θ, so that

$$c(y) = \Pr(\phi \in \hat{C}_Y(\theta) | Y = y). \tag{2.2}$$

In this paper we give methods for estimating $b(y)$ and $c(y)$. In the examples in Sections 3, 5 and 6 we compute or estimate $b(y)$, as $\tilde{\pi}(\theta|y)$ is relatively simple and we can compute $\tilde{C}_Y$. In the example in Section 7 we estimate $c(y)$.

We now give the estimators for $c(y)$. Estimators for $b(y)$ are similar but simpler as the estimate $\hat{C}_y$ is replaced by $\tilde{C}_y$ (exact for $\tilde{\pi}(\theta|y)$). Let $Q(\phi)$ be a proposal distribution which we discuss below. For $i = 1, \ldots, M$ we simulate $\phi_{(i)} \sim Q(\cdot)$, $y_{(i)} \sim p(\cdot|\phi_{(i)})$ and $\theta_{(i)} = (\theta_{i,1}, \ldots, \theta_{i,J})$ with $\theta_{i,j} \sim \tilde{\pi}(\cdot|y_{(i)})$ for $j = 1, \ldots, J$. Here $y_{(i)} \in \mathcal{Y}$ is a data vector and similarly $\phi_{(i)} \in \Omega$ and $\theta_{(i)} \in \Omega^J$ for $i = 1, \ldots, M$. We form an estimate $\hat{C}_{(i)} = \hat{C}_{y_{(i)}}(\theta_{(i)})$ of the approximate credible set using the sample set $\theta_{(i)}$ and use it to compute binary values

$$c_i = \mathbb{I}_{\phi_{(i)} \in \hat{C}_{(i)}}.$$

We have two natural choices for estimating $c(y)$ from the "data" $(c_i, y_{(i)})_{i=1,\ldots,M}$ with different strengths and weaknesses. Before we give these estimators we give an idealised, but often impractical, algorithm estimating $c(y)$ consistently. See Algorithm 1. In this algorithm we simulate $\phi_{(i)} \sim \pi(\cdot|y)$ for $i = 1, \ldots, M$, set $y_{(i)} = y$ and then simulate $\theta_{(i)}$, $\hat{C}_{(i)}$ and c_i as above. In this case our data are $(c_i, y)_{i=1,\ldots,M}$ and $\hat{c} = M^{-1} \sum_i c_i$ is unbiased and consistent for $c(y)$. Of course this is no use if we cannot simulate $\pi(\phi|y)$. Algorithm 1 helps make clear what realised coverage means. We used this algorithm in the examples in Sections 6 and 7 to demonstrate our estimators were working. Algorithm 1 will be useful when we can sample the exact target and the approximation is in the estimate $\hat{C}_y(\theta)$ itself, not the posterior. This seems relatively straightforward. We focus below on cases where Algorithm 1 cannot be implemented.

Algorithm 1 (in general unrealisable) estimation of realised coverage $c(y)$.

1: **for** $i = 1, \ldots, M$ **do**
2: Simulate $\phi_{(i)} \sim \pi(\cdot|y)$ and $\theta_{(i)} = (\theta_{i,1}, \ldots, \theta_{i,J})$ with $\theta_{i,j} \sim \tilde{\pi}(\cdot|y)$ for $j = 1, \ldots, J$.
3: Estimate a credible set $\hat{C}_{(i)} = \hat{C}_y(\theta_{(i)})$ from the posterior samples, and binary values $c_i = \mathbb{I}_{\phi_{(i)} \in \hat{C}_{(i)}}$.
4: **end for**
5: The estimated coverage is $\hat{c}(y) = M^{-1} \sum_i c_i$.

We now give the estimators. The first method we describe is logistic regression. The test distribution is $Q(\phi) = \pi(\phi)$ so the simulation step is that of Cook et al. (2006) and Yao et al. (2018). See Algorithm 2. In the triple (ϕ, y', θ), we have $m(\phi, \theta|y') = \pi(\phi|y')\tilde{\pi}(\theta|y')$ conditionally, so if we take any particular y' we cover $\phi \in \hat{C}_{y'}(\theta)$ with probability $c(y')$. We take a vector $s(y) \in \mathbb{R}^p$ of p summary statistics computed on the data and a vector $\gamma \in \mathbb{R}^p$ of regression parameters, and carry out logistic regression with $\tilde{c}(y') = \text{logistic}(s(y') \cdot \gamma)$ and $c_i \sim \text{Bernoulli}(\tilde{c}(y_{(i)}))$ independent observations

for $i = 1, \ldots, M$. Our coverage estimate is simply $\hat{c}(y) = \text{logistic}(s(y) \cdot \hat{\gamma})$ with $\hat{\gamma}$ the maximum likelihood estimator for γ. We found replacing linear logistic regression with a semi-parametric generalised additive model (a GAM) using methods outlined in Wood (2011) worked well in our examples. The vector s of summary statistics must be chosen with care. The examples in Sections 3 and 5 have a sufficient statistic so the choice of s is straightforward, and more generally we expect good results for exponential family models. In Section 6, the ABC-optimal rule given in Fearnhead and Prangle (2012) inspired the choice of s. Our regression approach did not give sensible estimates for a harder problem we tried (a large scale version of the example in Section 7). For high dimensional data vectors $y \in \mathbb{R}^n$ with n large the simulated data y' do not enclose the real data y and so we are making a large extrapolation of the coverage function $c(y)$.

Algorithm 2 Estimation of realised coverage $c(y)$ using logistic regression.

1: **for** $i = 1, \ldots, M$ **do**
2: Simulate $\phi_{(i)} \sim \pi(\cdot), y_{(i)} \sim p(\cdot|\phi_{(i)})$ and $\theta_{(i)} = (\theta_{i,1}, \ldots, \theta_{i,J})$ with $\theta_{i,j} \sim \tilde{\pi}(\cdot|y_{(i)})$ for $j = 1, \ldots, J$.
3: Estimate a credible set $\hat{C}_{(i)} = \hat{C}_{y_{(i)}}(\theta_{(i)})$ from the posterior samples, and binary values $c_i = \mathbb{I}_{\phi_{(i)} \in \hat{C}_{(i)}}$.
4: **end for**
5: Take p summary statistics on the data $s : \mathcal{Y} \to \mathbb{R}^p$. Carry out logistic regression of $c_i \sim s(y_{(i)})$ onto the data yielding regression coefficient $\hat{\gamma}$.
6: The estimated coverage is $\hat{c}(y) = \text{logistic}(s(y) \cdot \hat{\gamma})$.

The second method we describe is importance sampling (IS) with proposal distribution $Q(\phi) = \tilde{\pi}(\phi|y)$. Denote by $\delta(y, y')$ a distance function in the space of data $\mathcal{Y}$. For small $\rho > 0$ with $\Delta_y = \{y'; \delta(y, y') \leq \rho\}$ we begin by making an ABC-style approximation to $c(y)$. Define the probability $d(y')$ for $\phi \sim \pi(\cdot)$ and $Y \sim p(\cdot|\phi)$ as

$$d(y) = \Pr(\phi \in \hat{C}_Y(\theta) | Y \in \Delta_y) \tag{2.3}$$

$$= \int_{\Omega \times \Omega^J} \int_{\mathcal{Y}} \mathbb{I}_{\phi \in \hat{C}_{y'}(\theta)} \frac{\pi(\phi) p(y'|\phi) \mathbb{I}_{y' \in \Delta_y}}{\Pr(Y \in \Delta_y)} \tilde{\pi}(\theta|y') \, d\theta \, dy' \, d\phi. \tag{2.4}$$

Equation (2.4) uses the same abuse of notation we made in (2.2), since θ is again a generic set of J samples, equivalent to one of the sample sets $\theta_{(i)}, i = 1, \ldots, M$ in Algorithm 3 below. Also $\tilde{\pi}(\theta|y')$ represents the joint distribution of these J samples. For example, if θ is the first J samples output by an MCMC run, then $\tilde{\pi}(\theta|y')$ gives their joint distribution in (2.4).

Our plan is to estimate $d(y)$ in (2.4) using importance sampling, and then use this as an estimate for $c(y)$, the operational coverage of interest. We motivate our approach by describing an approach that did not work in our setting. We might simulate

$$(\phi, y', \theta) \sim \pi(\phi) p(y'|\phi) \mathbb{I}_{y' \in \Delta_y} \tilde{\pi}(\theta|y'), \tag{2.5}$$

using rejection with $\phi \sim \pi(\cdot)$ and $y' \sim p(\cdot|\phi)$, and keeping only pairs (ϕ, y') satisfying $y' \in \Delta_y$, and then $\theta \sim \tilde{\pi}(\cdot|y')$ as before. This approach is used to good effect in the

ABC-setting of Prangle et al. (2014) and characterises our different aims and methods. While Prangle et al. (2014) start with (1.2) and then restrict to $y' \in \Delta_y$ in order to stop the prior-approximation (the $a = 0$ case in (1.3)) satisfying a coverage test, we start with (1.1) and aim to estimate the operational coverage. For the purpose of removing the $a = 0$ solution to (1.3) it may be enough to take a rather large set Δ_y. However, for estimating $c(y)$, we need simulated data close to the real data. We would like the coverage $c(y')$ to be flat over $y' \in \Delta_y$, so that in turn $d(y) \simeq c(y)$ is a reasonable approximation. For high dimensional simulated data y' we do not hit Δ_y in the rejection stage if we use (2.5). We therefore use importance sampling $\phi \sim Q$ with proposal distribution $Q(\phi) = \tilde{\pi}(\phi|y)$. This pushes our ϕ values into areas of parameter space where the realised y' values are much closer to the data y. We weight samples (ϕ, y', θ) using the normalised weight function

$$w(\phi, y', \theta) \propto \tilde{p}(y|\phi)^{-1},$$

in order to get a consistent estimator for $d(y)$. This gives coverage Algorithm 3. This works well on simple problems, and even on the harder problems set out in Section 7 and 5. However, the harmonic estimator proved to be too unstable for the biggest problems we tried (again, a problem related to the example in Section 7 but involving a much larger data set). Developing better estimators is an obvious next step.

Algorithm 3 Importance sampler estimating the realised coverage $c(y)$.

1: **for** $i = 1, \ldots, M$ **do**
2: while $\delta(y_{(i)}, y) > \rho$, simulate $\phi_{(i)} \sim \tilde{\pi}(\cdot|y), y_{(i)} \sim p(\cdot|\phi_{(i)})$ then
3: simulate $\theta_{(i)} = (\theta_{i,1}, \ldots, \theta_{i,J})$ with $\theta_{i,j} \sim \tilde{\pi}(\cdot|y_{(i)})$ for $j = 1, \ldots, J$.
4: **end for**
5: **for** $i = 1, \ldots, M$ **do**
6: estimate a credible set $\hat{C}_{(i)} = \hat{C}_{y_{(i)}}(\theta_{(i)})$ from the posterior samples $\theta_{(i)}$, binary values $c_i = \mathbb{I}_{\phi_{(i)} \in \hat{C}_{(i)}}$ and normalised importance weights $w_i \propto \tilde{p}(y|\phi_{(i)})^{-1}$.
7: **end for**
8: The estimated coverage is $\hat{c}(y) = \sum_i w_i c_i$.

One promising choice of distance function for scalar θ (*i.e.* $\Omega = \mathbb{R}$) that seems well-adapted to our setting was suggested to us by Bardenet and Ryder (2019). The "Kolmogorov-Smirnov distance function" $\delta(y', y) = \|\hat{G}_y - \hat{G}_{y'}\|_\infty$ (hereafter, KS-distance) is based on the posterior Cumulative Distribution Function (CDF) $G_y(\theta)$ of $\tilde{\pi}(\theta|y)$ at y. We are going to sample $\theta \sim \tilde{\pi}(\cdot|y')$ anyway, and these samples may be used to form an empirical CDF $\hat{G}_y$. The downside of this is that we must simulate $\theta \sim \tilde{\pi}(\cdot|y')$ for all the data-vectors y' we simulate, not just the ones that satisfy $y' \in \Delta_y$, since we need these samples to compute $\delta(y', y)$ itself.

3 Coverage of a normal mean

The diagnostic tools we have described cannot be "fooled" in quite the same way checks based on the exchangeability of ϕ and θ in (1.2) can be. This point and some other strengths and weaknesses are illustrated by the following very simple example.

Suppose the prior is $\phi \sim \mathcal{N}(0,1)$, the observation model is $y \sim \mathcal{N}(\phi,1)$ (*i.e.* the data vector is a scalar, and we have just a single normal observation), so that the exact posterior is $\pi(\phi|y) = \mathcal{N}(\phi; \frac{y}{2}, \frac{1}{2})$. Suppose now the approximate model has a tempered likelihood,

$$\tilde{\pi}(\theta|y) \propto \mathcal{N}(\theta;0,1)\mathcal{N}(y;\theta,1)^v, \tag{3.1}$$

for some $v \geq 0$, that is,

$$\tilde{\pi}(\theta|y) = \mathcal{N}\left(\theta; \frac{vy}{1+v}, \frac{1}{1+v}\right). \tag{3.2}$$

The approximation is good when $v = 1$ (no approximation) and bad when $v = 0$ (the approximation $\tilde{\pi}(\theta|y)$ coincides with the prior $\pi(\theta)$).

Let Z_α and $\Phi(z)$ be respectively the $1 - \frac{1-\alpha}{2}$ quantile (recall, in this paper α is typically 0.9 or 0.95) and CDF of a standard normal and let

$$B_\pm = \frac{vy}{1+v} \pm Z_\alpha \sqrt{\frac{1}{1+v}}$$

so that $\tilde{C}_y = [B_-, B_+]$ is a credible set for θ given y, with nominal level α. Now, referring to the coverage probability $b(y)$ defined in (2.1), we have

$$b(y) = \Phi\left(\sqrt{2}(B_+ - y/2)\right) - \Phi\left(\sqrt{2}(B_- - y/2)\right),$$

so that $b(y) = \alpha$ when $v = 1$ and in general $b(y) \neq \alpha$ otherwise (see Figure 1).

Consider estimating $b(y)$ using Algorithms 2 and 3. The algorithms simplify slightly as we can calculate the approximate set $\tilde{C}_y = [B_-, B_+]$ we need, we do not need to draw samples $\theta \sim \tilde{\pi}(\cdot|y')$ in order to form an estimate $\hat{C}_{y'}$, so we are estimating $\hat{b}(y)$ in (2.1) rather than $\hat{c}(y)$ in (2.2). In Figure 1 we plot the exact operational coverage $b(y)$ and its estimate $\hat{b}(y)$ as functions of the summary statistic $s(y) = y$ for various values of v and ρ taking $M = 10000$ simulated "data points" $(c_i, y_{(i)})_{i=1}^M$. At top left we see logistic regression with a GAM accurately estimates the operational coverage at all data and all approximation values. The remaining graphs show the behavior of the IS-estimator, Algorithm 3. Again we have $M = 10000$ but now the data points are simulated using the importance sampling proposal distribution – the approximate posterior itself. At top right the "easy" case $\alpha = 0.9$ and $v = 1$ (*i.e.* no approximation) is in fact demanding as the estimator in Algorithm 3 is now unbiased but sharply skewed in distribution at large α. At bottom left and right we see evidence that IS is not consistent for $b(y) = \Pr(\phi \in \tilde{C}_Y | Y = y)$ as it is consistent for $\Pr(\phi \in \tilde{C}_Y | Y \in \Delta_y)$. However, the bias is reduced in the bottom row as ρ is decreased from $\rho = 1$ to $\rho = 0.3$. The two methods, logistic regression and IS, expose the poor approximation at $v = 0$ (approximation-is-prior, corresponding to $a = 0$ in (1.3)) very well. Notice that a test for exchangeability of ϕ and θ would not expose this poor approximation.

In Appendix A (Lee et al., 2019a) we show that if $0 \leq v < 2$ then, for this simple normal example, the un-normalised IS estimator in Algorithm 3 satisfies a Central Limit Theorem. If the approximate posterior $\tilde{\pi}(\theta|y)$ is under-dispersed with respect to the true posterior $\pi(\theta|y)$ then the performance of the simple IS estimator we are using may be poor. We expect this property to hold in a qualitative sense in other settings.

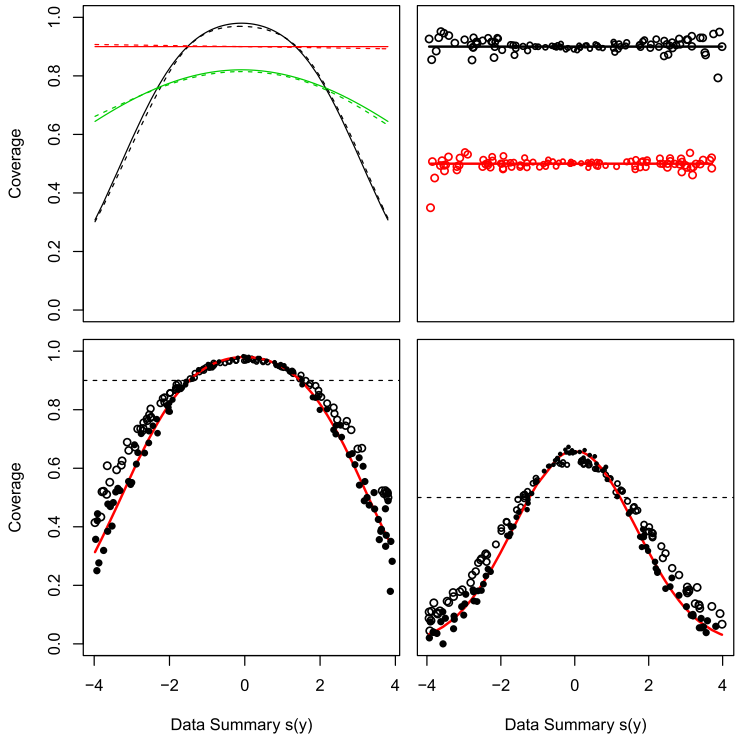

Figure 1: Tempered normal approximate likelihood. Y-axis is operational coverage. X-axis is scalar summary statistic $s(y) = y$. (Top Left) logistic regression (GAM) estimate of operational coverage $b(y)$ when $\alpha = 0.9$. Line types: truth $b(y)$, solid; estimated $\hat{b}(y)$, dashed. Colors: Red, $v = 1$, Green $v = 0.5$, Black $v = 0$. (Top Right) IS estimates of $b(y)$ at $v = 1$: lines give exact $b(y)$ at $\alpha = 0.9$ (black) and $\alpha = 0.5$ (red); points give IS estimates of $b(y)$, larger points have lower ESS. (Bottom Left) IS estimates of operational coverage with $v = 0$ and $\alpha = 0.9$: red line gives true operational coverage; open circles $\rho = 1$, full circles $\rho = 0.3$. (Bottom Right) As bottom left but with $\alpha = 0.5$.

4 Achieving the nominal level

The material in this section can be omitted at first reading. It is of independent interest, and highlights the connection between our work and Rodrigues et al. (2018). We have estimated the operational coverage $b(y)$ and the realised coverage $c(y)$ at the data for general credible sets respectively $\tilde{C}_y$ and $\hat{C}_y$ of fixed nominal level α. The framework above does not require the parameter θ to be a real scalar. We now restrict to $\Omega = \mathbb{R}$ and absolutely continuous distributions, and consider for now the level-α dependence of $c(y)$ specifically for lower tail credible intervals.

Let $\tilde{q}_{y'}(\alpha) = G_{y'}^{-1}(\alpha)$ be the level α quantile of $\tilde{\pi}(\theta|y')$ where $G_{y'}(\theta) = \int_{-\infty}^{\theta} \tilde{\pi}(\theta'|y')\, d\theta'$ is the CDF of θ at y' in the approximate posterior. Let $F_{y'}(\phi) = \int_{-\infty}^{\phi} \pi(\phi'|y')\, d\phi'$ be the

CDF for the true posterior given generic data y'. The operational coverage we achieve with our approximation $\tilde{\pi}(\theta|y)$ is a function of α at each y-value, and we write this as

$$b_y(\alpha) = \Pr(\phi \leq \tilde{q}_Y(\alpha)|Y = y). \tag{4.1}$$

This is the same as $b(y)$ but the dependence on the nominal level of coverage α is explicit. Equation (4.1) is the relation $b_y(\alpha) = F_y \circ G_y^{-1}(\alpha)$. This can be inverted to give the map from G_y to F_y at the data y,

$$F_y(\phi) = b_y \circ G_y(\phi).$$

Our coverage function $b_y(\alpha)$ is just the "distortion function" mapping the approximate CDF to the true CDF at the data. If we form estimates $\hat{b}_y(\alpha)$ of $b_y(\alpha)$ (using Algorithm 2 or 3) and $\hat{G}_y$ (the empirical CDF obtained using for example MCMC targeting $\tilde{\pi}(\theta|y)$) we may "recalibrate" G_y at the data y to better estimate F_y using the estimator

$$\hat{F}_y(\phi) = \hat{b}_y \circ \hat{G}_y(\phi). \tag{4.2}$$

This idea is set out in Rodrigues et al. (2018) who use it to map $\phi|y'$ to $\phi|y$ via the adjustment $\phi^{(adj)} = G_y^{-1} \circ G_{y'}(\phi)$. In our setting this sort of map would be effective. When we approximate $b_y(\alpha)$ with $\Pr(\phi \leq \tilde{q}_Y(\alpha)|Y \in \Delta_y)$ we assume $b_{y'}(\alpha)$ does not depend on y' for $y' \in \Delta_y$. If $\phi \sim \pi(\cdot)$ and $y' \sim p(\cdot|\phi)$ so that $\phi|y' \sim \pi(\cdot|y')$ then

$$\begin{aligned} \phi^{(adj)} &= G_y^{-1} \circ G_{y'}(\phi) \\ &= F_y^{-1} \circ b_y \circ b_{y'}^{-1} \circ F_{y'}(\phi) \\ &\simeq F_y^{-1} \circ F_{y'}(\phi), \end{aligned}$$

as $b_y \circ b_{y'}^{-1}(x) = x$ if $b_{y'}$ does not depend on y'. After adjustment, Rodrigues et al. (2018) have $\phi^{(adj)} \sim \pi(\cdot|y)$ (approximately). It is straightforward to check that b_y is invertible at each y. Rodrigues et al. (2018) use the empirical estimate $\hat{G}_y^{-1} \circ \hat{G}_{y'}(\phi)$ to implement the map.

We do not wish to make an adjustment of the kind Rodrigues et al. (2018) make, as we do not need to map samples θ at y' to samples $\phi^{(adj)}$ at the data y. We are interested in cases where we can generate approximately distributed samples at y by sampling $\theta \sim \tilde{\pi}(\cdot|y)$ itself. These samples could be recalibrated (at y) using (4.2). For example, if we seek a corrected median estimate we can replace the median estimate $\hat{G}_y^{-1}(0.5)$ for $\tilde{\pi}(\theta|y)$ with $\hat{G}_y^{-1} \circ \hat{b}_y^{-1}(0.5)$. Our aim is to provide an estimate $\hat{c}$ of the coverage of an approximate credible set $\hat{C}_y$, not an improved credible set. However we show how the correction may be made and give an example in Section 5.

Given Monte Carlo samples $\theta = (\theta_1, \ldots, \theta_J)$ distributed as $\tilde{\pi}(\cdot|y')$ (notation as (2.4)), and ordered so that $\theta_{i,j} < \theta_{i,j+1}$ for $j = 1, \ldots, J-1$, we estimate $\hat{q}_{y'}(\theta; \alpha) = \theta_{(\lceil \alpha J \rceil)}$. The realised coverage is

$$c_{y,J}(\alpha) = \Pr(\phi \leq \hat{q}_Y(\theta; \alpha)|Y = y).$$

Again we assume $c_{y,J}(\alpha) \simeq d_{y,J}(\alpha)$ where now

$$d_{y,J}(\alpha) = \Pr(\phi \leq \hat{q}_Y(\theta;\alpha)|Y \in \Delta_y)$$

is a function of the level. Reasoning as before,

$$d_{y,J}(\alpha) = \int_{\Omega\times\Omega^J}\int_{\mathcal{Y}} \mathbb{I}_{\phi\leq\hat{q}_{y'}(\theta;\alpha)} \frac{\pi(\phi)p(y'|\phi)\mathbb{I}_{y'\in\Delta_y}}{\Pr(Y'\in\Delta_y)}\tilde{\pi}(\theta|y')\,d\theta\,dy'\,d\phi.$$

We estimate this in Algorithm 4 using importance sampling draws from $\phi_{(i)} \sim \tilde{\pi}(\cdot|y_{(i)})$ and weighting by $1/\tilde{p}(y|\phi_{(i)})$ as before. We are estimating $c_{y,J}(\alpha)$ via a consistent estimator for $d_{y,J}(\alpha)$. In this setting it seems clear that following Bardenet and Ryder (2019) and using the distance function $\delta(y',y) = \|\hat{G}_y - \hat{G}_{y'}\|_\infty$ is desirable: if the CDF's are similar, at least for $y' \in \Delta_y$, then we may hope that the distortion functions $c_{y,J}(\alpha)$ and $c_{y',J}(\alpha)$ are similar, supporting our assumption $c_{y,J}(\alpha) \simeq d_{y,J}(\alpha)$.

Algorithm 4 Importance sampler estimating the realised coverage function $c_y(\alpha)$.

1: **for** $i = 1,\ldots,M$ **do**
2: while $\delta(y_{(i)},y) > \rho$, simulate $\phi_{(i)} \sim \tilde{\pi}(\cdot|y), y_{(i)} \sim p(\cdot|\phi_{(i)})$ then
3: simulate $\theta_{(i)} = (\theta_{i,1},\ldots,\theta_{i,J})$ with $\theta_{i,j} \sim \tilde{\pi}(\cdot|y_{(i)})$ for $j = 1,\ldots,J$, ordered so that $\theta_{i,j} < \theta_{i,j+1}$ for $j = 1,\ldots,J-1$.
4: **end for**
5: **for** $i = 1,\ldots,M$ **do**
6: compute the step functions $c_i(\alpha) = \mathbb{I}_{\phi_{(i)}\leq\theta_{i,\lceil\alpha J\rceil}}$ and normalised importance weights $w_i \propto \tilde{p}(y|\phi_{(i)})^{-1}$.
7: **end for**
8: The estimated coverage function at level α is $\hat{c}_y(\alpha) = \sum_i w_i c_i(\alpha)$.

Algorithm 4 makes Algorithm 3 redundant. We can use the function $\hat{c}_y(\alpha)$ output by Algorithm 4 to estimate the realised coverage $c(y)$ of our estimate $\hat{C}_y(\theta)$ by evaluating $\hat{c}_y(\alpha)$ at the value of α used to form $\hat{C}_y(\theta)$. However we can also correct $\hat{C}_y(\theta)$ to get a new interval with the required operational coverage. If we find the value $\tilde{\alpha}$ say, satisfying $\alpha = c_y(\tilde{\alpha})$ then $(-\infty, \tilde{q}_y(\tilde{\alpha})]$ covers ϕ with probability α. In practice we work with estimates, so we solve $\alpha = \hat{c}_y(\hat{\alpha})$ and estimate the credible set $\hat{C}_y(\theta) = (-\infty, \hat{q}_y(\theta;\hat{\alpha})]$ based on an adjusted level, in order to make the realised coverage match the desired nominal coverage. We give an example of this calculation in the next section (see the last paragraph and Figure 3 of Section 5). However, as noted above, this is a by-product of the analysis, not the essential point. We seek a quality guarantee, not a correction.

Algorithm 4, given here for lower-tail intervals, can be extended to handle equal-tailed intervals and HPD regions (Rodrigues et al. (2018) set this out in a general way). It is sufficient that credible sets at smaller nominal α are nested within credible sets at larger α. That makes $c_i(\alpha)$ an increasing step function and $\hat{c}_y(\alpha)$ non-decreasing.

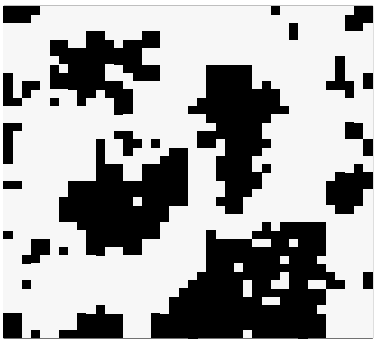

Figure 2: Ice floe data from Bornn et al. (2013).

5 Coverage of the Ising model smoothing parameter

The image in Figure 2 is a data set quoted from Bornn et al. (2013) where it was used to illustrate adaptive Wang-Landau simulation of a binary Markov Random Field. Those authors registered it by thresholding a larger grey-level image of ice floes published in Banfield and Raftery (1992).

In this section we fit a binary Markov random field (MRF) to these data and estimate the smoothing parameter, ϕ (also referred to as the inverse temperature, and usually denoted β, as we fit the Ising model). The data vector y records the $N \times N$ square array of binary values in the image in Figure 2, with $N = 40$. In the true observation model the MRF has a free boundary condition. This is a natural modelling choice but gives an intractable likelihood for ϕ. We approximate this with an observation model which has periodic boundary conditions. The likelihood for this second model is easily evaluated to machine precision. Foreshadowing our results, Figure 3 (left) shows the estimated coverage function $b(y)$. This is the probability our "wrong" credible set for ϕ with nominal coverage $\alpha = 0.95$ covers ϕ if ϕ is a draw from the prior and the data y is a draw from the Ising model with smoothing parameter ϕ. The coverage depends on a scalar sufficient statistic $s(y)$ defined below, so in Figure 3 (left) we plot $b(y)$ against $s(y)$. The coverage of our estimated credible interval $\tilde{C}_{y'}$ varies significantly over the space $\mathcal{Y}$ of data sets y', so it is important to estimate the operational coverage at the value of $s(y)$ corresponding to the data. Zhu and Fan (2018) calibrate an approximate fit to a Potts model using a coarsening procedure related to the real-space renormalisation group methods Gidas (1989) applies in image processing. Zhu and Fan (2018) give frequentist coverage probabilities at chosen values of the parameter ϕ.

The Ising model is a well known Markov model for a binary random field. Let G be a graph with edges E and vertices V. For $v \in V$ let $y_v \in \{0, 1\}$ be binary data at vertex v. Let $y = (y_v)_{v \in V}$ so that $y \in \mathcal{Y}$ with $\mathcal{Y} = \{0, 1\}^{N^2}$. Let $\langle u, v \rangle \in E$ denote a generic edge in G with vertices $u, v \in V$. Denote by

$$f(y; E) = \sum_{\langle u,v \rangle \in E} \mathbb{I}_{y_u \neq y_v}$$

the number of edges connecting non-equal neighbours on the graph. In our case G is a rectangular $N \times N$ lattice with $N = 40$ and a free boundary, $G_F = (E_F, V)$ say. On this graph interior vertices have degree 4, edge vertices have degree 3, and corner vertices have degree 2. We consider also lattices $G_P = (E_P, V)$ with periodic or toroidal boundary conditions. In this case the lattice is wrapped onto a torus and all vertices have 4 neighbours.

Let $\phi \geq 0$ be a positive scalar smoothing parameter. The Ising model distribution for a rectangular lattice with a free boundary is

$$p_F(y|\phi) = \frac{1}{Z(\phi; E_F)} \exp\left(-\phi f(y; E_F)\right), \tag{5.1}$$

where

$$Z(\phi; E_F) = \sum_{x \in \mathcal{Y}} \exp\left(-\phi f(x; E_F)\right) \tag{5.2}$$

is a normalising constant. The normalising constant $Z(\phi; E_F)$ is an intractable function of ϕ, for free boundary conditions, for N at all large. However, for the special case of periodic boundary conditions, $Z(\phi; E_P)$ is available from Beale (1996) in a simple closed form derived by Kaufman (1949) (for MatLab implementation see Lee et al. (2019b)).

Consider the problem of estimating the smoothing parameter ϕ for the data in Figure 2. Values of ϕ greater than about 2 are uninteresting for image modeling purposes as the image is essentially all 0's or all 1's under the prior. We take as prior $\pi(\phi) \propto \mathbb{I}_{\phi \in [0,2]}$. The posterior is

$$\pi(\phi|y) \propto \frac{1}{Z(\phi; E_F)} \exp\left(-\phi f(y; E_F)\right) \mathbb{I}_{\phi \in [0,2]}.$$

The likelihood for ϕ depends on the data y through the scalar quantity $s(y) = f(y; E_F)$ only, so this statistic is sufficient. This posterior is doubly intractable, due to the $Z(\phi; E_F)$-dependence. One approximate solution is to simply replace $Z(\phi; E_F)$ with $Z(\phi; E_P)$, which we can compute. Denote by

$$\tilde{\pi}(\theta|y) \propto \frac{1}{Z(\theta; E_P)} \exp\left(-\theta f(y; E_F)\right) \mathbb{I}_{\theta \in [0,2]}$$

the approximate posterior obtained on making this substitution. In this case the approximate posterior density and its CDF are readily evaluated using the formula for the partition function derived by Kaufman (1949), and we can compute $\tilde{C}_y$ to machine precision using the code in Lee et al. (2019b). The result for the data in Figure 2 is $\tilde{C}_y = [0.84, 0.90]$. This is exact for $\tilde{\pi}(\theta|y)$ but only approximate for $\pi(\phi|y)$. We would like to know the operational coverage $b(y)$ this approximation achieves.

We now run Algorithms 2 and 3 to estimate $b(y)$. As the credible interval under the approximation is available without simulation, the algorithms again simplify: as in Section 3, we estimate $b(y)$ rather than $c(y)$ in (2.1) and (2.2). We made $M = 1000$ simulations of ϕ and $y' \sim \tilde{\pi}(\cdot|\phi)$ in Algorithms 2 and 3. For our algorithm to be correct this simulation should be exact. However we simulated $y_{(i)}$ using a simple

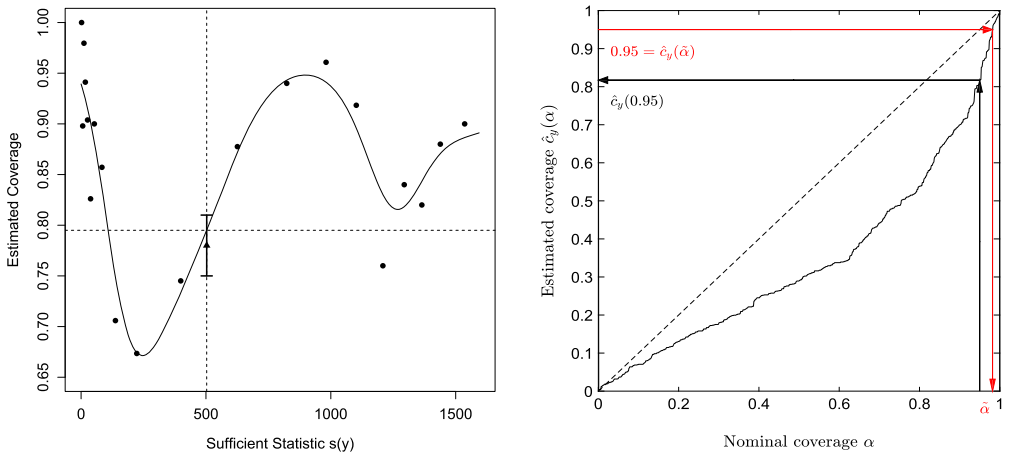

Figure 3: Estimation of coverage for the Ising model of Section 5. (Left) Estimated coverage as a function of the sufficient statistic $s(y)$: points are average coverage in x-axis bins each containing 50 data points; solid curve is GAM regression of coverage response on $s(y)$; vertical (horizontal) dashed line at the data value $s(y) = 503$ (resp. estimated operational coverage $\hat{c} = 0.80$) from the image in Figure 2; the error bar gives the IS-estimate $\hat{c} = 0.78$ and error $\hat{\sigma} = 0.03$. (Right) Solid line is estimated operational coverage $\hat{c}_y(\alpha)$ at data y plotted against nominal coverage α; dashed line is ideal operational coverage; black arrows give map from nominal coverage 0.95 to realised coverage 0.82; red arrows give inverse map from target realised coverage 0.95 to the nominal coverage $\tilde{\alpha} = 0.98$ which would achieve it.

single-site MCMC algorithm with a very large run-length. Although it is possible that this introduces another layer of bias, we took – for the purpose of this analysis – a very large run length and checked convergence carefully so that bias involved is negligible compared to the effect due to the boundary condition.

In our logistic regression in Algorithm 2 we use, as a covariate in the logistic regression, the summary statistic $s(y_{(i)}) = f(y_{(i)}; E_F)$ where $y_{(i)} \sim p(\cdot|\phi_{(i)})$. In Figure 3 (left) we plot the estimated operational coverage $\hat{b}(y')$ as a function of the sufficient statistic $s(y') = f(y'; E_F)$. This is the coverage we get over data space if we aim at a fixed nominal coverage equal 0.95 (*i.e.* $\alpha = 0.95$ is fixed, as in Section 2). The curve is a semi-parametric logistic regression (a GAM computed using the R function `gam()` in the package `mgcv`, see Wood (2011)) of the coverage response c_i in Algorithm 2 on the sufficient statistic, where $y_{(i)} \sim p(\cdot|\phi_{(i)})$ is the simulated data at $\phi_{(i)}$ and $\phi_{(i)} \sim \pi(\cdot)$ is a draw from the prior, for $i = 1, \ldots, M$. In this setting, with a sufficient statistic, this is a fairly reliable estimate of the true operational coverage function $b(y)$, interpolating the proportion of c-values equal to 1 in the neighbourhood of each $s(y)$-value. The value of the sufficient statistic at the data is $s(y) = 503$, so our best estimate of $b(y)$ at the real data y (*i.e.* the GAM fit at $s(y)$) is $\hat{b}(503) \simeq 0.80$. The big dip in the curve at small $s(y)$ is at data values typical for the Ising model phase transition, where the spatial

correlation length diverges, and the system is sensitive to boundary conditions. At large $s(y)$-values the spatial correlation length is short. At very small $s(y)$-values the state is mostly one color, with inclusions much smaller than the region. In these cases the estimated coverage is close to nominal.

In Algorithm 3, we used the KS-distance $\delta(y, y') = \|G_y - G_{y'}\|_\infty$ in our importance sampling estimation. We set the threshold distance at 0.5. This gave an effective sample size of 275 (out of 1000 samples). This (*i.e.* $\rho = 0.5$) may seem large, however it reflects the way $G_{y'}$ changes as y' varies. The shape (and we hope the distortion function b'_y) of $G_{y'}$ remains almost unchanged as y' varies. However, the overlap of $G_{y'}$ and G_y in Ω goes rapidly to zero as y' moves away from y, so large ρ corresponds to distributions relatively close together in Ω. We expect $b_{y'}$ to be insensitive to small variations of this sort, and since data y' with similar $b_{y'}$ functions to b_y are good data, we use it. We saw a clear dependence of weight variance on KS-distance. If we set the threshold distance just below 1 (the maximum possible) the ESS is reduced to 32 as there are some very large weights at larger KS-distances. Data y' at large KS-distance from y is associated with parameter values ϕ that have large IS-weights $1/\tilde{p}(y|\phi)$, so the KS-distance is helpful in stabilising our estimator in this case. Estimation of $b(y)$ using importance sampling, Algorithm 3 yields $\hat{c} = 0.78$ with standard deviation $\hat{\sigma} = 0.03$, where

$$\hat{\sigma}^2 = \sum_{i=1}^{M} w_i^2 (c_i - \hat{c})^2.$$

The convenience of semi-parametric logistic regression in this simple setting is striking. However, importance sampling was also straightforward.

In Figure 3 (right) we plot the estimated calibration function $\hat{c}_y(\alpha)$. This gives the operational coverage achieved by our estimator as a function of the nominal coverage we are targeting. This function was estimated as in Algorithm 4 by forming a weighted average of the binary step functions

$$c_i(\alpha) = \mathbb{I}_{\phi_{(i)} \leq \tilde{q}_{y_{(i)}}(\alpha)}$$

defined for $0 \leq \alpha \leq 1$. The black arrow follows the map from a nominal coverage of 0.95 to the realised coverage (about 0.82, not a perfect match for the value 0.78 we estimated using IS in Algorithm 3 due to Monte Carlo error, but note also that we have switched from equal-tailed to lower-tail credible intervals in moving from Figure 3 (left) to (right)). We can also ask, what nominal level would give operational coverage equal 0.95? This is the inverse map represented by the red arrows. We see we should have used $\alpha = 0.98$ if we wanted to cover $\phi \sim \pi(\cdot|y)$ 95% of the time.

6 Mixture-model parameters via Variational Bayes

Consider data from a mixture of two normal distributions

$$y \sim p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2). \tag{6.1}$$

We impose $0 < p < \frac{1}{2}$ to ensure identifiability, and wish to estimate the location of the secondary mode μ_1. To this end, we use Variational Bayes (VB, (Jordan et al., 1999)).

VB provides an analytical approximation to the posterior distribution $\pi(\cdot|y)$, by finding the parametric distribution which minimizes the Kullback-Leibler divergence

$$\tilde{\pi} = \arg\min_{Q \in \mathcal{Q}} D_{KL}\left(Q(\cdot)||\pi(\cdot|y)\right), \tag{6.2}$$

where $\mathcal{Q}$ is a parametrized set of distributions. In our example, the set $\mathcal{Q}$ is defined by imposing that the approximate posterior be of the form

$$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) \sim \mathcal{N}(\nu, \tau) \otimes \mathcal{N}(\nu', \tau') \otimes \mathcal{IG}(a, b) \otimes \mathcal{IG}(a', b') \otimes \mathcal{B}(a'', b'')$$

for some values of the scalars $(\nu, \tau, \nu', \tau', a, b, a', b', a'', b'')$, where $\mathcal{N}$, $\mathcal{IG}$ and $\mathcal{B}$ refer to the Normal, Inverse-Gamma and Beta distributions respectively. We use $\tilde{\pi}$ defined by (6.2) as an approximate posterior distribution. The computation of the optimal $\tilde{\pi}$ can be done very rapidly; we used the implementation of the R package vabayelMix (Teschendorff, 2006). We refer the reader to Blei et al. (2017) for a review of VB, including its application to a mixture of normals.

We implemented Algorithm 2 with $M = 10000$ synthetic data sets $y_{(1)}, \ldots, y_{(M)}$. For $i = 1, \ldots, M$, each data set $y_{(i)}$ is a set of size $n = 20$ simulated from the mixture model in (6.1) with parameters drawn from the priors $\mu_1^{(i)} \sim \mathcal{N}(0, 10)$, $\mu_2^{(i)} \sim \mathcal{N}(0, 10)$, $p^{(i)} \sim \mathcal{U}([0, \frac{1}{2}])$, and $\sigma_1^{(i)} = \sigma_2^{(i)} = 1$. For each synthetic data set $y_{(i)}$, we compute the VB approximate posterior, which we summarize by the set of statistics $s(y_{(i)}) = (|\hat{\mu}_1^{(i)} - \hat{\mu}_2^{(i)}|, \hat{p}^{(i)}, \hat{\sigma}_1^{(i)}, \hat{\sigma}_2^{(i)}, \frac{1}{\hat{\sigma}_1^{(i)}}, \frac{1}{\hat{\sigma}_2^{(i)}})$, where $\hat{\mu}_1^{(i)}$ is the expected value of μ_1 under the VB approximate posterior $\tilde{\pi}(\cdot|y_{(i)})$, and similarly for the other parameters. This is inspired by the ABC-optimal choice of Fearnhead and Prangle (2012); as with any ABC-like method, the choice of the summary statistics is crucial and including better summary statistics if available can vastly improve the inference: we also experimented with other summary statistics, including the data mean, standard deviation and various quantiles, but found that these statistics did not improve our estimates. We use the VB approximate marginal posterior for μ_1 to compute analytically a 90% credible interval $\hat{C}_{(i)} = \tilde{C}_{y_{(i)}}$ for $\mu_1|y_{(i)}$ and record the binary value $c_i = \mathbb{I}_{\mu_1^{(i)} \in \hat{C}_{(i)}}$.

We regressed (using a GAM as above) the coverage indicator c_i against the set of summary statistics $s(y_{(i)})$. These are not sufficient, so we expect some loss of precision. Once this regression is performed, it can be used (at no further computational cost) to estimate the coverage of different observed data sets for the same model. The output of the regression allows us to estimate the coverage of the VB approximate posterior given the output of VB for some observed data y. To evaluate the methodology, we estimated the coverage by simulating $N_{test} = 2000$ new "observed" data sets. For each data set $j = 1 \ldots N_{test}$, we used VB to compute an approximate HPD[1] interval $\hat{C}_j$ with nominal coverage $\alpha = 0.90$. We recorded the estimated coverage $\hat{c}_j \in [0, 1]$ given by Algorithm 2 as well as the binary value $c_j = \mathbb{I}_{\mu_1^{(j)} \in \hat{C}_j}$. We then computed the cross-entropy

$$\frac{1}{N_{test}} \sum_j -c_j \log(\hat{c}_j) - (1 - c_j) \log(1 - \hat{c}_j).$$

[1]Due to the nature of the VB approximation, an HPD interval is an equal-tailed credible interval.

A lower cross-entropy means we are better at estimating the operational coverage. We compare these estimated coverages to the nominal coverage $\forall j, \hat{c}'_j = 0.90$ and to the best constant estimator $\forall j, \hat{c}''_j = \frac{1}{N_{test}} \sum c_j = 0.719$. Algorithm 2 gave cross-entropy 0.435, $\hat{c}'_j = 0.90$ gave 0.773, and $\hat{c}''_j = 0.719$ gave 0.616. The fact that we outperform $\hat{c}'_j = 0.90$ indicates that Algorithm 2 estimates the operational coverage better than the nominal coverage. The fact that we outperform $\hat{c}''_j = 0.719$ indicates that we are able to detect in which parts of the space the coverage is higher or lower than average. This test is available in general. For this particular model, we can implement Algorithm 1 and compute the mean squared error as an additional check. We used the Gibbs sampler implementation of the R function `rnmixGibbs` from the package `bayesm` (Rossi and McCulloch, 2017) to target $\pi(\phi|y)$. This gives a consistent estimator of the operational coverage not always available in real applications. We treat this estimate as exact, as the MCMC sample size was large and the Monte Carlo error small. We generate an MCMC sample $(\theta_{j,1}, \ldots, \theta_{j,K}) \sim \pi(\cdot|y_{(j)})$. Convergence diagnostics show that $K = 1000$ is reasonable. We record as the true operational coverage the value $\bar{c}_j = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\theta_{j,k} \in \hat{C}_j}$. The mean squared error is approximately $\frac{1}{N_{test}} \sum (\hat{c}_j - \bar{c}_j)^2$ (ignoring Monte Carlo error). Using the nominal coverage 0.9 leads to mean squared error of 0.109; Algorithm 2 shows a 10-fold improvement with mean squared error 0.0106.

Results are shown in Figure 4, which plots this "true" operational coverage $\bar{c}_j$ against the estimate $\hat{c}_j$ given by Algorithm 2. Points are scattered close to the line $y = x$, indicating that our coverage estimator is reliable. Figure 5 (top) gives the true distribution of operational coverage across 1000 approximate credible intervals with nominal level 0.90. For unimodal posterior distributions, VB often provides a good estimate of the posterior mean but underestimates the posterior variance (Blei et al., 2017), *i.e.* approximate credible intervals may have operational coverage much lower than the nominal coverage. The distribution of estimated coverage shown in Figure 5 (bottom) is smeared out by Monte-Carlo error and estimator bias. Figure 4 shows that we overestimate the coverage a little when it is greater than about 0.75. This must be a consequence of the choice of summary statistics, as the GAM-based estimator will have negligible bias over the space $\mathcal{S}$ of summary statistics. The summary statistics map many data sets $y_{(i)} \in \mathcal{Y}$ to a point, $s(y_{(i)}) \in \mathcal{S}$, bringing their c_i-values with them. Notice the atom of mass at operational coverage equal one in Figure 5 (top) and (bottom) generated by data vectors in $A = \{y' \in \mathcal{Y} : c(y') \simeq 1\}$ say. This corresponds to data where the VB credible intervals are far too wide. Algorithm 2 detects this so both histograms show a spike at one. However, the estimated coverage at $s \in \mathcal{S}$ averages coverage from data in its pre-image $\mathcal{Y}_s = \{y' \in \mathcal{Y} : s(y') = s\}$. Since $c(y')$ need not be constant for $y' \in \mathcal{Y}_s$, we over-estimate $c(y')$ when y' is part of a pre-image $\mathcal{Y}_{s(y')}$ having more overlap with A. The same process lowers estimated coverage for some data points in A (top right of Figure 4, where $c(y) \simeq 1$ but $\hat{c} < 1$). Data points with lower coverage, where the distribution is flatter, are less affected.

7 Coverage of a partition of random effects

The `car90` data contain specifications of $n = 111$ cars, extracted from *Consumer Reports, 1990*. The dependence of car price on car specification is of interest. The

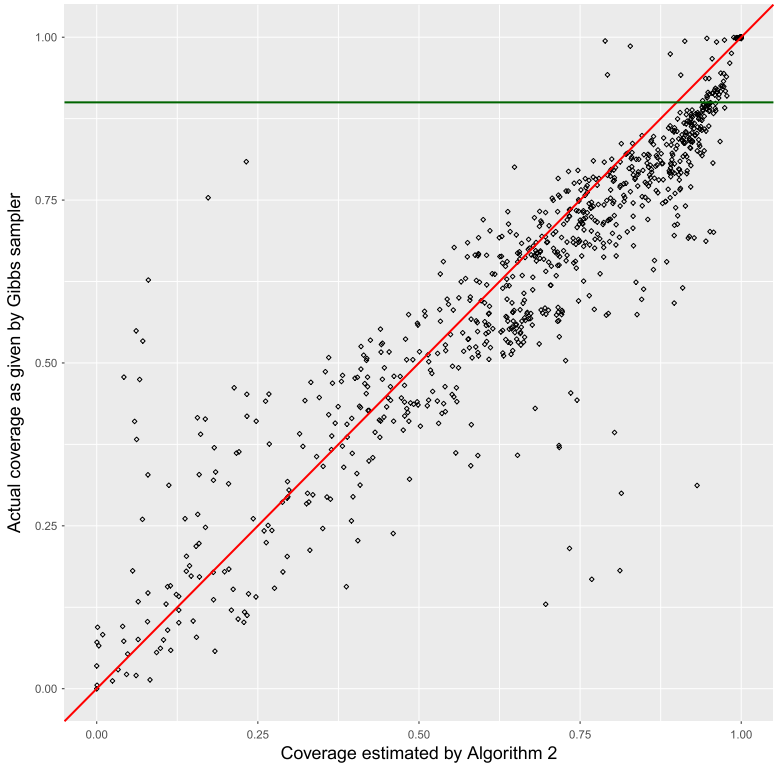

Figure 4: For the mixture of normals example in Section 6 and $j = 1 \ldots N_{test}$, true operational coverage $\bar{c}_j \simeq b(y_{(j)})$ is plotted on the y-axis against operational coverage $\hat{c}_j$ (on the x-axis) estimated by Algorithm 2. Most points are close to the $y = x$ red line. The green horizontal line at $y = 0.9$ shows the nominal coverage, which is far from the operational coverage in most cases.

dataset is available in the R package `rpart` (Therneau and Atkinson, 2018). We focus on the problem of clustering the levels of a categorical variable as part of the modeling. We use these data to illustrate an approximate method for fitting a Dirichlet process model for the clustering. The output of this analysis is a credible set of partitions of the levels indicating how the levels may plausibly be grouped. For this purpose we select from the original 33 variables the engine displacement in cubic inches ($x = (x_1, \ldots, x_n)$), the red line value (the maximum safe engine speed in rpm, $z = (z_1, \ldots, z_n)$) and the car type ($t = (t_1, \ldots, t_n)$ with $t_i \in \mathcal{T}$ for $i = 1, \ldots, n$ and $\mathcal{T} = \{\text{small, medium, large, van, compact, sporty}\}$).

Let $S = (S_1, \ldots, S_K)$ be a partition of $\mathcal{T}$ into K sets, with $K \in \{1, 2, \ldots, 6\}$, and for $i = 1, \ldots, n$ let s_i denote the partition for car i so that $s_i = k$ is equivalent to $t_i \in S_k$. In our model the overall effects due to type are assumed to fall in groups: we have a separate effect γ_k for each group $k = 1, \ldots, K$, and an effect for each type within each group,

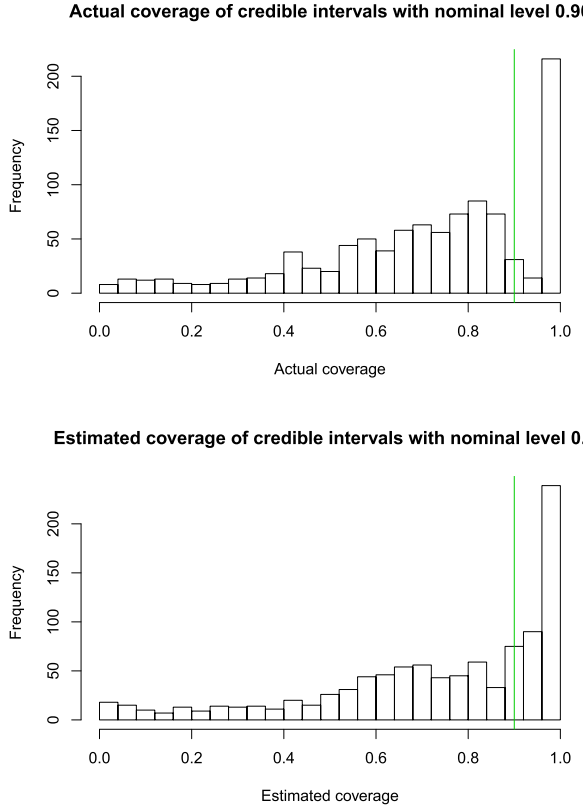

Figure 5: True and estimated distributions of VB coverage, $x-$ (lower) and $y-$ (upper) marginals of Figure 4. For 1000 data sets we estimate the operational coverage of VB at that data set using Algorithm 2 (lower, our estimate) and again using Algorithm 1 and an exact Gibbs sampler (upper, the true operational coverage). The top plot is a histogram of the actual coverage, which is often very far from the nominal level of 0.90 (green vertical line). The bottom plot shows a histogram of the estimated coverage using our method. Our method correctly identifies the tendency for VB to give low coverage.

$\eta_\tau, \tau \in \mathcal{T}$. The two random effect covariance matrices Σ_η and Σ_γ are assumed diagonal with diagonal elements $h_\eta\sigma^2$ and $h_\gamma\sigma^2$ respectively, where σ^2 is defined below as the response variance. The overall random effect, η'_{t_i} say, for observation i is $\eta'_{t_i} = \eta_{t_i} + \gamma_{s_i}$ (notice s_i is determined from t_i given S). Let $\gamma = (\gamma_1, \ldots, \gamma_K)$ and $\eta = (\eta_1, \ldots, \eta_6)$.

The model in this section clusters random effects "by covariance". If we integrate out $\gamma \in \mathbb{R}^K$ given the partition S then we are left with a model in which random effects in the same group have a higher covariance (*i.e.* $h_\gamma\sigma^2$) than random effects in different groups (where the covariance zero). Malsiner-Walli et al. (2018) and Pauger and Wagner (2018) treat the same problem in a similar way but use a different parameterisation and prior. Malsiner-Walli et al. (2018) use a mixture of spiky normal distributions in their

prior for label effects. Pauger and Wagner (2018) take the overall random effects for type $\eta'_\tau, \tau \in \mathcal{T}$ and introduce covariance terms off-diagonal in the random effects covariance matrix $\Sigma_{\eta'}$. They operate directly on the elements of the covariance matrix in order to explore the model space.

Given the partition S the price y in \$1000 dollars is modelled using the following random and fixed effects, for $i = 1, \ldots, n$,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \eta_{t_i} + \gamma_{s_i} + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The parameter priors are

$$\pi(\sigma^2) \propto 1/\sigma^2\,, \quad \beta_0, \beta_1, \beta_2 \sim \mathcal{N}(0, h_b\sigma^2)\,, \quad \eta_\tau \sim \mathcal{N}(0, h_\eta\sigma^2), \tau \in \mathcal{T}\,.$$

The partition S is unknown. Let $\mathcal{P}$ denote the space of partitions of our six types (there are 203 distinct partitions). We take a Chinese restaurant process (CRP) prior $\pi(S)$ for S, with clustering parameter $\alpha_{CRP} = 1$

$$\pi(S) = \frac{\Gamma(\alpha_{CRP})}{\Gamma(\alpha_{CRP} + n)} \alpha_{CRP}^K \prod_{k=1}^{K} \Gamma(n_k)$$

where n_k is the number of datapoints in partition k. We are modeling the random effects via a Dirichlet Process $G_\gamma \sim DP(\alpha, H), \gamma_k \sim G_\gamma$ with base density

$$H(\gamma_k) = \mathcal{N}(\gamma_k; 0, h_\gamma\sigma^2), k = 1, \ldots, K\,.$$

Integrating over the DP random measure G_γ yields the prior

$$\pi(\gamma, S|h_\gamma, \sigma^2) = \pi(S) \prod_{k=1}^{K(S)} \mathcal{N}(\gamma_k; 0, h_\gamma\sigma^2).$$

The scale parameter priors are $h_\eta, h_\gamma, h_b \sim Inv\chi^2(1)$. Let $h = (h_\eta, h_\gamma, h_b)$.

We are interested in the marginal posterior distribution of $S|y$ and estimating a HPD credible set for S. Let

$$\psi = (\beta_0, \beta_1, \beta_2, \sigma, \gamma, \eta)$$

denote the vector of parameters besides S and h. It is often convenient (*i.e.* in models slightly more complex that this) to work directly with the marginal (or "collapsed") posterior

$$\pi(S|y) \propto p(y|S)\pi(S),$$

where $p(y|S)$ is the intractable marginal likelihood

$$p(y|S) = \int p(y|S, h, \psi)\pi(\psi|h, S)\pi(S, h)\, d\psi\, dh.$$

However $p(y|S, h)$ is available in closed form and there are a number of ways one might then proceed to solve the problem without further approximation (for example using asymptotically exact MCMC). Here we simply set $h_\eta = h_\gamma = h_b = 10$, that is we define

$$\tilde{p}(y|S) = p(y|S, h)|_{h_\eta = h_\gamma = h_b = 10}$$

Partition, S	$G(S\|y)$
(Compact,Large,Medium,Small,Sporty,Van)	0.21
(Compact,Large,Medium,Sporty,Van),(Small)	0.25
(Compact,Medium,Small,Sporty,Van),(Large)	0.29
(Compact,Large,Medium,Small,Sporty),(Van)	0.32
(Compact,Large,Medium,Small,Van),(Sporty)	0.36
⋮	⋮
(Compact,Van),(Large,Sporty),(Medium,Small)	0.95

Table 1: HPD set for partition S in Section 7 using $\tilde{\pi}(S|y)$. Rows are partitions sorted by posterior probability with the largest first. The second column is the cumulative sum (*i.e.* the CDF $\hat{G}_y(s)$). There are 144 partitions in this HPD set.

and

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

We implemented MCMC targeting $\tilde{\pi}(S|y)$ using Metropolis Hastings MCMC updating one level of car type at each update. We define and estimate the empirical CDF $\hat{G}_{y'}(S),\ S \in \mathcal{P}, y' \in \mathcal{Y}$ and associated KS distance as follows. For $S \in \mathcal{P}$ let $\hat{\pi}(S|y)$ be an estimate of the approximate posterior formed from the MCMC output, and let $>_y$ be the (random) order on partitions determined by $S >_y S' \Leftrightarrow \hat{\pi}(S|y) > \hat{\pi}(S'|y)$. This order is always fixed by the real data y. The empirical cdf of the approximate posterior at $y' \in \mathcal{Y}$ is $\hat{G}_{y'}(S) = \sum_{S' \geq_y S} \hat{\pi}(S'|y')$ and the estimated KS distance is $\delta(y', y) = \|\hat{G}_y - \hat{G}_{y'}\|_\infty$.

The level $\alpha = 0.95$ HPD set is shown in Table 1. We would like to use our calibration check to see if this Monte Carlo HPD set is reliable. We estimate the coverage probability $c(y)$ by estimating $d(y)$ in (2.3) using importance sampling, Algorithm 3, with $M = 100000$ samples and the KS-distance function. In this multi-parameter setting we need ψ and h in order to simulate $p(y|S, h, \psi)$. These can often be sampled from their priors, so the only importance re-weighting comes with replacing $S \sim \pi(\cdot)$ with $S \sim \tilde{\pi}(\cdot|y)$. Here σ has an improper prior, so we take $\phi = (\psi, S)$ in Algorithm 3 and sample $\psi, S|y \sim \tilde{\pi}(\cdot|y)$, where $\tilde{\pi}(\psi, S|y) \propto \pi(\psi, S, h|y)|_{h=10}$ is the approximate posterior for all the parameters. When we use this importance sampling proposal distribution, the IS weight is $w(S, \psi) \propto 1/\tilde{p}(y|S, \psi)$.

For each of 30 threshold values ρ from 0.11 to 1, we estimated the operational coverage and the effective sample size. Results are summarised in Figure 6. In order to establish ground truth, we additionally estimate the operational coverage (at $c(y) \simeq 0.98$) using Algorithm 1. This gives an unbiased and consistent estimate of $c(y)$ with very small uncertainty. We regard this as the truth (horizontal line in graph at left in Figure 6). This check would not be available in general (if we can implement Algorithm 1 we can sample $\pi(\phi|y)$). The coverage-probability estimate varies sharply with ρ and approaches the true operational coverage $c(y) \simeq 0.98$ at small ρ with an ESS dropping from 50 to 15 in the last two (leftmost) point estimates. We see a bias-variance trade-off here. As $\rho \to 0$, we drop samples from the estimate, and increase estimator

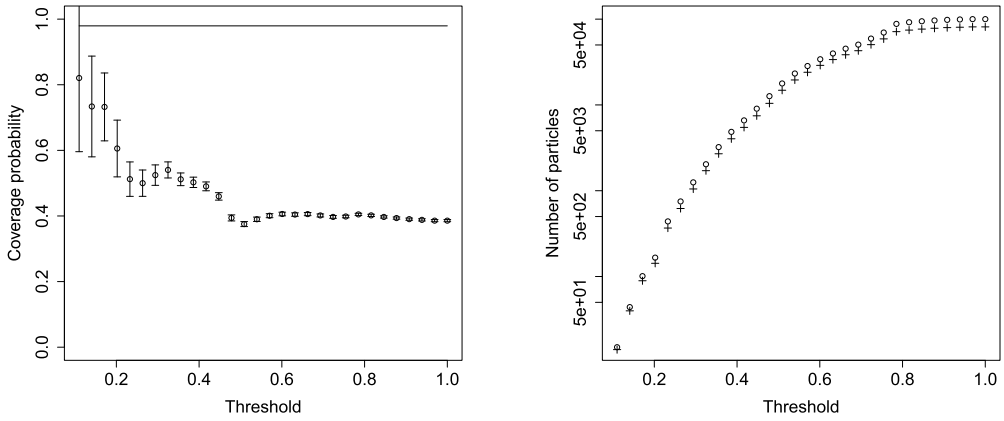

Figure 6: (Left) Coverage probability estimates $\hat{c}(y)$ for Section 7. For each threshold value ρ a KS-distance $\delta(y', y)$ and an operational coverage probability $c(y)$ are estimated. Error bars are 2-sigma. Horizontal line (at 0.98) is a high precision estimate of $c(y)$ using Algorithm 1. (Right) Effective Sample Sizes (ESS) using Importance sampling. The number of particles in the Δ_y neighbourhood (circle) and the ESS (cross) of the weighted samples surviving the cut, as a function of ρ.

variance. However, the bias decreases. When ρ is large the estimator averages coverage over the whole data space. This is wrong, as the operational coverage achieved by the approximate posterior at the data is much better than its average over all possible data. However, as $\rho \to 0$, and $d(y)$ approaches $c(y)$, the estimator mean approaches the true operational coverage $c(y)$ conditioned on the data.

8 Conclusions

We have presented two methods for estimating the operational coverage of approximate credible sets. Our examples show that the operational coverage can be far from the nominal coverage. When we are in control of the precision of our approximation it may be convincing simply to check that credible sets are stable as precision is increased. However when we make a fixed approximation, as we do in Sections 5, 6 and 7, this standard check is no longer available, and in all cases a measure of the operational posterior coverage our posterior approximation achieves will be of interest.

Depending on the setting, Algorithm 2 or 3 may be easier to implement. We have found that both algorithms are relatively straightforward, and both apply to a wide range of posterior approximations. The principal weakness of our approach is judging the reliability of our reliability checks. If we estimate a poor coverage, $\hat{c} \ll \alpha$ far from the nominal level, we should be concerned. We found in practice that, when our coverage estimators failed, the fact that they had failed was obvious (very small ESS or unreliable extrapolation in high dimension) and showed up in straightforward stability checks. In the example in Section 5, standard logistic regression checks provide good evidence that

our reliability estimate is itself reliable. However, the estimators we give are simple and could be improved a great deal.

Approximate inference schemes are essential tools in Bayesian analysis, and of rising importance. Accompanying a credible set with a calibration measure provides a generic and easily understood check, or alternatively warns us the approximation has failed. The non-negligible computational time these methods take seems a reasonable price to pay for such a check.

Supplementary Material

Calibration procedures for approximate Bayesian credible sets: SUPPLEMENT
(DOI: 10.1214/19-BA1175SUPPA; .pdf).

Supplementary material – code
(DOI: 10.1214/19-BA1175SUPPB; .zip).

References

Banfield, J. D. and Raftery, A. E. (1992). "Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering about Principal Curves." *Journal of the American Statistical Association*, 87(417): 7–16. 1256

Bardenet, R. and Ryder, R. J. (2019). In preparation. 1251, 1255

Beale, P. (1996). "Exact Distribution of Energies in the Two-Dimensional Ising Model." *Physical Review Letters*, 76: 78–81. 1257

Besag, J. (1975). "Statistical analysis of non-lattice data." *The Statistician*, 24(3): 179–195. 1245

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational inference: A review for statisticians." *Journal of the American Statistical Association*, 112(518): 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 1260, 1261

Bornn, L., Jacob, P., Moral, P., and Doucet, A. (2013). "An Adaptive Interacting Wang-Landau Algorithm for Automatic Density Exploration." *Journal of Computational and Graphical Statistics*, 22(3): 749–773. MR3173740. doi: https://doi.org/10.1080/10618600.2012.723569. 1256

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). "Validation of software for Bayesian models using posterior quantiles." *Journal of Computational and Graphical Statistics*, 15: 675–692. MR2291268. doi: https://doi.org/10.1198/106186006X136976. 1245, 1246, 1249

Fearnhead, P. and Prangle, D. (2012). "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):

419–474. MR2925370. doi: https://doi.org/10.1111/j.1467-9868.2011.01010.x. 1245, 1250, 1260

Geweke, J. (2004). "Getting it right: Joint distribution tests of posterior simulators." *Journal of the American Statistical Association*, 99: 799–804. MR2090912. doi: https://doi.org/10.1198/016214504000001132. 1245, 1246

Gidas, B. (1989). "A Renormalization Group Approach to Image Processing Problems." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(2): 164–180. 1256

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine learning*, 37: 183–233. 1245, 1259

Kaufman, B. (1949). "Crystal Statistics II. Partition function evaluated by spinor analysis." *Physical Review*, 76: 1232–1243. 1257

Lee, J. E., Nicholls, G. K., and Ryder, R. J. (2019a). "Supplementary material – Appendices." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1175SUPPA. 1252

Lee, J. E., Nicholls, G. K., and Ryder, R. J. (2019b). "Supplementary material – code." URL https://github.com/robinryder/calibration. *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1175SUPPB. 1248, 1257

Malsiner-Walli, G., Pauger, D., and Wagner, H. (2018). "Effect fusion using model-based clustering." *Statistical Modelling*, 18(2): 175–196. MR3770129. doi: https://doi.org/10.1177/1471082X17739058. 1263

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). "Approximate Bayesian computational methods." *Statistics and Computing*, 22(6): 1167–1180. MR2992292. doi: https://doi.org/10.1007/s11222-011-9288-2. 1245

Menendez, P., Fan, Y., Garthwaite, P., and Sisson, S. (2014). "Simultaneous adjustment of bias and coverage probabilities for confidence intervals." *Computational Statistics & Data Analysis*, 70: 35–44. MR3125476. doi: https://doi.org/10.1016/j.csda.2013.08.016. 1247

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference." In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 362–369. Morgan Kaufmann Publishers Inc. 1245

Monahan, J. F. and Boos, D. D. (1992). "Proper Likelihoods for Bayesian Analysis." *Biometrika*, 79(2): 271–278. MR1185129. doi: https://doi.org/10.1093/biomet/79.2.271. 1246

Pauger, D. and Wagner, H. (2018). "Bayesian Effect Fusion for Categorical Predictors." *Bayesian Analysis*. Advance publication. URL https://doi.org/10.1214/18-BA1096. MR3934089. doi: https://doi.org/10.1214/18-BA1096. 1263, 1264

Prangle, D., Blum, M. G. B., Popovic, G., and Sisson, S. A. (2014). "Diagnostic tools for approximate Bayesian computation using the coverage property." *Australian &*

New Zealand Journal of Statistics, 56(4): 309–329. MR3300163. doi: https://doi.org/10.1111/anzs.12087. 1245, 1247, 1251

Pritchard, J. K., T, S. M., Perez-Lezaun, A., and Feldman, M. W. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." *Molecular Biology and Evolution*, 16(12): 1791–1798. 1245

Rodrigues, G., Prangle, D., and Sisson, S. (2018). "Recalibration: A post-processing method for approximate Bayesian computation." *Computational Statistics & Data Analysis*, 126: 53–66. MR3808389. doi: https://doi.org/10.1016/j.csda.2018.04.004. 1247, 1248, 1253, 1254, 1255

Rossi, P. and McCulloch, R. (2017). *bayesm: Bayesian inference for marketing/micro-econometrics*. R package version 3.1-0.1. 1261

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). "Validating Bayesian inference algorithms with simulation-based calibration." arxiv.org/abs/1804.06788. 1246, 1247, 1248

Teschendorff, A. E. (2006). *vabayelMix: Variational Bayesian Mixture Modelling*. R package version 0.3. 1260

Therneau, T. and Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-12. 1262

Wood, S. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society (B)*, 73(1): 3–36. MR2797734. doi: https://doi.org/10.1111/j.1467-9868.2010.00749.x. 1250, 1258

Wood, S. N. (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, 466(7310): 1102–1104. 1245

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Yes, but Did It Work?: Evaluating Variational Inference." In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, 5581–5590. 1245, 1246, 1247, 1249

Zhu, W. and Fan, Y. (2018). "A Novel Approach for Markov Random Field With Intractable Normalizing Constant on Large Lattices." *Journal of Computational and Graphical Statistics*, 27(1): 59–70. MR3788301. doi: https://doi.org/10.1080/10618600.2017.1317263. 1256