Diagnostic tools for approximate Bayes

Clarkson, Jason
Oxford

Liu, Bryan
Imperial

Monod, Mélodie
Imperial

Abstract

In this paper we consider the problem of validating the correctness of models and computation in Bayesian analysis. We investigate a number of diagnostic tools — Simulation-based Calibration, Pareto Smoothed Importance Sampling, and Calibrated Approximate Bayesian Inference. We attempt to reproduce the results in the original work, and point out cases where they might work less favourably.

1 Introduction

Consider the usual objects in Bayesian inference: a prior distribution $\phi \sim \pi(\cdot)$, and a posterior distribution $\pi(\cdot|y)$ given an observed data y generated from $p(\cdot|\phi)$. We aim to do inference on the posterior density which describes how the latent variables vary, conditioned on a set of observations y. Often the posterior distribution is analytically intractable, and can only be estimated. Thus, we seek to approximate the posterior. As model and algorithms become more complex, the need to verify the correctness of their computation mounts.

Let's assume that one has formulated a model (i.e., mathematically defined beliefs about unknown parameter θ and generative model). Several options for inferring the posterior are available. They include classical methods such as Markov chain Monte Carlo (MCMC) or faster under model complexity, but less-studied methods, Variational Inference (VI). While the objective, approximating the posterior, remains the same, the procedure to achieve it differs. In this report, we present three diagnostics to evaluate the quality of the Bayes posterior approximation from either an MCMC or a VI. One should note that the diagnostics do not test if the computational methods have converged. This is assumed and on the responsibility of the scientist.

Throughout the report, we use as toy example a Bayesian linear regression, where the data are generated as:

$$\beta \sim \mathcal{N}(0, 10^2), \tag{1}$$

$$\alpha \sim \mathcal{N}(0, 10^2), \tag{2}$$

$$y \sim \mathcal{N}(\alpha + \beta x, 1.2^2), \tag{3}$$

where covariate $x = 0{:}4$. We can then build a Bayesian model to estimate the posterior for β and α given some simulated data.

2 Related concepts

2.1 Markov Chain Monte Carlo

In Markov Chain Monte Carlo (MCMC), we sample from an ergotic chain to collect samples from a stationary distribution that approximates the posterior (Robert and Casella, 2004).

2.2 Variational Inference

Variational inference (VI) consider a family of simple densities and find the member closest to the posterior. Specifically, it optimizes a loss function to obtain density $q(\theta|\psi)$, parametrized by ψ, which is the closest to the posterior. This turns approximate inference into optimization (Blei et al., 2017). Closeness is measured by Kullback-Leibler (KL) divergence defined as,

$$\mathrm{KL}\big(q(\theta|\psi)||\pi(\theta|y)\big) = \int_\Theta \log \frac{q(\theta|\psi)}{\pi(\theta|y)}\, q(\theta|\psi) d\theta \tag{4}$$

Minimizing the KL divergence directly is not possible because of the potentially intractable normalization constant $p(y)$. Instead, it is equivalent to maximizing the evidence lower bound (ELBO):

$$\mathrm{ELBO}(\psi) = \int_\Theta \log \frac{p(\theta, y)}{q(\theta|\psi)}\, q(\theta|\psi) d\theta \tag{5}$$

$$= \int_\Theta \log p(y|\theta) q(\theta|\psi) d\theta - \mathrm{KL}\big(q(\theta|\psi)||\pi(\theta)\big) \tag{6}$$

VI then solves

$$\psi^* = \underset{\psi}{\text{argmax}}\ \text{ELBO}(\psi),$$
$$\text{s.t. supp}(q(\theta|\psi)) \subseteq \text{supp}(p(\theta|y)). \tag{7}$$

where the second line specifies the support matching constraint implied in the KL divergence. By maximizing the ELBO, we encourage the optimization process to choose a candidate distribution which (1) explains the observed data well and (2) is similar to the prior distribution.

2.3 Importance Sampling

Let h be a density from a distribution that is absolutely continuous with π. Then one can write,

$$\pi(x) = \frac{w(x)h(x)}{\int w(x)h(x)dx}, \tag{8}$$

$$\text{with} \quad w(x) = \frac{\gamma(x)}{h(x)}, \tag{9}$$

where γ is the un-normalized distribution of π. We say that h is the importance distribution. Assume we can obtain n i.i.d samples $X_i \sim h$, then we approximate,

$$\hat{\pi}(\varphi(X)) = \sum_{i=1}^{n} W_i \varphi(X_i) \quad \text{and} \tag{10}$$

$$\hat{\pi}(dx) = \sum_{i=1}^{n} W_i \delta_{X_i}(dx), \tag{11}$$

where $W_i = \frac{w(x_i)}{\sum_{j=1}^{n} w(x_j)}$.

3 Simulation-based Calibration

We begin by introducing Simulation-Based Calibration (SBC), a diagnostic procedure that evaluates whether posterior samples generated from any Bayesian algorithms are reflective of the exact posterior distribution.

Consider the Bayesian computation example in Section 1, where we compute samples from the prior $\phi \sim \pi(\cdot)$, likelihood conditioned on the sampled prior $y \sim p(\cdot|\phi)$, and the estimated posterior $\theta \sim \tilde{\pi}(\cdot|y)$. (Cook et al., 2006) observed the self-consistency between an data-averaged posterior[1] and the prior distribution: assuming the posterior $\pi(\cdot|y)$ is exact, integrating it over all possible sampled data (likelihood) and ground truth (prior) should return the prior distribution:

$$\pi(\theta) = \int \pi(\theta|y)p(y|\phi)\pi(\phi)\,\mathrm{d}y\mathrm{d}\phi. \tag{12}$$

[1]Average of the posteriors w.r.t. to the generated data

(Talts et al., 2018*a*) further proved that the self-consistency implies that the rank[2] of the prior sample ϕ relative to the exact posterior samples $\{\theta_1, \cdots, \theta_L\} \sim \pi(\cdot|y)$ will follow a discrete uniform distribution, as they are identically distributed. They argue that the observation can be used to guard against inaccurate computation of the posterior or misimplemented models during Bayesian analyses. The errors will lead to a discrepancy between the data-averaged (estimated) posterior and the prior distributions, and hence non-uniformity in the ranks.

The observations leads to the development of SBC: In each iteration we draw, in turn, 1) a prior sample, 2) simulated data conditioned on the prior sample, and 3) L posterior samples conditioned on the data. We then compute the rank of the prior sample relative to the posterior samples as described above. This is repeated N times to obtain an empirical distribution of the ranks for each parameter of interest.[3] Talts et al. suggested visualising the rank distributions with histograms, where a symmetric $\cap$-shape, a symmetric $\cup$-shape, and asymmetry in the rank distributions indicates the data-averaged posterior is overdispersed, underdispersed, and biased relative to the prior distribution respectively.

3.1 Implementation & reproduction of key results

We implemented the modified SBC (Algorithm 2 in (Talts et al., 2018*a*)[4]) in `R`,[5] and run it using the linear regression example in Section 1. In the case where the prior is correctly specified in the inference model (i.e. same as that in the data generation process), we are able to obtain a similar histogram to that shown in the original work (see Figure 1).

Moreover, we are able to reproduce the $\cup$-shape and asymmetric rank distributions featured in the original work in Figure 2. This is obtained when the prior we used to build our posterior (model prior) is severely underdispersed and biased, respectively, than that used to generate prior samples for the linear regression example (data prior). This leads to an underdispersed and biased data-averaged posterior even in the presence of the data.

[2]# of posterior samples less than the prior sample.

[3]For multi-dimensional inference, we inspect the rank distribution of each parameter in turn, potentially reusing the computed posterior samples.

[4]The modified procedure aims to combat autocorrelation in the posterior samples generated in the underlying MCMC chain by uniformly thinning the samples. This generally removes spikes in the histogram that is not caused by model misspecification.

[5]Using `rstan` to obtain posterior samples.

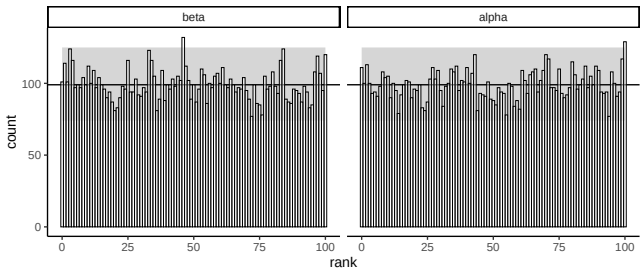
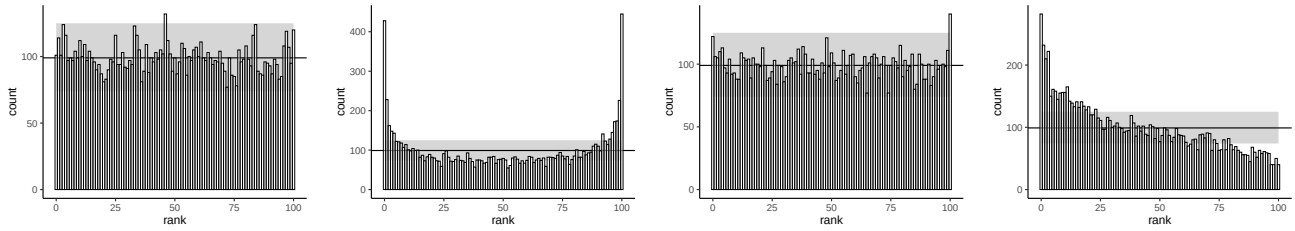


Figure 1: The rank distribution produced by Simulation-Based Calibration (SBC) when the model is specified correctly. Both the model and data generating process assumes $\beta, \alpha \sim \mathcal{N}(0, 10^2)$, and the latter generates five data points under $y \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$ for some one-dimensional covariate X before we fit the model to obtain a posterior. 10,000 rank samples are generated in the procedure, with each prior sample compared against 100 posterior samples to obtain one rank sample. The grey shaded area represents the 99% interval expected from a uniform histogram.



(a) $\beta \sim \mathcal{N}(0, 10^2)$ (b) $\beta \sim \mathcal{N}(0, 2^2)$ (c) $\beta \sim \mathcal{N}(0, 100^2)$ (d) $\beta \sim \mathcal{N}(100, 10^2)$

Figure 2: The rank distributions under Simulation-Based Calibration (SBC) for one parameter, when the prior used to build our posterior (model prior) is the same and different than the one used to generate the data for that particular parameter. We define the data generating process as $\beta, \alpha \sim \mathcal{N}(0, 10^2)$ and $y_{1:5} \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$, with X, a one-dimensional covariate, taking five different values. We modify the model prior by changing the distribution of β as shown in the sub-captions, while keeping α. SBC is able to produce a symmetric $\cup$-shape and asymmetric rank distribution when the model prior is underdispersed and biased respectively. In the case where the model prior is overdispersed, SBC returns a uniform rank distribution.

We are unable to show the ∩-shape when the model prior is overdispersed though. We believe this is due to the misspecified model prior being uninformative and thus quickly "corrected" by the simulated data (which are generated from the correct data prior) to produce a posterior similar to the data prior. This observation suggests that SBC is not set up to detect such inconsistency, as it is not wrong to be uncertain on your initial belief.

3.2 Potential gaps

While the results above appear to be promising, we also encounter a number of cases that indicate potential gaps in SBC. The gaps may prevent SBC to become a practically useful diagnostic tool, where we have "automated diagnostics that can flag certain parameters for closer inspection"(Talts et al., 2018*a*). Further work is required to understand the underlying cause(s) and provide workaround(s).

We believe SBC may not be able to determine specifically which model parameter(s) is/are misspecified. One would expect that when we misspecify the model prior for one parameter, it would only affect the rank distribution associated to that parameter but not that of other parameters, so that we can focus our investigation on the misspecified parameter while leaving other ones alone. Figure 3 shows this might not the case for the linear regression example: when we intentionally make the model prior for β underdispersed / biased, the rank distributions for α also suggests underdispersion / bias in the data-averaged posterior.[6] The observation suggest SBC at its current form may flag too many false positives to be useful, and might lead to practitioners attempting to "wrong a right" while the true problem lies elsewhere.

Moreover, we believe the detection power of SBC may be limited in applications where the model prior is misspecified and a large number of observed data are involved. This is based on the observation where increasing the number of data points (y) simulated when we obtain each rank sample appears to stabilise the rank distribution. The stabilisation effect is shown in Figure 4 using the linear regression example with different number of data points simulated per rank sample. This is concerning as any error introduced in model specification can be masked by such stabilisation, and more work is required to understand the effect's impact.

Finally, we note the authors observing that global summaries are a natural but ineffective options, due to the fact that deviation from uniformity tends to occur in only a few systematic ways for rank distributions, as mentioned earlier. It will be interesting to see if a statistical test utilising beta-binomial distributions (the discrete version of the Beta(α, β) distribution) can step up to the task. In the case where the rank distribution is uniform, we should expect that $\alpha = \beta = 1$. Otherwise we should expect to see $\alpha \neq \beta$, $\alpha = \beta < 1$, and $\alpha = \beta > 1$ for biased, underdispersed, and overdispersed data-averaged posterior respectively, as compared to the true prior.

[6] It is worth noting that the rank distribution for α suggest the bias of the data-averaged posterior goes in the opposite direction than that of β, perhaps as an attempt by the model to compensate for the bias in β.

4 Pareto Smoothed Importance Sampling

Secondly, we introduce the Pareto Smoothed Importance Sampling (PSIS) diagnostic, which could be view as a post-adjustment for VI approximation (Yao et al., 2018). The best estimator $q(.|\psi^*)$, estimated by VI, is used as a proposal for importance sampling.

4.1 Importance sampling captures the divergence between the target and proposal

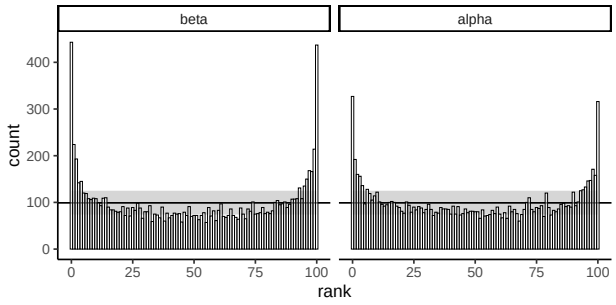
Using IS allows to harness the sampling weights' essence as a way to evaluate the closeness between the target and the proposal distribution. In our context the importance weights defines in 2.3 are,

$$w_s = \frac{p(\theta_s, y)}{q(\theta_s|\psi^*)} \tag{13}$$

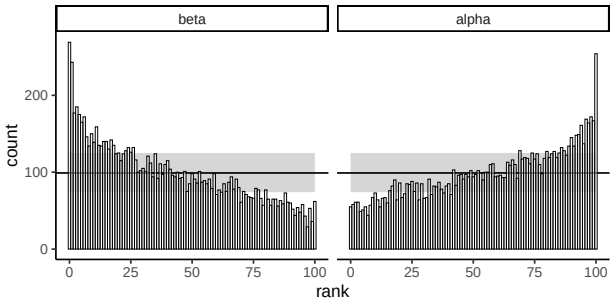
for $\theta_s \in \{\theta_1, ..., \theta_S\}$ evaluation of the proposal. They capture 'how close' the proposal is from the target. Intuitively, if w_s is around one, the two distributions are close at θ_s. Conversely, extreme w_s (towards zero or infinity) indicates a discrepancy at this point. If the later behavior is often observed, the importance weights distribution may have a few finite moments. The success of plain importance sampling depends entirely on how many moments the importance ratios w_s possess. Especially, an IS estimator has finite variance if the second moments of the importance weights distribution is finite. The latter is also a necessary condition for the Lindeberg-Lévy central limit theorem to holds (Koopman et al., 2009). However, the existence of the variance of the importance weights is by no means guaranteed.

4.2 Using PSIS to estimate the number of importance weights finite moments

From the original idea of Koopman et al. (2009), revisited in Vehtari et al. (2017) and Vehtari et al. (2015),

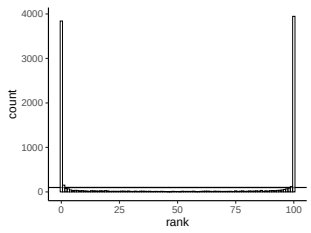


(a) $\beta \sim \mathcal{N}(0, 2^2)$

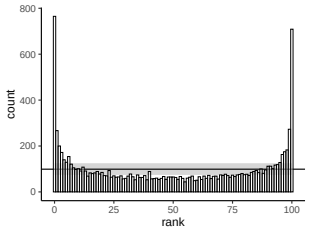


(b) $\beta \sim \mathcal{N}(100, 10^2)$

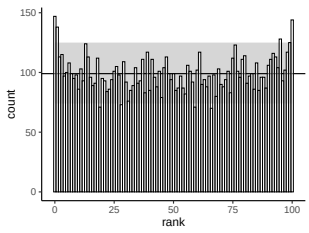
Figure 3: The rank distributions under Simulation-Based Calibration (SBC) for all parameters, when one parameter of the prior used to build our posterior (model prior) is different to the one used to generate the data (data prior). We define the data generating process as that in Figure 2, and modify the model prior by only changing the distribution of β as shown in the sub-captions. One would expect the histograms for α show the ranks are uniformly distributed, but that is not the case.



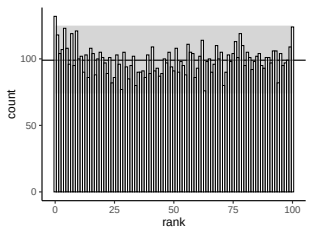
(a) $s = 2$



(b) $s = 10$



(c) $s = 25$



(d) $s = 50$

Figure 4: The rank distributions under Simulation-Based Calibration (SBC) for a misspecified parameter, when we vary the number of simulated data generated per rank sample (s). We define the data generating process as $\beta, \alpha \sim \mathcal{N}(0, 10^2)$ and $y_{1:s} \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$, with X, a one-dimensional covariate, taking s different values. We make β underdispersed in the prior used to build our posterior by specifying $\beta \sim \mathcal{N}(0, 1^2)$. One can see as s increase, the rank distribution converges to a uniform distribution.

a 'smoothing procedure' can be applied on the importance weights to estimate the number of existing moments of their upper tail distribution. Specially, the latter is fitted with a Generalized Pareto distribution (GP). The GP density is,

$$p(y|u,\sigma,k) = \begin{cases} \frac{1}{\sigma}\big(1 + k(\frac{y-u}{\sigma})\big)^{-\frac{1}{k}-1}, & k \neq 0 \\ \frac{1}{\sigma}\exp\big(\frac{y-u}{\sigma}\big), & k = 0 \end{cases} \quad (14)$$

where u is a lower bound parameter, y is restricted to the range (u, ∞), $\sigma \in \mathbb{R}^+$ is a non-negative scale parameter, and $k \in \mathbb{R}$ is an unconstrained shape parameter. Inference procedure fit the Pareto distribution for the M largest sample defined in Vehtari et al. (2015) as follow,

$$M = \begin{cases} 3\sqrt{S}, & S > 225, \\ S/5, & S \leq 225. \end{cases} \quad (15)$$

We fix parameter u by restraining the weights with a lower bound and fit a GP to model the distribution of $w_s|w_s > u$. It is not restraining to focus only on the large weights because two distributions sharing discrepancies points have a mix of very small and very large weights (i.e, if there exist a point on the support where one attribute more probability mass than the other, then there must be another point for which the opposite is true.) if the sample size is large enough. GP possesses the interesting property of having $\lfloor 1/k \rfloor$ finite moments when $k > 0$. By examining the shape parameter of the fitted Pareto distribution, denoted $\hat{k}$, we can obtain sample-based estimates of the existence of the moments (Koopman et al., 2009). Indeed, $\hat{k}$ approximates

$$k = \inf\Big\{k' > 0 : E_q\Big[\Big(\frac{p(\theta|y)}{q(\theta|\psi^*)}\Big)^{1/k'}\Big] < \infty\Big\}. \quad (16)$$

Note here that it is equivalent to use the conditional distribution because $p(y)$ is a constant. Having this estimate allow us to infer whether the tails of the importance ratio are heavy, or equivalently if the target and the proposal distributions are close. Indeed, this concept is related to VI optimization as we can express the estimator in terms of divergence between the distributions. Taking the logarithm and discarding the constant gives the Rényi divergence (Rényi and et al, 1961),

$$k = \inf\Big\{k' < 0 : D_{1/k'}(p||q) < \infty\Big\} \quad (17)$$

$$\text{where } D_\alpha(p||q) = \frac{1}{\alpha - 1}\, log \int_\Theta p(\theta|y)^\alpha\, q(\theta|\psi^*)^{1-\alpha}\, d\theta. \quad (18)$$

The Rényi divergence is monotonic increasing of order α. That is, the divergence between the true posterior and VI approximation decreases with the existence of subsequent moments (i.e., $1/k$ decreases). When $k = 1$ (i.e., only the mean exist), (1) the Rényi divergence is equal to the KL divergence and (2) the divergence is infinite - we conclude that the approximation from VI is terrible. For other cases, Vehtari et al. (2017) and Yao et al. (2018) have defined the following thresholds:

- $\hat{k} < 0.5$: The central limit theorem holds suggesting that PSIS converges at relatively quick rate $N^{1/2}$. The VI posterior approximation is close enough to the true posterior.
- $0.5 < \hat{k} < 0.7$: The time until convergence is finite, indicating useful IS samples. The VI posterior approximation is not perfect but can be helpful.
- $0.7 > \hat{k}$: IS samples should not be trusted with large weights dominating estimations. The VI posterior approximation is not good.

4.3 Implementation on a linear regression

Yao et al. (2018) investigated the quality of VI estimation for linear regression under different optimisation tolerance levels. They showed that $\hat{k}$ can be viewed as a convergence test. In this report, we explore the behavior of the diagnostics, for different model specification, under the linear regression framework presented in Section 1. We proceed as follow:

1. We use `ADVI` (Automatic differentiation variational inference) from `stan` package (version 2.19.1) (Kucukelbir et al., 2015) to approximate the posterior using VI. The algorithm optimizes the variational objective under the Gaussian distribution family. The threshold of relative ELBO is set to a conservative value of 10^{-5} (default is 10^{-2}) and η to 0.05. The former produced the best results in Yao et al. (2018).
2. We compute the importance sampling weights w_s in (13) for $S = 10^5$ posterior evaluations.
3. We fit the Generalized Pareto distribution with the `loo` package (Vehtari et al., 2017, Vehtari et al., 2015) and obtain $\hat{k}$.
4. We compare the empirical mean of the VI posterior evaluation on the parameter β to the true posterior mean[7] (that is tractable) by evaluating the Root Mean Square Error (RMSE).

[7] true posterior refers to the posterior obtained from the correct model specification defined in Section 2.3.

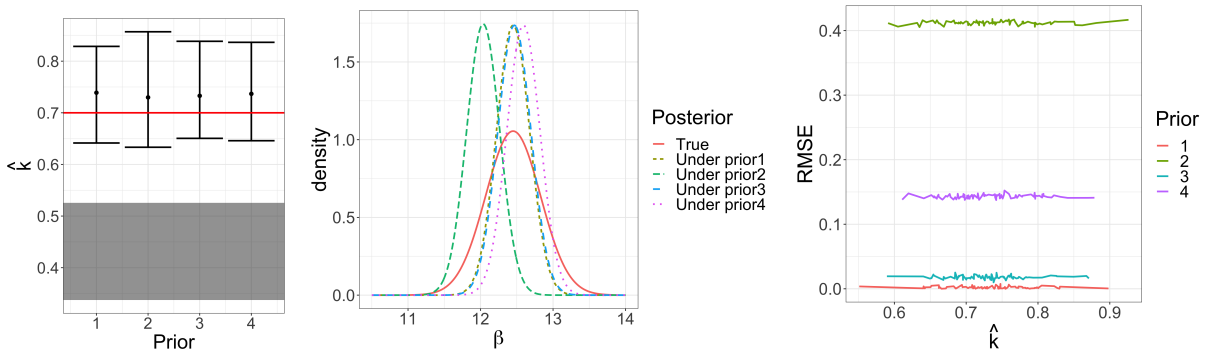


Figure 5: We explore the PSIS diagnostics under VI optimization for the bayesian linear model presented in Section 1. The prior of α is fixed to a gaussian $0, 10$. We compare the prior under four prior specification for β: (1) $\beta \sim \mathcal{N}(0, 10)$, (2) $\beta \sim \mathcal{N}(0, 2)$, (3) $\beta \sim \mathcal{N}(0, 100)$, (4) $\beta \sim \mathcal{N}(100, 10)$. The left-hand panel shows the value of $\hat{k}$ obtained with the 10^5 posterior evaluation for the four model specifications. Specifically, we plot the median (dot) and 95% credible interval (line) for 100 simulations. The dashed area represents the credible interval of the diagnostics evaluated under the true posterior. The second panel shows VI posterior approximation of β compared to the true posterior. Lastly, the right-hand panel presents the RMSE of β against $\hat{k}$.

We use different priors for β: (1) $\beta \sim \mathcal{N}(0, 10)$, (2) $\beta \sim \mathcal{N}(0, 2)$, (3) $\beta \sim \mathcal{N}(0, 100)$, (4) $\beta \sim \mathcal{N}(100, 10)$. The same prior on α is used, $\alpha \sim \mathcal{N}(0, 10)$. To account for the uncertainty of stochastic optimization and $\hat{k}$ estimation, steps above are repeated 100 times for each prior.

We do not find the same estimates $\hat{k}$ as in Yao et al. (2018) because our version of the `loo` package is more recent than theirs (R package version 2.1.0 rather than 2.0.0).

The left-hand panel shows that the estimated Generalized Pareto shape parameter $\hat{k}$ under VI approach is not significantly different from 0.7 and is nowhere near the estimation under the true posterior (grey area). We conclude that VI approximation should not be trusted under this straightforward framework. Talts et al. (2018*b*) also found that ADVI failed under this scheme.

Like the VI approximation, the true posterior is a Gaussian. This gives an opportunity, in the middle panel, to investigate which of the two optimized parameters is responsible for the distributions discrepancy. Two approximations, under prior 2 and 4, are off the true mean. This behavior was expected as the optimization encourages support close to the prior. Indeed more probability mass is concentrated around 0 in the first case, so the mean is under-estimated ; and around 100 in the second case, so the mean is over-estimated. This behaviour is due to model mispecification and should lessen with more observations. However, the mean of the posterior is not responsible for the bad diagnostics. Indeed, the right-hand panel shows that the RMSE does not depend on the value of $\hat{k}$. In other words, a good location estimator does not produce a smaller value of $\hat{k}$.

In the middle panel, we also observe that all VI approximations have lighter tails than the true posterior. This is symptomatic of the optimization objective, the KL divergence in (4). In fact, the optimization penalizes placing mass on $q(.|\psi)$ where $\pi(.|y)$[8] has little mass. Therefore, VI approximation is compacted around modes. The discrepancy in the tails engenders huge importance sampling weights, and result in a great shape parameter. The VI approximation should not be trusted because it gives over-confident parameters. It should be noted that model mispecification does not seem to impact the posterior variance.

In this example model mispecification does not worsen the diagnostics. A possible reason is that the rate difference at which the tails have to decrease for the existence of two subsequent moments is not constant. Loosely speaking, losing the second moment requires less extreme observation than losing the first moment. The intuition can be obtained by looking at the GP distribution cdf for different k.

First, the bad diagnostics are not caused by the approximated mean of the Gaussian but by its variance. In this example, to successfully match the marginal variances and achieve better diagnostics, $q(.|\psi)$ would have to expand into territory where $\pi(.|y)$ has little mass. Second, the diagnostics fails to notice model mi-specification. Instead it detects a systematic behavior of VI and stagnates just above 0.7. Third, the prior has on influence on the mean but not on the variance of the approximated posterior. Fourth, the thresholds defined in Vehtari et al. (2017) and Yao et al. (2018) are over constraining. Indeed even the diagnostics evaluated under the true posterior has a

[8]This is the approximated posterior dependant on the prior specified

significant chance to be greater than 0.5. Lastly, we wonder if this diagnostics is punitive enough for VI. Indeed, the proposal evaluation are concentrated around its approximated posterior mode, where the target has a smaller density (i.e., the weights are small). If points were extended to the tails of the target, we would obtain much extreme observation and perhaps higher k.

5 Calibrated Approximate Bayesian Inference

We now turn our attention to specific case where the goal is to compute a posterior-credible set for the parameter using an appromixation $\tilde{\pi}(\cdot|y)$. The use of an approximation will likely introduce some error into our estimate of a credible set. MCMC based methods only produce samples from the approximate posterior, meaning credible intervals must be approximated via Monte-Carlo sampling methods; This introduces a further source of bias. The goal of is to estimate the bias in the coverage of a credible interval computed using these approximations compared to one computed using the true posterior. The paper under examination in this text is (Xing et al., 2019).

5.1 Definition of the Problem

We define the exact α-level credible set as the set C_y such that:

$$\alpha = \int_{\Omega} \chi_{\phi \in C_y} \pi(\phi|y) d\phi \tag{19}$$

The credible set covers the true parameter ϕ with probability α if ϕ is drawn from the prior, and the data really was generated using the observation model we have selected.
Of course, in practice we usually do not have access to the true posterior nor can we compute exact credible sets. In such a situation one usually proceeds by taking J samples $\underline{\theta} = (\theta_1, \dots, \theta_J)$, $\theta_j \sim \tilde{\pi}(\cdot|y)$ and then uses them to compute an estimate $\tilde{C}_y(\underline{\theta})$. The quantity we are interested in is the realised coverage of our approximate credible set (all Monte-Carlo error included), that is $c(y) := \mathbb{P}(\phi \in \tilde{C}_Y(\underline{\theta})|Y = y)$.

We start by noticing the conditional distribution of $(\phi, \underline{\theta}|y)$ in the generative model is given by:

$$m(\phi, \underline{\theta}|y) = \pi(\phi|y)\tilde{\pi}(\underline{\theta}|y) \tag{20}$$

writing $\tilde{\pi}(\underline{\theta}|y)$ for the joint distribution of $\underline{\theta} \in \Omega^J$. We can use m to write $c(y)$ as an expectation:

$$c(y) = \int_{\Omega^J} \int_{\Omega} \chi_{\phi \in \tilde{C}_Y(\underline{\theta})} m(\phi, \underline{\theta}|y) d\phi d\underline{\theta} \tag{21}$$

This expectation involves the true posterior, something we do not have access to. However we can estimate expectations with respect to densities we know up to a normalizing constant using Monte Carlo methods (recall we have access to the likelihood and prior).

5.2 Annealed Importance Sampling

The method of Annealed Importance Sampling (AIS) defined in (Neal, 1998) merges the ideas of MCMC and Simulated Annealing. A motivating example for AIS is the case where the posterior density is multi-modal. If the modes are sparse and peaked, traditional MCMC algorithms may stay within the mode they are initialized closest to and hence will not provide accurate samples. The idea of annealed importance sampling is to 'melt' the target distribution and instead begin by proposing samples from some other 'helpful' distribution that may provide better initial mixing, then gradually cool the temperature, moving back towards the target and providing final samples from regions of high probability in the desired distribution.

Suppose the goal is to sample from a distribution $p_0(x)$. We define a sequence of annealed distributions $p_j(x) \propto p_0(x)^{\beta_j} p_n(x)^{1-\beta_j}$, $1 = \beta_0 > \cdots > \beta_n$. Note that initially we are only proposing from $p_n(\cdot)$. On each iteration of the algorithm we start at $p_n(x)$ and work backwards. The algorithm runs an MCMC chain targetting $p_j(x)$ for a number of steps, using the final state of the j^{th} chain $x^{(j)}$ as the initial state of the $(j-1)^{th}$ chain. Starting with $p_n(x)$ we progressively decreases the temperature, concentrating the probability mass again around the modes and ending up back at $p_0(x)$, the real target and a sample $x := x^{(0)}$

To make it explicit: the algorithm runs a chain targetting each $p_j(x)$ for $j = n, \dots, 0$ to produce a *single sample*, and we repeat this whole procedure many times to produce a set of samples. If we run this procedure several times, due to the (hopefully) easier exploration from p_n in the early phases we expect to see the final chains finish at different points across the sample space, for example in distinct modes, and hence overall better mixing.

The proof that the method is valid (i.e. produces samples from $p_0(\cdot)$) comes from viewing each iteration as an importance sampling method on the extended state space $(x^{(0)}, \dots, x^{(n)})$. One can show that (under regularity conditions) the samples/weights produced by AIS are valid importance samples from the true joint density of the state space (and so marginally $x^{(0)} \sim p_0(\cdot)$ as required).

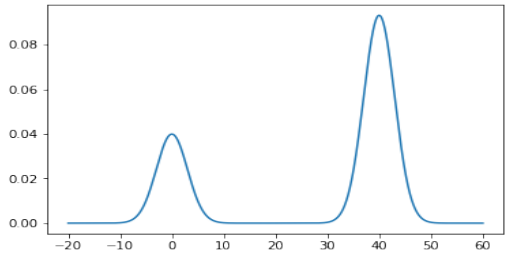


Figure 6: Plot of $f(x)$

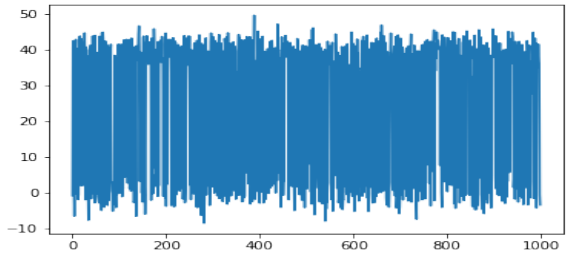


Figure 8: AIS trace-plot for 1000 samples

5.3 Numerical Results

To illustrate the effectiveness of the method, we run the algorithm on a (somewhat contrived) example of a Gaussian mixture model with sparse modes. Suppose we want to sample $X \sim f(\cdot)$, where:

$$f(x) = 0.3\,\mathcal{N}(x; 0, 3) + 0.7\,\mathcal{N}(x; 40, 3) \tag{22}$$

We run a simple random walk Metropolis-Hastings algorithm targeting this density, initialising the chain at $x_0 = 1$ with proposal variance $\sigma_p^2 = 2$.

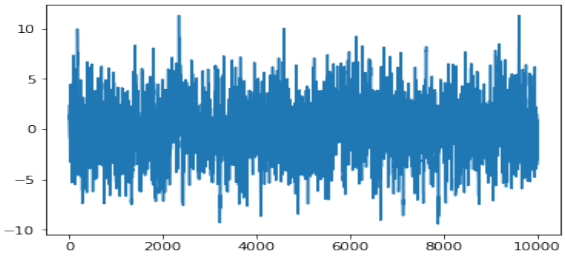


Figure 7: Simple MH trace-plot for 10000 samples

We see from the traceplot in figure 7 the chain never finds the other mode at 40, reporting a poor estimate of the mean $\hat{\mu} = 0.044$.

We now sample from f using an AIS scheme, where we cleverly choose our initial distribution $f_n(x) = \mathcal{N}(x; 20, 2)$. This is in the middle of the two modes, so the idea is that some chains will go left to the smaller mode around zero and some will go right towards the larger.

It is clear from the trace-plot shown in figure 8 that the AIS scheme has much better mixing properties (it explores both modes), providing a much better estimate of the mean at $\hat{\mu} = 23.96$.

5.4 Computing the Approximate Coverage

The novel idea in the paper is to apply AIS to compute the integral in 21. The reason this is helpful is we can use our approximation $\tilde{\pi}$ as our 'cleverly' chosen $p_n(\cdot)$. If our approximation is close to the true posterior then one would expect to see good convergence properties. The authors prove that the Markov Chain given by their specific choices for $p_j(\cdot)$ give importance samples that amount to a consistent estimator for $c(y_{obs})$.

5.5 Evaluation of the Method

The method appears to solve several problems associated with the previous work in Lee et al. (2018), in particular that the samples produced have a much higher ESS. Other technical simplifications allow for computational speedup. There are however some disadvantages. Firstly the goal is to estimate the bias induced by approximation, yet the estimate is itself another approximation and so to be rigorous one should also adjust for the error in the approximation of the error of the approximation, and so on and so forth. There is also the issue that if the posterior approximation is poor, then the starting point for AIS will not introduce useful information and may actually lead to slow mixing.

6 Conclusion

We investigate a number of diagnostics tools designed to detect miss-specified models and compute the error in the implementation of Bayesian algorithms. They include Simulation Based Calibration, Pareto Smoothed Importance Sampling, and Calibrated Approximate Bayesian Inference. We are able to reproduce a number of key result presented in the original works while encountering a number of edge cases using a simple Bayesian linear regression example. Further work is required to understand the underlying causes for these edge cases, and if any workaround is required and/or exists.

References

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.

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**(3), 675–692.
URL: *https://doi.org/10.1198/106186006X136976*

Koopman, S. J., Shephard, N. and Creal, D. (2009), 'Testing the assumptions behind importance sampling', *Journal of Econometrics* **149**(1), 2–11.

Kucukelbir, A., Ranganath, R., Gelman, A. and Blei, D. M. (2015), 'Automatic Variational Inference in Stan', *arXiv e-prints* p. arXiv:1506.03431.

Lee, J. E., Nicholls, G. K. and Ryder, R. J. (2018), 'Calibration procedures for approximate bayesian credible sets'. Advance publication.
URL: *https://doi.org/10.1214/19-BA1175*

Neal, R. M. (1998), 'Annealed importance sampling'.

Rényi, A. and et al (1961), 'On measures of entropy and information. 547-561.', *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* .

Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer Texts in Statistics.

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

Talts, S., Betancourt, M., Simpson, D., Vehtari, A. and Gelman, A. (2018*b*), 'Validating bayesian inference algorithms with simulation-based calibration'.

Vehtari, A., Gelman, A. and Gabry, J. (2017), 'Practical bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and Computing* **27**(5), 1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. and Gabry, J. (2015), 'Pareto Smoothed Importance Sampling', *arXiv e-prints* p. arXiv:1507.02646.

Xing, H., Nicholls, G. and Lee, J. (2019), Calibrated approximate Bayesian inference, *in* K. Chaudhuri and R. Salakhutdinov, eds, 'Proceedings of the 36th International Conference on Machine Learning', Vol. 97 of *Proceedings of Machine Learning Research*, PMLR, Long Beach, California, USA, pp. 6912–6920.
URL: *http://proceedings.mlr.press/v97/xing19a.html*

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018), 'Yes, but Did It Work?: Evaluating Variational Inference', *arXiv e-prints* p. arXiv:1802.02538.

Appendix

A Self-consistency of the data-averaged posterior and the prior

In this section, we provide a step-by-step derivation of the self-consistency property of the data-averaged posterior and the prior, as described in (Cook et al., 2006) and (Talts et al., 2018*a*). The authors observed that integrating the exact posterior $\pi(\cdot|y)$ over all priors $\phi \sim \pi(\cdot)$ and their associated likelihood $y \sim \pi(\cdot|\phi)$ (i.e. all possible ground truths) returns the prior distribution:

$$\pi(\theta) = \int \pi(\theta|y)p(y|\phi)\pi(\phi)\,\mathrm{d}y\,\mathrm{d}\phi \tag{23}$$

We begin the derivation from the RHS, applying Bayes' theorem to the first two terms we have

$$\int \frac{p(y|\theta)\pi(\theta)}{p(y)}\frac{\pi(\phi|y)p(y)}{\pi(\phi)}\pi(\phi)\,\mathrm{d}y\,\mathrm{d}\phi. \tag{24}$$

Note the $\pi(\phi)$s and the $p(y)$s cancel, and we can rearrange the terms to obtain

$$\int \pi(\theta|y)p(y|\phi)\pi(\phi)\,\mathrm{d}y\,\mathrm{d}\phi, \tag{25}$$

which is the joint distribution of θ, y and ϕ:

$$\int \pi(\theta, y, \phi)\,\mathrm{d}y\,\mathrm{d}\phi. \tag{26}$$

Integrating over ϕ and y we obtain the marginal as $\pi(\theta)$, which is the LHS of Equation (23).