Diagnosing embedded Laplace approximation

Hyunji Moon

October 21, 2020

Abstract

1 Diagnosing approximation algorithm with SBC and PSIS

We aim to introduce a framework for validating and correcting approximation algorithm. Simulation-based calibration (SBC) (and PSIS further) is applied on embedded Laplace approximation. SBC is one of the few tools that could examine choice of computational method (Talts et al., 2018). Two things should be noted for SBC. First, SBC validates the combination of approximate algorithm, model, and prior. Therefore, when SBC test fails, non-unifrom SBC rank histogram for example, it does not pinpoint the cause of the problem; only the fact that something is wrong. Algorithm, likelihood, and prior should be tested with an extra care on their dependency which requires principled workflow as in Gabry et al. (2019) and Schad et al. (2020). As this paper is focused on calibrating the algorithm, prior and likelihood model are fixed, but the fact that SBC test results on algorithm validity corresponds to only the subset of the entire model space should be noted. Second, to do SBC, reasonably informative priors are needed to avoid simulating unrealistic parameter values. SBC tests not only likelihood but also prior; and it is highly prior-sensitive. In real world data-fitting situation, the existence of data offsets the effect of mis-specified prior. But in SBC, data itself is generated from the prior upon which the model is fitted to retrieve the parameters. This relation between simulated dataset and retrieved parameters would be elaborated in section 2.2. Example of setting the SBC prior could be found in Talts et al. (2018) where gaussian process priors were specified to ensure $\Pr(\rho < 0.1) = \Pr(\sigma > 1) = 0.1$ given the data, to test INLA.

Probabilistic nature of this calibration has advantages to the worst-case frequentist calibration: it's easier to (approximately) implement in practice using simulations and it focuses on the model configurations consistent with our prior model instead of more extreme model configurations. However, its ability to test the average performance on multiple components (algorithm, prior, and likelihood) is not always helpful, especially when the aim is to test a specific

parameter settings and data. Good performance on average might not hold for a particular realization of the data and if the observed data is not well represented by the model, diagnostic would be ineffective. This is the price SBC pays for averaging over all priors (Yao et al., 2018) which leads to the need for PSIS validation where diagnosis is more specificed in terms or data and parameters.

Margossian et al. (2020) proposed an algorithm that efficiently differentiate the Laplace approximation and showed the benefits of running HMC only on the hyperparameter space by marginalizing out the parameters. This could achieve both greater speed and less tuning efforts, and therefore accuracy, compared to the standard model especially for latent gaussian models with Poisson (section 2)and Bernoulli (section 3) likelihood. SBC could diagnose the validity of this algorithm, which we would refer to as embedded Laplace approximation, and further provide a feedback for improvement. Embedded Laplace approximation aims to approximate the full HMC model where HMC is run on joint hyperparameter and parameter space. The process of SBC for approximate algorithm is as follows. First, parameters are simulated from the given prior. Second, data is simulated based on the likelihood of standard model and its simulated parameter. Third, simulated data is fitted with approximation algorithm which leads to posterior. Lastly, posterior and prior sample is compared.

$$
\begin{aligned}
&\tilde{\theta_0} \sim \pi(\theta) \\
&\tilde{y} \sim \pi(y \mid \tilde{\theta_0}) \\
&\left\{\tilde{\theta}_1, \ldots, \tilde{\theta}_L\right\} \sim \pi_G(\theta \mid \tilde{y}) \\
&r\left(\left\{f\left(\tilde{\theta}_1\right), \ldots, f\left(\tilde{\theta}_L\right)\right\}, f(\tilde{\theta})\right) = \sum_{l=1}^{L} \mathbb{I}\left[f\left(\tilde{\theta}_l\right) < f(\tilde{\theta})\right] \in [0, L]
\end{aligned}
\tag{1}
$$

When $\pi\left(\theta_l \mid y\right) = \pi\left(\theta_0 \mid y\right)$, the rank statistics are proven to be uniform (Appendix 4.1). In other words, degree of SBC histogram uniformity would indicate the accuracy of the approximate algorithm: posterior inference. Any discrepancy of its inference would affect the rank ($\int \mathrm{d}u(y)[u(y)]^r[1-u(y)]^{L-r}$ term from the proof) with the manifestation of its type, such as the direction of bias and dispersion.

2 Gaussian process with a Poisson likelihood

2.1 SBC test results

Vanhatalo et al. (2010) models the mortality count across Finland and out of 911 countries we have randomly selected 20 counties for SBC. As prior settings are solely the responsibility of the modeler, different distribution, location, and scale of α and ρ priors are experimented. Slight difference was observed but for clarity, we suggest some representative results; see Appendix 4.2 for SBC histogram on different prior settings. Although, some ragged patterns are observed, both

hyperparameters (α and ρ) and parameter (θ) are within the bound of credible interval (Figure 1). The interval signifies the 0.005 quantile to the 0.995 quantile of the Binomial $\left(N, (L+1)^{-1}\right)$ distribution. Bins is the number of bins whose values is 20, following the recommendation of (Talts et al., 2018). It is expected that, on average, the counts in only one bin a hundred will deviate outside this shaded interval. However, when observed with empirical cumulative distribution function (ECDF) which helps detect the deviation at both extremes (Talts et al., 2018), slight deviation was observed for θ, the latent parameter. This bias implies embedded Laplace approximation tends to underestimate the parameter. Note that the intervals are based on repeated predictive values from uniform distribution (Talts et al., 2018).

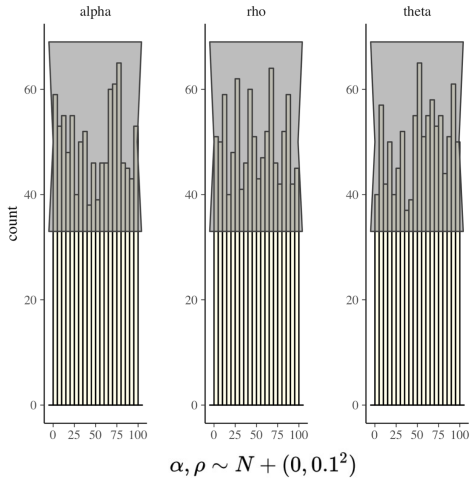

Figure 1: SBC histogram for disease map data with half-Normal prior

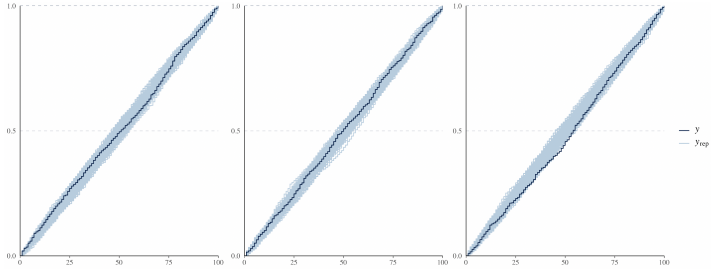

Figure 2: SBC ECDF for disease map data with half-Normal prior

2.2 Understanding SBC with simulated y and parameter

Figure 3 shows twenty sets of simulated y and the original mean y_e (red) from $\pi(y_i \mid \theta) = \text{Poisson}\left(y_e^i e^{\theta_i}\right)$. Poisson likelihood model with half normal prior $N + (0, 1^2)$ for α and ρ was used to simulate y. Extreme realized value from simulation set 7 (green) and 20 (orange) are highlighted. Table 1 compares fitting time, log-likelihood, and retrieved parameter values, for three different simulated y datasets; set with short fitting time and two sets where extreme y values were simulated. Fitting time varied in scales of hundreds (shortest: 0.4s, longest: 76s) and much higher log probability was observed for the set with shorter fitting time. Simulation set with the most extreme value had highest α and lowest ρ. As large α and small ρ contribute to higher frequency outcomes, retrieved hyperparameter values are understandable.

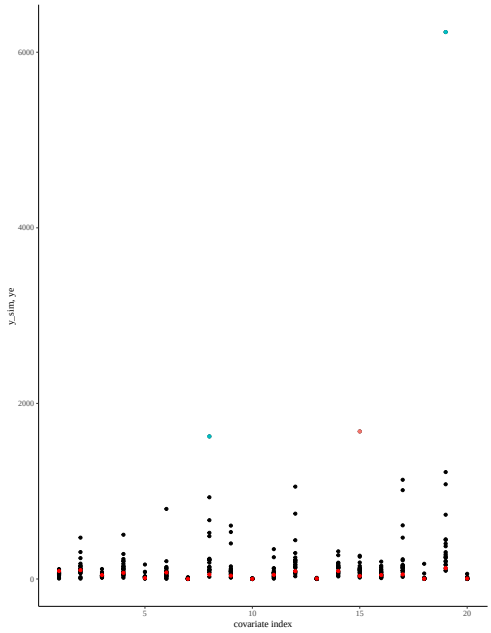

Figure 3: simulated y values

Table 1: simulated y sets comparison

simulation set	fitting time	lp	α	ρ
4	0.43s	-80	0.25	0.72
20	1.5S	-109	1.44	0.66
7	76s	-121	1.21	0.49

As our main concern is testing the algorithm, different prior settings are not discussed in detail in this paper. However, we note that when heavy tailed prior was experimented, inverse-gamma for example, unreasonable values were simulated more frequently leading to SBC test failure (see Appendix 4.2).

3 General linear regression model with a regularized horseshoe prior

For Bernoulli likelihood, general linear regression model with a regularized horseshoe prior is used. Figure 4, shows SBC histogram of selected four hyperparameters. λ_1 and λ_{86} each represents coefficient for minor and major cancer factor. Figure 5 shows ECDF for the parameters. Though log scaled parameters are quantities of interest, log is a monotone function and therefore, SBC rank comparison was performed on the original scale. While slight overestimation could be checked for τ, other parameters are within the bounds. p is a probability vector of each patient having a cancer and is the function of latent parameter. Components corresponding to five patients were tested. Both SBC histogram and ECDF imply that the approximation overestimates the cancer probability (Figure 6).

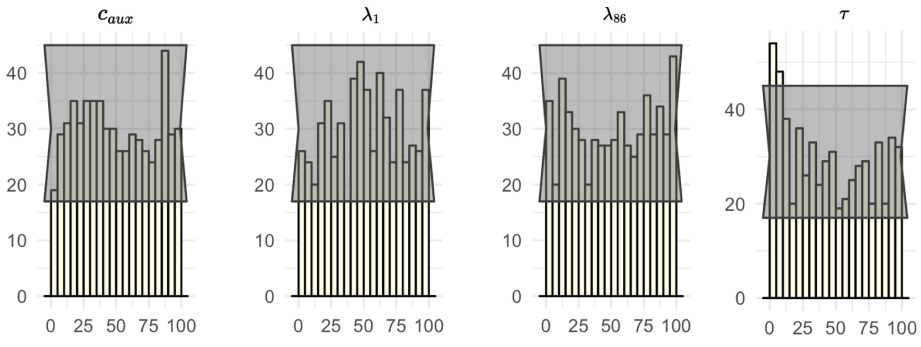

Figure 4: SBC histogram for hyperparamters in prostate cancer data

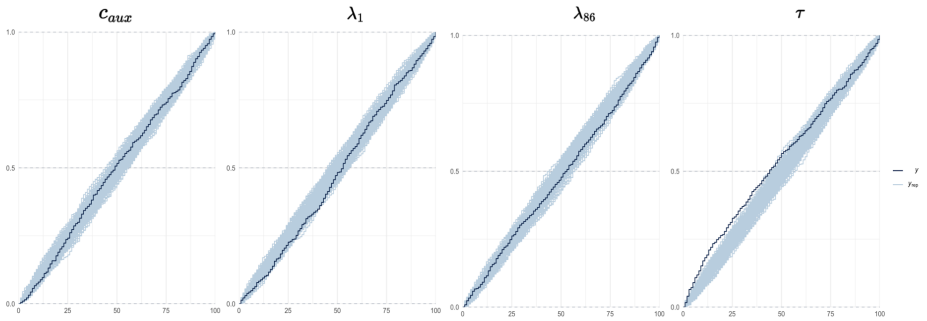

Figure 5: SBC ECDF for hyperparamters in prostate cancer data

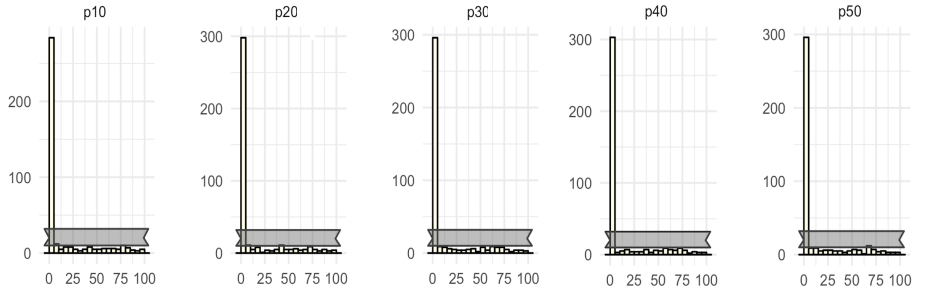

Figure 6: SBC histogram for cancer probability

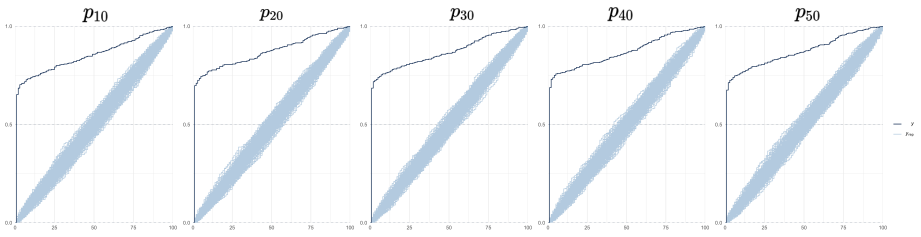

Figure 7: SBC ECDF for cancer probability

Several prior and data averaged posterior samples were compared for each parameter. For the latter, mean values of posterior samples are used for the plot. Note that underscored are prior samples. As can be seen from Figure 8, both major and minor factors showed the tendency to overestimate the values even though SBC histogram and ECDF test did not flagged any problem. This could be the result of SBC being underpowered. For the probability of certain patients getting a cancer, the approximation algorithm showed tendency to balance out the probability between two extremes, 0 and 1, which is not surprising as Laplace approximation is known to avoid extreme values of the probability.

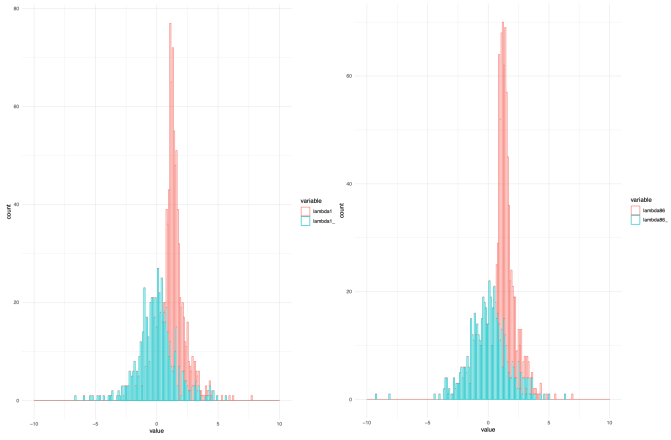

Figure 8: Compare λ prior and data averaged posterior samples

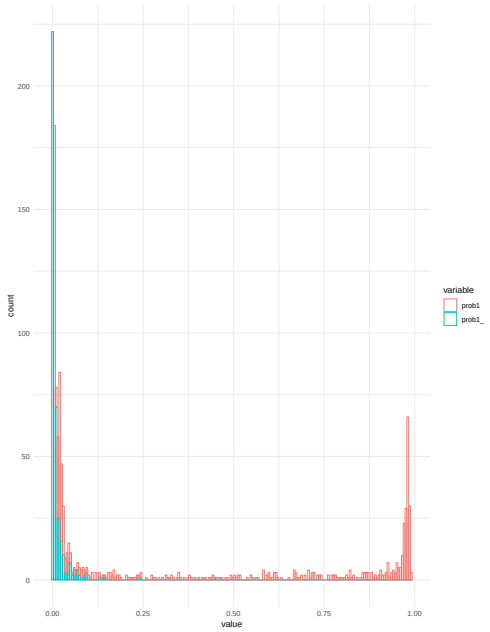

Figure 9: Compare cancer probability prior and data averaged posterior samples

4 Appendix

4.1 SBC Proof

The probability of the rank, r, of a prior sample, θ_0, relative to L posterior samples $\left\{\tilde{\theta}_1, \ldots, \tilde{\theta}_L\right\}$, conditioned on an intermediate data simulation, $\tilde{y}$

$$\begin{aligned} \tilde{\theta}_0 &\sim \pi(\theta) \\ \tilde{y} &\sim \pi\left(y \mid \tilde{\theta}_0\right) \\ \left\{\tilde{\theta}_1, \ldots, \tilde{\theta}_L\right\} &\sim \pi(\theta \mid \tilde{y}) \end{aligned}$$

$$\begin{aligned}
\pi(r) &= \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \frac{L!}{r!(L-r)!} \mathbb{P}\left[\theta_l < \theta_0\right] \cdot \mathbb{P}\left[\theta_l \geq \theta_0\right] \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \mathbb{P}\left[\theta_l < \theta_0\right] \cdot \mathbb{P}\left[\theta_l \geq \theta_0\right] \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \left[\prod_{l=1}^{r} \int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right] \left[\prod_{l=r+1}^{L} \int_{\theta_0}^{\infty} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right] \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \left[\int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right]^r \left[\int_{\theta_0}^{\infty} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right]^{L-r} \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \left[\int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right]^r \left[1 - \int_{0}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid \theta_0, y\right)\right]^{L-r}
\end{aligned}$$

is given by

Because the posterior samples are independent of the prior sample conditioned on the data we have

$$\pi\left(\theta_l \mid \theta_0, y\right) = \pi\left(\theta_l \mid y\right)$$

$$\begin{aligned}
\pi(r) &= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(y, \theta_0\right) \left[\int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^r \left[1 - \int_{0}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^{L-r} \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}\theta_0 \mathrm{d}y \pi\left(\theta_0 \mid y\right) \pi(y) \left[\int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^r \left[1 - \int_{0}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^{L-r} \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}y \pi(y) \int \mathrm{d}\theta_0 \pi\left(\theta_0 \mid y\right) \left[\int_{-\infty}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^r \left[1 - \int_{0}^{\theta_0} \mathrm{d}\theta_l \pi\left(\theta_l \mid y\right)\right]^{L-r}
\end{aligned}$$

Because we're simulating model configurations and observations from the same model from which our posterior is constructed

$$\pi\left(\theta_l \mid y\right) = \pi\left(\theta_0 \mid y\right)$$

In particular we can consider the change of variables

$$u(y) = \int_{-\infty}^{\theta_0} \mathrm{d}\theta \pi(\theta \mid y)$$

which gives

$$\begin{aligned}
\pi(r) &= \frac{L!}{r!(L-r)!} \int \mathrm{d}y \pi(y) \int \mathrm{d}u(y) [u(y)]^r [1 - u(y)]^{L-r} \\
&= \frac{L!}{r!(L-r)!} \int \mathrm{d}y \pi(y) \frac{r!(L-r)!}{(L+1)!} \\
&= \frac{1}{L+1} \int \mathrm{d}y \pi(y) \\
&= \frac{1}{L+1}
\end{aligned}$$

4.2 SBC on Different Prior Settings

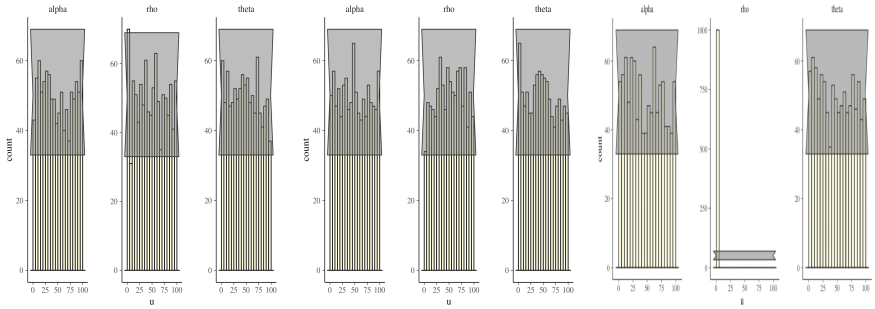

Figure 10: SBC histogram with different prior parameters and shape

SBC histograms corresponding to different prior distributions, normal, t, and invese gamma are compared in Figure10. ρ parameter for inverse gamma prior is interesting. Heavy tail prior could have caused large variance in simulated y values and therefore imposed small values on data averaged posterior sample of ρ.

References

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

C. C. Margossian, A. Vehtari, D. Simpson, and R. Agrawal. Hamiltonian monte carlo using an adjoint-differentiated laplace approximation. *arXiv preprint arXiv:2004.12550*, 2020.

D. J. Schad, M. Betancourt, and S. Vasishth. Toward a principled bayesian workflow in cognitive science. *Psychological methods*, 2020.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv:1804.06788v1*, 2018.

J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15): 1580–1607, 2010.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018.