Approximate posterior recalibration

Tiffany Cai, Philip Greengard, Ben Goodrich, and Andrew Gelman

23 Apr 2024

Abstract

Bayesian inference is often implemented using approximations, and resulting posterior uncertainty intervals can then be too narrow, not fully capturing the uncertainty in the model. We address the question of how to adjust these approximate posteriors so that they appropriately capture uncertainty. We introduce two methods that extend simulation-based calibration checking (SBC) to widen approximate posterior uncertainty intervals so as to aim for marginal calibration. We demonstrate these methods in several experimental settings, and we discuss the challenge of calibration using posterior inferences.

1. Introduction

When we fit large, complicated models using approximate algorithms, there is concern about the calibration of uncertainty statements. We would like our inferences to recover true parameter values or population quantities, to the extent that they are identifiable from the data and model, and we would like our uncertainty statements to capture the errors in estimation.

Assume the following scenario of Bayesian computation. A true scalar parameter θ has been drawn by Nature from its prior distribution, $p(\theta)$. Data y are then drawn from $p(y|\theta)$. An analyst observes the data and knows the prior and data model and hence can write the posterior, $p(\theta|y) \propto p(\theta)p(y|\theta)$. By construction, Bayesian posterior intervals are calibrated, attaining nominal coverage with respect to the prior. If we (a) draw parameters θ^l for $l = 1, \ldots, L$ from the prior, (b) generate data y^l from θ^l, and (c) construct posterior intervals for θ^s_{post} conditioned on y^l, then those intervals should contain the true parameter θ^l the correct portion of the time, averaged over samples of θ^l [Cook et al., 2006], a procedure that has been formalized as simulation-based calibration checking (SBC) [Talts et al., 2020, Modrák et al., 2024].

However, often the exact posterior cannot be obtained because of computational reasons, so that the analyst instead uses an approximate computational procedure such as variational inference [Blei et al., 2017]. In contrast to intervals obtained from exact posteriors, intervals obtained from approximate posteriors are not generally expected to produce correct coverage with respect to the prior, even with a large number of simulations. At a practical level, we would still like to use approximate posteriors, especially when exact posterior inference is too computationally costly. However, we can adjust approximate intervals with the aim of achieving posterior calibration.

This paper makes two contributions. First, we propose recalibration methods based on simulation-based calibration checking for widening approximate posterior intervals so that they are calibrated with respect to the prior in Section 3. Essentially, we pick a class of potential adjustments, and choose the adjustment that best calibrates posteriors according to SBC. We demonstrate these recalibration methods in several settings in Section 4.

Second, we explore nuances in the problem of posterior recalibration, including the distribution with respect to which we would like to be calibrated. In particular, Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] propose and evaluate methods to do this, but in place of step (a) above (drawing θ^l from the prior), they draw θ^l from the posterior given observed data, $p(\theta|y)$. The advantage of using posterior draws is that the recalibration is focused on the zone of parameter space that is of applied interest given the problem at hand. The disadvantage is that we should no longer

expect calibration, even if the approximate inference were exact, that is, where exact posteriors are generated in step (c) above. Nonetheless, Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] find their procedure to work reasonably well to recalibrate variational inferences for a hierarchical model of applied interest. In Section 6 we work out the details of simulation-based recalibration for a one-dimensional normal-normal model, a simple setting where it is possible to obtain exact posterior draws. It turns out that when using this exact computation, inferences averaging over the prior predictive distribution are calibrated (up to Monte Carlo error), but inferences averaging over the posterior predictive distribution are not, outside of extreme cases. Performing posterior predictive recalibration results in inferences that differ from the exact posterior by doing less pooling toward the prior.

We conclude by discussing implications for simulation-based calibration and consider why posterior recalibration could still work well for hierarchical models, even while being so far off in a simple non-hierarchical setting.

Our work is related to that of Rodrigues et al. [2018], who propose a method for recalibrating inferences constructed using the ABC procedure, and Yu et al. [2021], who propose a method of calibration checking and recalibration based on moments rather than quantiles.

2. Simulation-based calibration checking

In this section, we introduce notation and describe simulation-based calibration checking [Cook et al., 2006, Talts et al., 2020, Modrák et al., 2024], which we extend in our recalibration procedures. We will use the notation M for the model that we are trying to fit, θ for the parameters in the model, y for the data, M' for the approximate computation, and $g(\theta)$ for a scalar quantity of interest. For simplicity, we take $g(\theta) := \theta$, though our methods are easily extensible to more general g.

A natural way to check if the posteriors are correct is through simulation-based calibration checking as follows:

Algorithm 0 *Simulation-based calibration checking*

Requires Models M, M', constants L, S

1: Obtain L independent draws $\theta^1, \ldots, \theta^L$ of θ from its prior distribution in model M.
2: **for** $l = \{1, \ldots, L\}$, in parallel **do**
3: Take θ^l, one of the draws of θ from step 1.
4: Draw $y^l \sim P(y|\theta^l)$ from the data model in M.
5: Fit M' to y^l to obtain S independent approximate posterior samples, $\theta^{l,s}_{\text{post}},\ s = 1, \ldots, S$.
6: Determine the quantile q^l of θ^l in this distribution, that is, $q^l = \frac{1}{S}\sum_{s=1}^{S} 1_{\theta^l > \theta^{l,s}_{\text{post}}}$
7: **end for**
8: If the model is fit exactly (that is, if $M' \equiv M$) and the simulation draws are independent, then $q^1, \ldots, q^L$ should be i.i.d. random draws from the uniform distribution on $[0, 1]$.
9: Compare $q^1, \ldots, q^L$ to the uniform distribution on $[0, 1]$. A discrepancy indicates miscalibration.

Simulation-based calibration checking [Talts et al., 2020, Modrák et al., 2024] relies on the well-known observation that the data-averaged posterior and the prior are self-consistent [Cook et al., 2006]: for the prior distribution $\pi(\theta)$ defined over parameters of interest θ, and $\theta^l \sim \pi(\theta)$, and $y^l \sim \pi(y|\theta^l)$, then by construction $(\theta^l, y^l) \sim \pi(\theta, y)$ so that

$$\theta^l \sim \pi(\theta|y^l). \tag{1}$$

In other words, the parameter θ^l, which was drawn from $\pi(\theta)$, can also be thought of as being drawn from the posterior $\pi(\theta|y^l)$. Checking calibration is ultimately verifying that θ^l appears to be sampled from the posterior $\pi(\theta|y^l)$, which can be done by comparing θ^l with independently drawn samples from the actual posterior, $\theta^{l,s}_{\text{post}} \sim \pi(\theta|y^l)$. From this we obtain the condition used in SBC to check for calibration:

$$\pi(\theta) = \int\int \pi(\theta^l)\,\pi(y^l|\theta^l)\,\pi(\theta|y^l)\,d\theta^l\,dy^l. \tag{2}$$

In this paper, we go beyond the above SBC procedure. We don't want to just identify problems with calibration; we would also like to obtain roughly calibrated intervals.

3. Proposed calibration procedures

The idea is as follows: because the exact posterior satisfies data-averaged self-consistency (1), the exact posterior satisfies various specific self-consistency properties. Now suppose we have a way to obtain approximate posteriors. We can choose a specific self-consistency property and learn an adjustment so that these approximated posteriors satisfy this specific self-consistency property, too. Then we can apply the adjustment we learned in the previous step on an approximated posterior for a specific dataset y_D that we care about. The general algorithm is as follows:

Algorithm 1 *Approximate posterior calibration (general)*

Requires Models M, M', constants L, S, data y_D, specific self-consistency property

1: Obtain L independent draws $\theta^1, \ldots, \theta^L$ of θ from its prior distribution in model M.
2: **for** $l = \{1, \ldots, L\}$, in parallel **do**
3: Take θ^l, one of the draws of θ from step 1.
4: Draw $y^l \sim P(y|\theta^l)$ from the data model in M.
5: Fit M' to y^l to obtain S independent approximate posterior simulations, $\theta^{l,s}_{\text{post}},\ s = 1, \ldots, S$.
6: **end for**
7: Calculate the adjustment needed to satisfy the specific self-consistency property.
8: Fit M' to y_D to obtain an approximate posterior.
9: Apply the adjustment from Step 7 to the approximate posterior from Step 8.

In the above, we have not specified what "Calculate the adjustment needed to satisfy the specific property" is, as this step can vary with the specific property desired. Thus from the above we have a class of procedures to adjust a posterior approximation method, each corresponding to a way of obtaining an appropriate adjustment for a posterior, which consists of

1. A class of adjustments, and
2. A way to choose an adjustment from that class of adjustments.

For the sake of simplicity, in our methods, our class of adjustments will be just a fixed width scaling σ for posteriors, where the scaling is around the mean, so that if the samples from the approximate posterior are $\theta^{l,s}_{\text{post}}$ for $s = 1, \ldots, S$, the corresponding sample from the adjusted posterior is

$$\sigma(\theta^{l,s}_{\text{post}} - \overline{\theta}^l_{\text{post}}) + \overline{\theta}^l_{\text{post}}, \tag{3}$$

where $\overline{\theta}^l_{\text{post}}$ is the mean of $\theta^{l,s}_{\text{post}}$ for $s = 1, \ldots, S$ and σ is chosen by the method.

Now we discuss two ways of choosing adjustments. One can easily think of variations, but we stick to these for simplicity.

3.1. Nominal coverage method

The objective is for $(1-\alpha)$ intervals to achieve nominal coverage. From (1), the quantiles $q(a) := P(\theta^l \geq \theta^{l,s}_{\text{post}})$, with θ^l_{post} should be uniform across replications indexed by l if $\theta^{l,s}_{\text{post}}$ are draws from the true posterior distribution of θ given y^l. Thus, $(1-\alpha)$ of the time, θ^l should be contained within the interval from the $\alpha/2$th quantile to the $(1-\alpha/2)$th quantile of θ^l_{post}. In other words, the property we aim to satisfy is

$$q(1-\alpha/2) = 1-\alpha/2 \quad \text{and} \quad q(\alpha/2) = \alpha/2.$$

If we only have samples $\theta^{l,s}_{\text{post}}$ for $s = 1, \ldots, S$ that are from an approximate posterior, and if we believe that only the posterior variance requires correction **[TC: probably rephrase this: global vs local correction]**, then we can scale the posterior samples $\theta^{l,s}_{\text{post}}$ by some σ around its mean where the σ is chosen to minimize the following objective:

$$(q(1-\alpha/2) - q(\alpha/2)) - (1-\alpha))^2.$$

One way to find the best adjustment is to do a grid search over potential adjustment values σ to find the minimum value of the objective above.

3.2. Method using z-scores

As before, let $\theta^l \sim \pi(\theta)$, $y^l \sim \pi(y|\theta^l)$. If $\theta^{l,s}_{\text{post}} \sim \pi(\theta|y^l)$ are draws from the true posterior, we can think of θ^l as being drawn from $\pi(\theta|y^l)$, from which we have draws $\theta^{l,s}_{\text{post}}$, we can consider the z-score for θ^l, $z^l := (\mathbb{E}(\theta^{l,s}_{\text{post}}) - \theta^l)/\text{sd}(\theta^{l,s}_{\text{post}})$. Across $l = 1, \ldots, L$, the z-scores z^l should have mean 0 and variance 1. Thus, the property we aim to satisfy is

$$\text{Self-consistency:} \quad \mathbb{E}\left(\frac{\mathbb{E}(\theta^{l,s}_{\text{post}}) - \theta^l}{\text{sd}(\theta^{l,s}_{\text{post}})}\right) = 0 \;\text{ and sd}\left(\frac{\mathbb{E}(\theta^{l,s}_{\text{post}}) - \theta^l}{\text{sd}(\theta^{l,s}_{\text{post}})}\right) = 1, \tag{4}$$

where the inner expectation and sd are over posterior samples $s = 1, \ldots, S$ for a single l, and the outer expectation and sd are over draws $l = 1, \ldots, L$.

For some posterior approximation methods, if we believe that only the posterior variance requires correction, it can be reasonable to focus on just the second property. Thus, if $\theta^{l,s}_{\text{post}}$ are approximate posterior samples, a way to satisfy (4) is to consider an adjusted version of posterior samples

$$\theta^{\text{adj},l,s}_{\text{post}} := \overline{\theta}^l_{\text{post}} + s_z(\theta^{l,s}_{\text{post}} - \overline{\theta}^l_{\text{post}}) \tag{5}$$

where s_z is the standard deviation of z^l across $l = 1, \ldots, L$, while $\overline{\theta}^l_{\text{post}}$ is the mean of $\theta^{l,s}_{\text{post}}$ across $s = 1, \ldots, S$. By construction, if we assume the first self-consistency property holds, then this adjusted posterior satisfies the second self-consistency property in (4).[1]

[1] If we adjust the approximated posterior with both a scale and a shift to satisfy both parts of (4), then the adjustment would be

$$\theta^{\text{adj},l,s}_{\text{post}} := \overline{\theta}^l_{\text{post}} + s_z(\theta^{l,s}_{\text{post}} - \overline{\theta}^l_{\text{post}}) - \overline{z}s^l_{\text{post}} \tag{6}$$

where $\overline{z}, s_z$ are the mean and standard deviation of z^l for $l = 1, \ldots, L$, while $\overline{\theta}^l_{\text{post}}, s^l_{\text{post}}$ are the mean and standard deviation of $\theta^{l,s}_{\text{post}}$ for $s = 1, \ldots, S$. To see this, write $\mu^l_{\text{post}} := \mathbb{E}[\theta^{l,s}_{\text{post}}]$, $s^l_{\text{post}} := \text{sd}(\theta^{l,s}_{\text{post}})$, $z^l := \frac{\mu^l_{\text{post}} - \theta^l}{s^l_{\text{post}}}$. Consider

3.3. Comparison

In the following sections, we will compare the z-score method with the nominal coverage method. The z-score method does not directly aim for nominal coverage for specific $(1-\alpha)$ intervals; there is only one σ overall. This can be a disadvantage if the posterior width scaling *should* vary by α. However, if the optimal posterior width scaling does not vary by α, the z-score method has the advantage of being comparatively more stable to estimate than the nominal coverage method, especially for α that are close to 0 or 1. Unlike the nominal coverage method, the z-score method can be used to select σ having to search for the best value. Empirically, the two methods seem to give similar values for σ in our experiments.

4. Calibration experiments

We demonstrate and evaluate our procedure by applying it to several examples that we already understand well, where we know the ground truth posterior and can use it to evaluate our procedure.

4.1. A simple example: a one-parameter Gaussian model

To provide a concrete and simple example, we apply APC to the following model:

$$\begin{aligned} \theta &\sim \text{normal}(0, 1) \\ y|\theta &\sim \text{normal}(\theta, 1) \end{aligned} \tag{8}$$

For simplicity, we choose $g(\theta) = \theta$. With this simple model we have the benefit of a formula for the posterior,

$$\theta \sim \text{normal}\left(\frac{1}{N+1}\sum_{i=1}^{N} y_i,\ \frac{1}{\sqrt{N+1}}\right). \tag{9}$$

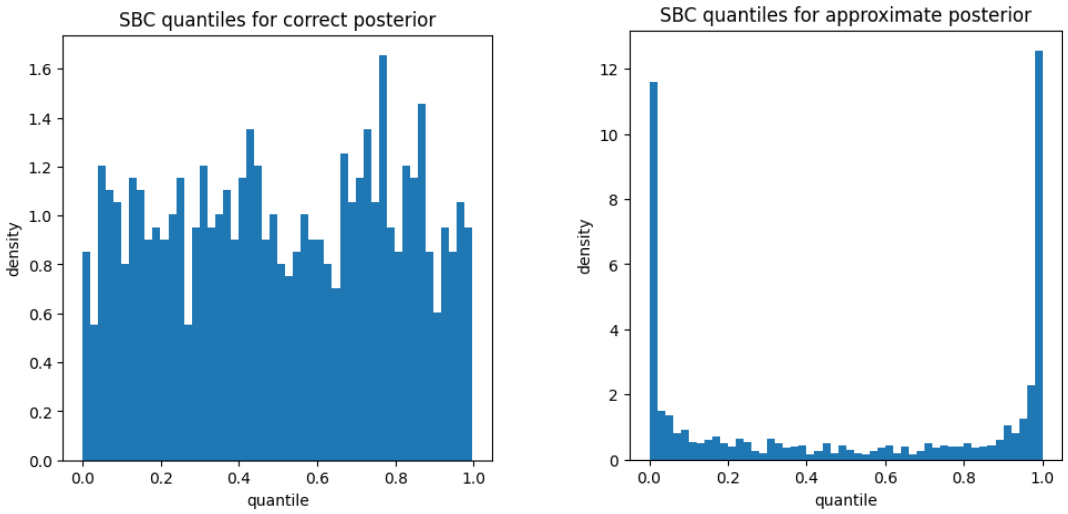
The primary purpose of our method is to calibrate confidence intervals for functions of the posterior when the posterior is difficult to sample from and only approximate inference methods are available. In the above model we're far from that regime—we can easily sample from the posterior with for example, HMC, or even i.i.d. draws. We use the above model primarily to build intuition.

We now demonstrate the calibration procedure applied to this simple model. We artificially introduce a sampling error by narrowing the posterior standard deviation by a scaling factor of 3. Without knowing the ground truth, we can tell that the approximated posterior is not calibrated, by using quantiles from simulation-based calibration checking (SBC) [Talts et al., 2020], which we have also repeated here in Algorithm 0. Then, to calibrate $(1-\alpha)$ confidence intervals, we find scale adjustments using the nominal coverage method using a grid search, for scaling values $2, 2.01, 2.02, \ldots, 4.98, 4.99, 5$. Indeed, using this method, we nearly recover the true posterior for each value of α in Table 1, as the scaling found is usually close to 3. After calibration, the adjusted posteriors obtain almost nominal coverage (Table 1). We also find posterior scaling adjustments

adjusted posterior samples $\tilde{\theta}^{l,s}_{\text{post}} := \sigma^l(\theta^{l,s}_{\text{post}} - \mu^l_{\text{post}}) + \mu^l_{\text{post}} + c^l$, so that $\tilde{\mu}^l_{\text{post}} := \mathbb{E}[\tilde{\theta}^{l,s}_{\text{post}}] = \mu^l_{\text{post}} + c^l$ and $\tilde{\mu}^l_{\text{post}} := \text{sd}[\tilde{\theta}^{l,s}_{\text{post}}] = \sigma^l s^l_{\text{post}}$. Then the z-scores for adjusted posteriors are

$$\tilde{z}^l := \frac{\tilde{\mu}^l_{\text{post}} - \theta^l}{\tilde{s}^l_{\text{post}}} = \frac{\mu^l_{\text{post}} - \theta^l + c^l}{\sigma^l s^l_{\text{post}}} = \frac{1}{\sigma^l}(z^l + \frac{c^l}{s^l_{\text{post}}}). \tag{7}$$

In order to satisfy the conditions in (4), i.e. that $\mathbb{E}(\tilde{z}^l) = 0$ and $\text{sd}(\tilde{z}^l) = 1$, then $\sigma^l = \text{sd}(z^l)$ is the standard deviation across z^l (one shift for all posteriors) and $c^l = -\mathbb{E}(z^l)s^l_{\text{post}}$.



(a) *SBC quantiles for true posterior*

(b) *SBC quantiles for posterior with sampling error*

Figure 1: *SBC quantiles for model* (8) *when sampling from the exact posterior (left) and posterior with sampling error (right)*

$1-\alpha$	posterior width scaling σ	adjusted coverage
0.95	3.10	0.952
0.90	2.98	0.896
0.80	3.07	0.795
0.50	3.17	0.491

Table 1: *Posterior scaling of* $(1-\alpha)$*-level posterior intervals for the simple model* (8) *using the nominal coverage method.*

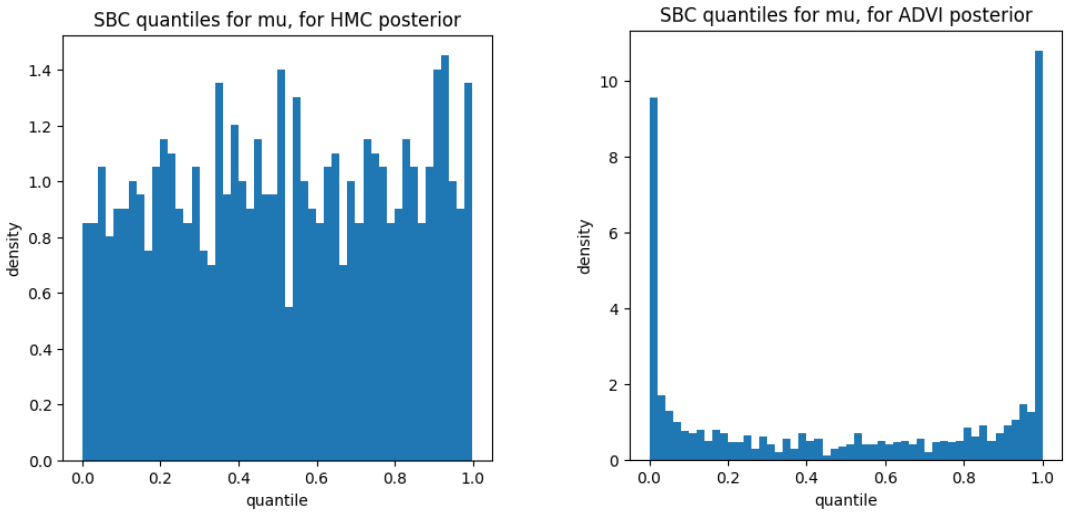
using the z-score method, which gives $\sigma = 3.083$ which is also close to the scaling to recover the true posterior (3), and is more computationally efficient than the nominal coverage method as it does not require grid search. The adjusted posteriors obtain almost nominal coverage by this method, as well (Table 2).

$1-\alpha$	posterior width scaling σ	adjusted coverage
0.95	3.08	0.951
0.90	3.08	0.909
0.80	3.08	0.796
0.50	3.08	0.475

Table 2: *Posterior scaling of* $(1-\alpha)$*-level posterior intervals for the simple model* (8) *using the* z*-score method.*

4.2. Hierarchical model (8 schools) and ADVI

We use the hierarchical modeling example for the 8 schools model, as defined and discussed in Chapter 5 of Gelman et al. [2013].



(a) SBC quantiles with HMC

(b) SBC quantiles with ADVI

Figure 2: *SBC quantiles for the eight schools model when sampling with HMC (left) and ADVI (right). The HMC posterior is calibrated, while the ADVI posterior is not.*

Here, we will compute the posterior distribution using automatic differentiated variational inference (ADVI), a fast approximate method. We use ADVI to obtain approximate posteriors for the 8 schools example, and then we use our methods to adjust the approximate posteriors. We focus on the parameter μ, the overall mean of the effect size across all schools. Again, we choose $g(\theta) = \theta$. Later, to sanity check, we compare the unadjusted and adjusted ADVI posteriors to a posterior computed using HMC on a well-parameterized model, which we consider to be ground truth after checking simulation-based calibration on that posterior [Modrák et al., 2024, Talts et al., 2020]. We use ADVI on a centered parameterization, and HMC on a non-centered parameterization for the 8 schools example, as it is easier to produce a correct posterior using a non-centered parameterization than a centered parameterization, for the purposes of demonstrating our posterior adjustment method.

Although the calibration process requires repeated computation, the posterior inferences can all be run in parallel, so that the total run time can be comparable to that of just one posterior inference computation. Furthermore, the calibration adjustment can be computed for multiple $(1 - \alpha)$ intervals after approximate posteriors are computed, and also, once an adjustment is computed for $(1 - \alpha)$ intervals, the adjustment can be directly applied to multiple datasets y.

We draw $L = 1000$ θ^l's, and for each θ^l, we generate $S = 1000$ $\theta^{l,s}_{\text{post}}$'s. From Figure 2 where we calculated SBC quantities [Talts et al., 2020], we see that the approximate posterior from ADVI is not calibrated. Then, the appropriate scaling we find using grid search over σ are displayed in Table 3. We compare to the z-score method in Table 4. We see that the posterior width scaling σ is similar under the nominal coverage method and the z-score method. Accordingly, the $(1 - \alpha)$ intervals for the corresponding adjusted posteriors also have similar coverage.

Now we contextualize the posterior adjustments we found by comparing the ADVI (approximate) and HMC ("ground truth") posteriors. When we compare the ADVI posterior variances to the HMC ("ground truth") posterior variances in Figure 3, we find that the ADVI posterior variances are about twice as wide. This is smaller than the scalings in Table 3 and 4. Nevertheless, the larger scaling we found is sensible because in Figure 3 the ADVI posteriors are shifted (not only

$1-\alpha$	posterior width scaling σ	adjusted coverage
0.95	2.41	0.949
0.90	2.41	0.987
0.80	2.56	0.802
0.50	2.46	0.498

Table 3: *Posterior scaling of* $(1-\alpha)$*-level posterior intervals from ADVI for the 8 schools example using the nominal coverage method.*

$1-\alpha$	posterior width scaling σ	adjusted coverage
0.95	2.48	0.958
0.90	2.48	0.909
0.80	2.48	0.788
0.50	2.48	0.505

Table 4: *Posterior scaling of* $(1-\alpha)$*-level posterior intervals from ADVI for the 8 schools example using the z-score method.*

scaled) from the HMC posteriors, where the shift varies by θ^l and appears to be centered around 0. Because of this shift, widening ADVI posteriors by only a factor of 2 isn't enough for posterior intervals to contain the true draw θ^l the correct proportion of the time. Thus, in the absence of a way to adequately correct for the variation in shift, the next best thing to do is to widen posterior intervals by the scaling factors that we found.

5. Posterior recalibration

Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] have reported success following a similar procedure for a hierarchical model, but averaging over the posterior predictive distribution rather than averaging over the prior in Step (1) in Algorithm (1). Their approach has the benefit of focusing the recalibration on the zone of parameter space that is consistent with the data. However, posterior predictive recalibration cannot work in general. We demonstrate with a simple normal-normal example, working out the details of prior and posterior recalibration. Applying the recalibration procedure to the posterior predictive distribution has the effect of reducing the pooling toward the prior.

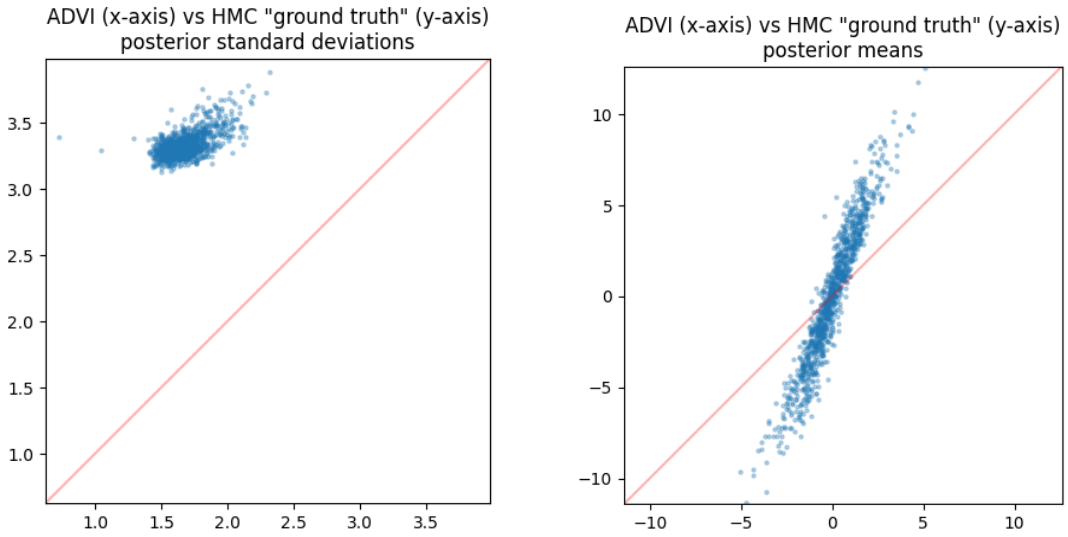
6. Understanding posterior recalibration for a one-parameter normal model

We can examine the differences between prior and posterior recalibration by considering a simple non-hierarchical model with only one parameter. In this scenario, the model contains no internal replication, and there is no reason to expect posterior recalibration, even approximately.

We consider the following simple normal-normal model with one scalar parameter and a point estimate:

$$\begin{aligned} \theta &\sim \text{normal}(0,1) \\ y|\theta &\sim \text{normal}(\theta,\sigma), \end{aligned} \tag{10}$$

with σ assumed known. This setup is more general than it might look, as y can be taken to be not just one observation but rather as any unbiased estimate such as a sample mean or regression



(a) Posterior standard deviation for ADVI vs. HMC ("ground truth")

(b) Posterior mean for ADVI vs. HMC ("ground truth")

Figure 3: *Posterior mean and standard deviation for ADVI vs. HMC*

coefficient. A proper prior is necessary here, as otherwise it would not be possible to perform prior predictive evaluation.

For our example, it is trivial to draw directly from the posterior, so we can evaluate the properties of calibration checking without needing to define an approximate posterior, g.

6.1. Prior calibration checking

First we check that prior calibration checking works here, as we know it should in general, from Cook et al. [2006]. With prior predictive checking, the replications $l = 1, \ldots, L$ have the following distribution:

$$\theta^l \sim \text{normal}(0, 1) \tag{11}$$

$$y^l|\theta^l \sim \text{normal}(\theta^l, \sigma) \tag{12}$$

$$\theta^{l,s}_{\text{post}}|y^l \sim \text{normal}(\mu^l_{\text{post}}, \sigma_{\text{post}}). \tag{13}$$

where

$$\mu^l_{\text{post}} = \frac{y^l}{1+\sigma^2} \quad \text{and} \quad \sigma_{\text{post}} = \left(\frac{\sigma^2}{1+\sigma^2}\right)^{1/2}.$$

In the limit of large S, the z-score of θ^l within the simulations of $\theta^{l,s}_{\text{post}}$ is

$$z^l = \frac{1}{\sigma_{\text{post}}}(\theta^l - \mu^l_{\text{post}}). \tag{14}$$

The corresponding quantiles are $\Phi(z^l)$, but since we are working here entirely with normal distributions it will be simpler to stick with z-scores.

For prior calibration checking, we must now look at the distribution of these z-scores, averaging θ^l and y^l over (11) and (12). The result is that $\theta^l - \mu^l_{\text{post}} \sim \text{normal}(0, \sigma_{\text{post}})$, hence the z-score (14) has a unit normal distribution in the limit of large number of simulation draws S.

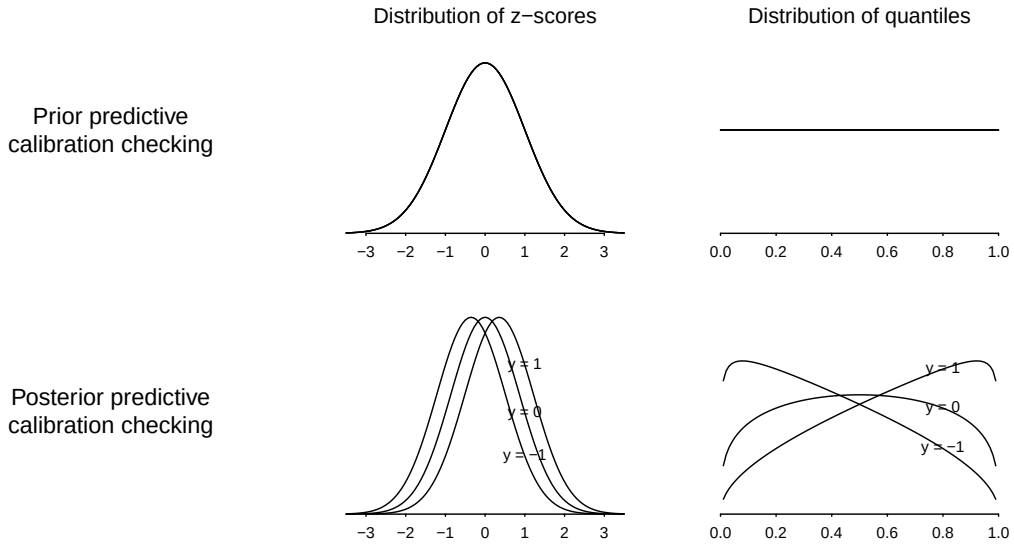


Figure 4: *Distribution of z-scores and quantiles in simulation-based calibration checking when the correct model is being fit for the simple normal example, under two scenarios, both for* $\sigma = 1$. Top row: *Simulating the parameter* θ^l *from the prior distribution.* Bottom row: *Simulating* θ^l *from the posterior distribution, in which case the distributions depend on the data,* y. *Curves show distributions conditional on three possible data values.*

6.2. Posterior calibration checking

Next we look at the properties of calibration checking when starting with draws from the *posterior* distribution. All is the same as before except that the prior (11) is replaced with the posterior (13),

$$\theta^l \sim \text{normal}(\mu_{\text{post}}, \sigma_{\text{post}}). \tag{15}$$

The z-score of θ^l within the simulations $\theta_s, s = 1, \ldots, S$ is still given by (14); the only difference is that, instead of averaging over (11) and (12), we average over (15) and (12). In this posterior predictive distribution, some algebra leads to

$$z^l = \frac{1}{\sigma_{\text{post}}}(\theta^l - \mu^l_{\text{post}}) \sim \text{normal}\left(\frac{\sigma}{(1+\sigma^2)^{3/2}}\, y, \frac{\sqrt{\sigma^4+\sigma^2+1}}{1+\sigma^2}\right). \tag{16}$$

For no value of y will this be a unit normal distribution (for $0 < \sigma < \infty$), thus we would *not* see predictive calibration (a uniform distribution of the L quantiles), even if the computation is exact.[2]

[2]To derive (16), let z_1 and z_2 be independent draws from the unit normal distribution and then do the following steps:

1. Draw $\theta^l \sim \text{normal}(y/(1+\sigma^2), \sigma/\sqrt{1+\sigma^2})$; equivalently, $\theta^l = \frac{y}{1+\sigma^2} + \frac{\sigma}{\sqrt{1+\sigma^2}} z_1$
2. Draw $y^l|\theta^l = \text{normal}(\theta^l, \sigma)$; equivalently, $y^l = \theta^l + \sigma z_2$
3. Compute $z^l = \frac{1}{\sigma_{\text{post}}}(\theta^l - \mu^l_{\text{post}})$; here, $\mu^l_{\text{post}} = \frac{y^l}{1+\sigma^2} = \frac{\theta^l + \sigma z_2}{1+\sigma^2}$. Then $\theta^l - \mu^l_{\text{post}} = \theta^l\left(1 - \frac{1}{1+\sigma^2}\right) - \frac{\sigma}{1+\sigma^2} z_2$. Then substitute $\theta^l = \frac{y}{1+\sigma^2} + \frac{\sigma}{\sqrt{1+\sigma^2}} z_1$ and use $\frac{1}{\sigma_{\text{post}}} = \frac{\sqrt{(\sigma^2+1)}}{\sigma}$ to rewrite

$$z^l = \frac{\sqrt{\sigma^2+1}}{\sigma}\left(\frac{y}{1+\sigma^2}\frac{\sigma^2}{1+\sigma^2} + \frac{\sigma}{\sqrt{1+\sigma^2}}\frac{\sigma^2}{1+\sigma^2} z_1 - \frac{\sigma}{1+\sigma^2} z_2\right)$$

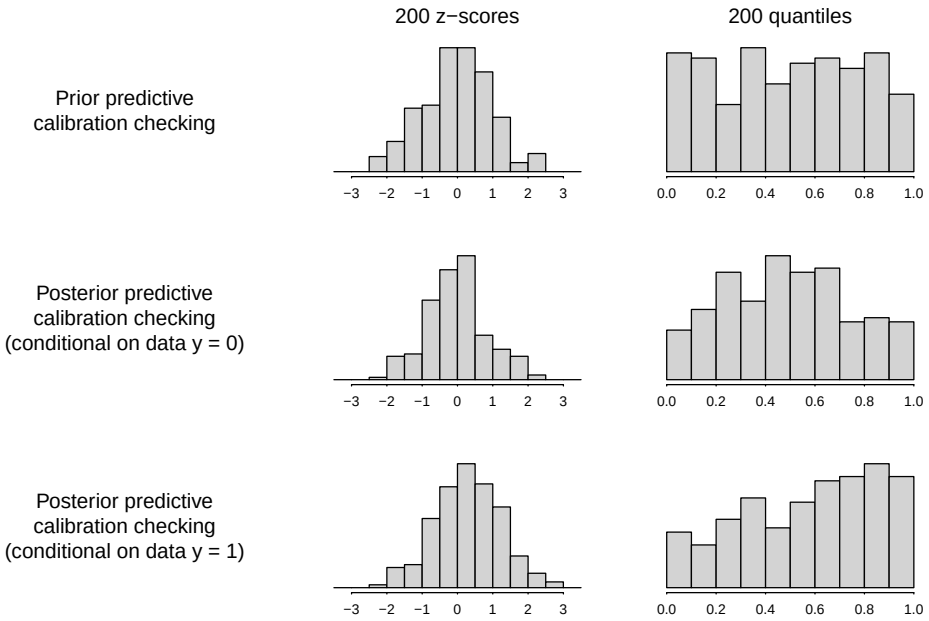


Figure 5: *Simulation-based calibration checking using 200 replications of 1000 simulation draws for each, for* $\sigma = 1$. Top row: *Simulating the parameter* θ^l *from the prior distribution.* Center and bottom rows: *Simulating* θ^l *from the posterior distribution, in which case the distributions depend on the data,* y. *As predicted by theory, the prior predictive simulations show calibration and the posterior predictive simulations do not, for* $\sigma = 1$.

The bottom row of Figure 4 shows the distribution of the z-scores and quantiles for different values of y, for $\sigma = 1$. For comparison, the top row of the figure shows the corresponding curves for the calibrated case of prior predictive checking.

Figure 5 shows corresponding results from a simulation with $\sigma = 1$, $L = 200$ and $S = 1000$, first drawing θ^l from the prior, which correctly shows calibration, then drawing from two different posteriors, one conditional on data $y = 0$, the other conditional on $y = 1$. The posterior z-scores and quantiles do not show calibration, despite that there is no approximation in the simulations.

These z-scores *are* a standard normal when $\sigma = 0$ and when $\sigma \to \infty$, indicating that posterior calibration checking is valid in the extreme cases in which the prior or the data are uninformative.

6.3. Prior recalibration

The idea of simulation-based recalibration is to use any miscalibration found in the checking to adjust the simulation draws. Here we work with one of the methods of Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] using a location-scale shift. The first step is to perform calibration checking, in which we obtain z-scores $z^l, l = 1, \ldots, L$, and summarize the miscalibration by $\bar{z}$ and s_z, the mean and standard deviation of these L values. The next step is to alter the fitting procedure by dilating the simulations $\theta^{l,s}_{\text{post}}$ by the factor s_z and shifting them by the relative value $\bar{z}$. In this affine transformation, draws from the approximate posterior θ^l_{post} are replaced by these adjusted draws:

$$\theta^{\text{adj},l,s}_{\text{post}} = \bar{\theta}^l_{\text{post}} + s_z(\theta^{l,s}_{\text{post}} - \bar{\theta}^l_{\text{post}}) + \bar{z}s^l_{\text{post}}\,, \tag{17}$$

which has mean $\frac{\sqrt{\sigma^2+1}}{\sigma}\frac{y}{1+\sigma^2}\frac{\sigma^2}{1+\sigma^2} = \frac{\sigma}{(1+\sigma^2)^{3/2}y}$ and variance $\frac{\sigma^2+1}{\sigma^2}\left(\frac{\sigma^6}{(1+\sigma^2)^3} + \frac{\sigma^2}{(1+\sigma^2)^2}\right) = \frac{\sigma^4+\sigma^2+1}{(1+\sigma^2)^2}$.

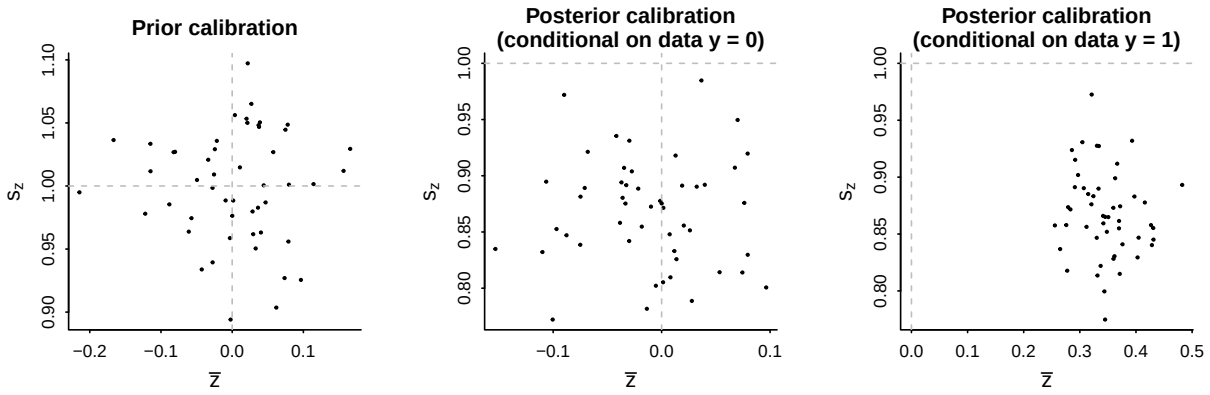


Figure 6: *Estimated mean and scale shifts from simulation-based calibration using 50 replications of 1000 simulation draws. In each scatterplot, the dots correspond to 100 independent simulations of the calibration process.* Left: *Simulating the parameter θ^l from the prior distribution; adjustments are close to the null adjustment, $(\bar{z}, s_z) = (0, 1)$.* Center and right: *Simulating θ^l from the posterior distribution, in which case the distributions depend on the data, y. In these cases, the recalibration has a large effect on posterior inferences.*

where $\bar{\theta}^l_{\text{post}}$ and s^l_{post} are the mean and standard deviation of the S values, $\theta^{l,s}_{\text{post}}$. Under calibration, $\bar{z}$ and s_z should be approximately 0 and 1, respectively, in which case the adjustment should essentially do nothing. If the fitting procedure is off in its first two moments, this recalibration should correct for that. This is the same adjustment and conditions as in Equations (4) and (7), except when draws are from the posterior, rather than the prior.

If we apply the calibration procedure to the prior predictive distribution, there should be essentially no effect. More precisely, there will be some random adjustments because of finite number of simulations—Monte Carlo error—but these adjustments should be minor.

We check this by performing this recalibration with $L = 200$ prior simulation draws and $S = 1000$ draws from each posterior, replicating the entire procedure 50 times. For each, we calculate $\bar{z}$ and s_z. The left plot in Figure 6 shows the results; most of the time the mean shift is less than 0.1 and the scale shift is less than 5%.

6.4. Posterior recalibration

What happens if we try to calibrate based on the posterior quantile? From a Bayesian standpoint, this is the wrong thing to do, but we can see what happens. As before, we set $L = 100$ and $S = 1000$ and then loop the entire process 100 times to see what could happen.

The center and right plots in Figure 6 show the results conditional on data $y = 0$ and $y = 1$, respectively. To understand what is happening, we consider a typical value for $(\bar{z}, s_z)$ in each case:

- Given data $y = 0$, the correct posterior distribution is $\theta|y \sim \text{normal}(0, 0.71)$. Shifting this by $\bar{z}s_{\text{post}} = 0$ and scaling it by $s_z = 0.87$ yields an adjusted posterior of $\text{normal}(0, 0.61)$.

- Given data $y = 1$, the correct posterior distribution is $\theta|y \sim \text{normal}(0.5, 0.71)$. Shifting this by $\bar{z}s_{\text{post}} = 0.35 \cdot 0.71 = 0.25$ and scaling it by $s_z = 0.87$ yields an adjusted posterior of $\text{normal}(0.75, 0.61)$.

Those simulations show what could happen with $L = 200$ in two special cases, $y = 0$ and $y = 1$.

For general y, we can work out analytically the adjustment that would occur in the limit of large L and S. Given θ^l and y^l, the z-score of θ^l in the limit of large S is $z^l = \sqrt{2}(\theta^l - y^l/2)$,

from (15). We can figure out the mean and standard deviation of this distribution, averaging over the posterior predictive distribution, in two steps. First we average over $y^l|\theta^l \sim \text{normal}(\theta^l, 1)$. Propagating that uncertainty, z^l has mean $\theta^l/(2\sqrt{2})$ and standard deviation $1/(2\sqrt{2})$. Next we average over $\theta^l \sim \text{normal}(y/2, 1/\sqrt{2})$. Propagating that uncertainty, z^l ends up with mean $y/(2\sqrt{2})$ and standard deviation $\sqrt{3}/2$.

If we apply posterior recalibration in our problem, it will change the posterior inferences in two ways. First, there will be much less partial pooling. Instead of the posterior mean of θ being $y/2$, it becomes $(\frac{1}{2} + \frac{1}{2\sqrt{2}}\frac{1}{\sqrt{2}})y = \frac{3}{4}y$. Second, uncertainty will be understated. Instead of the posterior standard deviation of θ being $\frac{1}{\sqrt{2}} = 0.71$, it becomes $\frac{1}{\sqrt{2}}\frac{\sqrt{3}}{2} = 0.61$.

To understand how this happens, start with the actual posterior, $\theta|y \sim \text{normal}(y/2, 1/\sqrt{2})$. In this case, simulated data y^i will be centered around $y/2$, and posterior inferences for θ_{post} will be again partially-pooled toward zero and will be centered around $y/4$. The values of θ drawn from the posterior distribution will be systematically farther from the prior, compared to draws from θ_{post}, and the calibration procedure will then adjust for this by pulling the posterior simulations away from the prior, thus reducing the amount of pooling. In this case, the recalibrated intervals pool only half as much as the correct posterior intervals.

6.5. Prior or posterior recalibration?

Our motivation for simulation-based recalibration of inferences is that we are often using approximate computational algorithms, and it makes sense to check these algorithms using simulation-based experimentation, checking that the computation can approximately recover true parameter values. In general it is not possible to correct an approximate distribution in high dimensions. The hope with simulation-based recalibration is that it could be possible to attain approximately nominal coverage for scalar summaries, one at a time, without aiming to solve the impossible problem of calibrating the joint distribution.

Bayesian inference is automatically calibrated when averaging over the prior distribution. However, in complicated models, the prior distribution can be broad and include regions of parameter space that are not at all supported by the data. For any particular dataset, we will not care about the performance of the inference algorithm in these faraway places; rather, we want to focus our checking effort on the region of parameter space that is consistent with the model and data: the posterior distribution.

However, Bayesian inferences will not be calibrated when averaging over the posterior, as is well known in theory and demonstrated in the present paper. It is an open question whether it would be possible to recalibrate the posterior to adjust for systematic errors in the recalibration procedure itself. Similar issues arise with the bootstrap, another class of procedures that uses simulation to correct for inferential biases [Efron, 1982]: the general idea is intuitive, but the procedure itself introduces another source of variability.

The fundamental problem with posterior recalibration—that, from a Bayesian perspective, we would not expect or even want intervals to have nominal coverage averaging over the posterior—is similar to the calibration problems of posterior predictive p-values [Gelman et al., 1996]. Methods have been proposed to recalibrate predictive p-values to be uniformly distributed [Robins et al., 2000, Hjort et al., 2006]; it has also been suggested that a uniform distribution of p-values is not necessarily desirable for the goal of predictive model evaluation [Gelman, 2013].

6.6. Why posterior calibration can approximately work for multilevel models

We conjecture that posterior recalibration works so well in the multilevel example of Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] because of the structure of the model being fit. Consider a hierarchical model with hyperparameters ϕ, local parameters $\alpha_1, \ldots, \alpha_J$, and local data $y_1, \ldots, y_J$. When drawing from the posterior, we can draw from the α_j's for the J existing groups (in which case the inference for each α_j will be conditional on the observed y_j) or for J new groups (in which case the inference for each α_j will be from its prior, conditional on ϕ).

Now suppose you have enough data so that the hyperparameters ϕ are precisely estimated in the posterior, so that $p(\phi|y_{1:J})$ can be approximated by a delta function at some $\hat{\phi}$. In that case, $p(\alpha|y_{1:J}) \approx p(\alpha|\hat{\phi}, y_{1:J})$. Furthermore, posterior draws of α for J new groups are essentially the same as prior draws, since $p(\alpha|y_{1:J}) \approx p(\alpha|\hat{\phi}, y_{1:J}) \approx p(\alpha|\hat{\phi})$. In addition, if the model is correct and J is large enough, then posterior draws for α for the J existing groups will approximate a set of J draws from the prior. Thus, posterior recalibration checking should look a lot like prior recalibration checking for a hierarchical model.

6.7. A halfhearted defense of posterior recalibration in this example

Our result is distressing—the adjustment shifts the posterior distribution and also makes it overconfident! If we want to make an argument in favor of this procedure, we could say that practitioners often have a desire to perform less partial pooling than is recommended by Bayesian inference, in part because of concern about calibration, that intervals will not have nominal coverage conditional on the true parameter value. In straight Bayesian inference, intervals have nominal coverage conditional on the data. Calibration that creates nominal coverage under the posterior could be thought of as a sort of compromise, a Bayesian analogue to classical coverage, and a step toward some form of generalized Bayesian inference [Yao, 2024].

We are not fully convinced by this argument, as we went into this problem with the simpler goal of improving approximations to Bayesian inference. Still, we have seen examples where going outside the Bayesian framework can systematically improve predictions in a model-open setting [Yao et al., 2018, 2022], so we are open to the idea that posterior recalibration could serve some robustness goal. Alternatively, perhaps posterior recalibration could itself be recalibrated in some way.

7. Discussion

In this work we explored simple methods to recalibrate approximate posteriors that are based on simulation-based calibration checking [Modrák et al., 2024, Talts et al., 2020]. We also explored a surprising feature of a posterior recalibration method by Gershunskaya and Savitsky [2023] and Savitsky et al. [2024] that calibrates from the posterior, rather than from the prior. First we discuss the limitations of these simple methods from Section 3.

7.1. Limitations of proposed calibration methods

To be able to use a method with confidence, it is necessary to know where it fails and what its limitations are.

Some limitations are due to inherent assumptions around posterior calibration. First, we assume the model is correct, as calibration is defined with respect to data that are generated from parameters drawn from the prior. Second, as calibration is defined on average over parameters drawn from the prior, each individual adjusted posterior may not be correct.

Other limitations are due to details of our method. For example, it's possible that one constant width rescaling around the posterior mean is not the appropriate adjustment: perhaps the correct rescaling varies over data draws, or perhaps the posterior intervals should be widened but by a different amount for different $(1 - \alpha)$ intervals. In the absence of additional knowledge, a naive rescaling around the posterior mean seems like a reasonable, if imperfect, choice. Our method also only looks at rescaling, though it can be extended to include shifts, as well, which we leave to future work. Lastly, we focus on posterior calibration for one parameter at a time. We do not try to calibrate multiple parameters at a time.

7.2. Open questions of proposed calibration methods

Many open questions remain. For example, if we are to find the best rescaling, what is the most sample-efficient way to do so? When should we use either of the two methods that we propose in Section 3? If we are to find the best scale *and* shift, how would we do that? What can be proved about our methods?

Our work builds on simulation-based calibration checking [Talts et al., 2020, Modrák et al., 2024]. It seems doubtful that there could be any general way to take simulations from an approximation and transform them into simulations from the target distribution—if that could be done, one would already do so, and we know approaches such as importance resampling fall apart in high dimensions. In the present paper we hare aiming only for interval coverage of parameters or quantities of interest, one at a time. Even if our recalibration procedure works perfectly, it will not yield a set of recalibrated joint simulations, a point that also arises in the work of Fearnhead and Prangle [2012] and Li et al. [2017].

7.3. Looking forward

Bayesian inference is automatically calibrated averaging over the prior predictive distribution; from that perspective, calibration is a concern only to the extent that the prior and data models might be wrong (that is, robustness of inferences) and if there might be problems with computation. Posterior recalibration does not play any role in formal Bayesian inference; indeed, with proper and nondegenerate priors, Bayesian inferences cannot be calibrated under the posterior predictive distribution for the same reason that the posterior mean cannot be a classically unbiased estimate. The contribution of the present paper is to demonstrate and explore posterior miscalibration in a simple example.

That said, there can be reasons for studying posterior recalibration. From the standpoint of Bayesian workflow, it can make sense to study the performance of approximate computation in the zone of parameter space that is consistent with the data. More generally, calibration of forecasting and uncertainty estimation is a goal in its own right, not just restricted to Bayesian inference [Gneiting et al., 2007, Cockayne et al., 2022]. Finally, posterior recalibration could be a useful tool in hierarchical models, as suggested by Gershunskaya and Savitsky [2023] and Savitsky et al. [2024].

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

J. Cockayne, M. M. Graham, C. J. Oates, T. J. Sullivan, and O. Teymur. Testing whether a learning procedure is calibrated. *Journal of Machine Learning Research*, 23:9213–9248, 2022.

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

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society of Industrial and Applied Mathematics, 1982.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). 74:419–474, 2012.

A. Gelman. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7:2595–2602, 2013.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 7:733–807, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, third edition, 2013.

J. Gershunskaya and T. Savitsky. Calibration procedure for estimates obtained from posterior approximation algorithms, with application to domain-level modeling. *U.S. Bureau of Labor Statistics*, 2023.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268, 2007.

N. L. Hjort, F. A. Dahl, and G. H. Steinbakk. Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101:1157–1174, 2006.

J. Li, D. J. Nott, Y. Fan, and S. A. Sisson. Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Computational Statistics and Data Analysis*, 106:77–89, 2017.

M. Modrák, A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari. Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. *Bayesian Analysis*, 2024.

J. M. Robins, A. van der Vaart, and V. Ventura. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95:1143–1156, 2000.

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

T. Savitsky, J. Gershunskaya, and A. Gelman. Simulation-based calibration of uncertainty intervals under approximate bayesian estimation. *U.S. Bureau of Labor Statistics*, 2024.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. 2020. URL `http://www.stat.columbia.edu/~gelman/research/unpublished/sbc.pdf`.

Y. Yao. Meta-Bayes. *Flatiron Institute*, 2024.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13:917–1003, 2018.

Y. Yao, G. Pirš, A. Vehtari, and A. Gelman. Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, 17:1043–1071, 2022.

X. Yu, D. J. Nott, M.-N. Tran, and N. Klein. Assessment and adjustment of approximate inference algorithms using the law of total variance. *Journal of Computational and Graphical Statistics*, 30:977–990, 2021.