Understanding posterior recalibration for a simple example*

Andrew Gelman†, Philip Greengard, Julie Gershunskaya, Terrance Savitsky, Ben Goodrich

21 Dec 2023

Abstract

Approximate computation algorithms for Bayesian inference can be evaluated using simulation-based calibration checking (SBC). When SBC reveals a problem, posterior inferences for parameters and scalar summaries can be *recalibrated* based on the departure of quantiles from their desired uniform distribution. Gershunskaya et al. (2023) have reported success following this procedure for a hierarchical model, but averaging over the posterior predictive distribution, an approach which has the benefit of focusing the calibration 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 calibration. Applying the recalibration procedure to the posterior predictive distribution has the effect of reducing the pooling toward the prior.

1. Background

1.1. Simulation-based calibration checking

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, $p(\theta|y) \propto p(\theta)p(y|\theta)$. For computational reasons, the analyst cannot easily work with the posterior and so summarize it using draws θ^s_{post}, $s = 1, \dots, S$, from an approximate posterior, $g(\theta|y)$, obtained using some computational procedure such as variational inference (Kucukelbir et al., 2017).

One way to evaluate this computational procedure is through simulation-based calibration checking (SBC; Modrák et al., 2023), which proceeds as follows. First, repeat the following computations L times: (a) Simulate a "true parameter value" θ^l from $p(\theta)$, (b) simulate a dataset y^l from $p(y|\theta^l)$, and (c) perform approximate inference to obtain S independent draws $\theta^s_{\text{post}}, s = 1, \dots, S$, from $g(\theta|y)^l$. Second, for each of the L replications compute the quantile of the true θ^l amid the S draws θ^s_{post}. If the simulations are calibrated, the L values of this quantile should be approximately uniformly distributed in the set $\{0, 1, \dots, S\}$. The distribution of the L quantiles can be compared visually to a uniform density or summarized analytically, for example by performing the normal-cdf transformation, $z^l = \Phi^{-1}(\frac{1}{S+1}(\text{quantile}^l + 0.5))$, and comparing the mean and variance of these quantiles to their approximate expected values of 0 and 1 under calibration.

*We thank the U.S. Bureau of Labor Statistics and National Science Foundation for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York.

1.2. Recalibration

The next logical step is to rescale the approximate simulation draws to approximately calibrate the process. Gershunskaya et al. (2023) propose and evaluate methods to do this, but with another twist: 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, replacing $g(\theta|y)$ by $p(\theta|y)$ in step (c) above. Nonetheless, Gershunskaya et al. (2023) find their procedure to work reasonably well to recalibrate variational inferences for a hierarchical model of applied interest.

In the present paper we work out the details of simulation-based recalibration for a simple one-dimensional normal-normal model, a simple setting where it is possible to obtain exact posterior draws. 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. Posterior predictive recalibration results in inferences that differ from the posterior by doing less pooling toward the prior. After going through this example, we discuss implications for simulation-based calibration and consider why posterior recalibration could work well for hierarchical models, even while being so far off in a simple non-hierarchical setting.

1.3. Connection to general ideas of statistical workflow

Simulation-based experimentation is a general way to understand and evaluate statistical methods: Start with an assumption about the world, simulate data, apply your statistical procedure to the simulated data, compare to the "true" or assumed model. Perform that procedure many times in parallel, and you would like something close to nominal coverage of uncertainty intervals.

In general, calibration checking can be performed to address a range of concerns: nonidentified or weakly-identified parameters, misfit of model to data, computational approximations, and flat-out computation or data errors. Even in the ideal setting where we assume error-free data and computation and a fully-specified Bayesian model, there is still the question in predictive checking of attributing misfit to data or prior distribution (Sinharay and Stern, 2003). From a workflow perspective, simulation-based checking can be viewed as a sort of omnibus test, with further steps required to resolve any problems that are found.

When working within a parametric model and thinking from a classical perspective, we would like our inference to recover, to its nominal precision, the true parameter vector θ. From a Bayesian perspective, however, simulation-based calibration should only work when averaging over the prior distribution. For example, if the assumed true θ is in the outskirts of the prior, then we would expect Bayesian inference to pull any estimated θ toward massy areas of the prior, so that with small sample size or weak identification, this true but extreme θ would not be well recovered.

How this all fits into Bayesian workflow is unclear, given that it is standard practice to use a weak priors as a placeholder, replacing it with a more informative model if necessary. So it could make sense to take a model with a weak ("broad" or "consensus") prior but then perform simulation-based experimentation with a stronger prior that is more "realistic" or focused on the problem at hand. Indeed, the basic approach of supposing a particular "true" value of θ could be considered as a special or limiting case of the use of a sharper prior for evaluation than for modeling.

2. Understanding posterior calibration for a one-parameter normal model

We can examine the differences between prior and posterior calibration 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 calibration, even approximately.

We consider a simple normal-normal model with one scalar parameter and one data point.

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

This setup is more general than it might look, as y can be taken to be not just one observation but rather as a point estimate with the problem scaled so its standard error equals 1. 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.

2.1. 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 simulations have the following distribution:

$$\theta^l \sim \text{normal}(0, 1) \tag{2}$$
$$y^l|\theta^l \sim \text{normal}(\theta^l, 1) \tag{3}$$
$$\theta^s_{\text{post}}|y^l \sim \text{normal}(y^l/2, 1/\sqrt{2}). \tag{4}$$

In the limit of large S, the position of θ^l within the simulations of θ^s_{post} is

$$z^l = \sqrt{2}(\theta^l - y^l/2). \tag{5}$$

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 (2) and (3). The result is that $\theta^l - y^l/2 \sim \text{normal}(0, 1/\sqrt{2})$, hence the z-score (5) has a unit normal distribution in the limit of large number of simulation draws S.

2.2. 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 (2) is replaced with:

$$\theta^l \sim \text{normal}(y/2, 1/\sqrt{2}), \tag{6}$$

where y is the observed data. The z-score of θ^l within the simulations $\theta_s, s = 1, \ldots, S$ is still given by (5); the only difference is that, instead of averaging over (2) and (3), we average over (6) and (3). In this posterior predictive distribution, $z^l = \sqrt{2}(\theta^l - y^l/2) \sim \text{normal}(y/(2\sqrt{2}), \sqrt{3}/2)$. For no value of y will this be a unit normal distribution, thus we would *not* see predictive calibration (a uniform distribution of the L quantiles), even if the computation is exact.

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

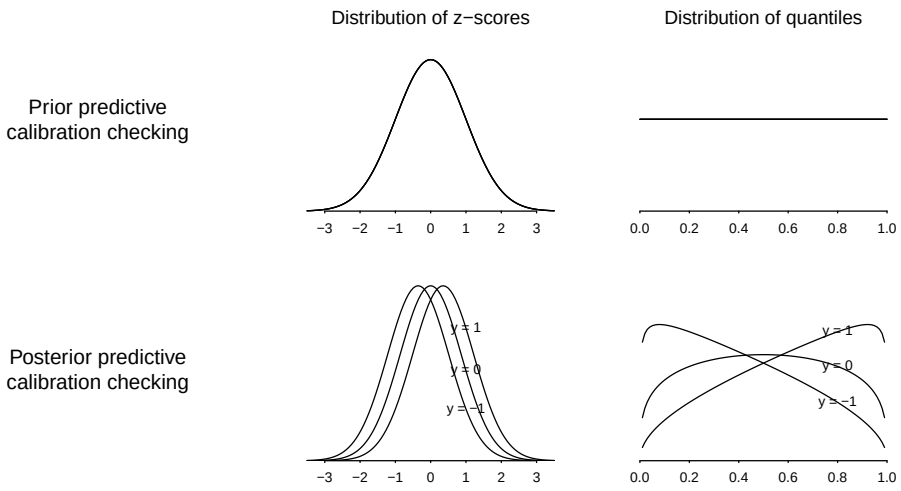

Figure 1: *Distribution of z-scores and quantiles in simulation-based calibration checking when the correct model is being fit for the simple normal example, under two scenarios.* Top row: *Simulating the parameter θ^l from the prior distribution.* Bottom row: *Simulating θ^l from the posterior distribution,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

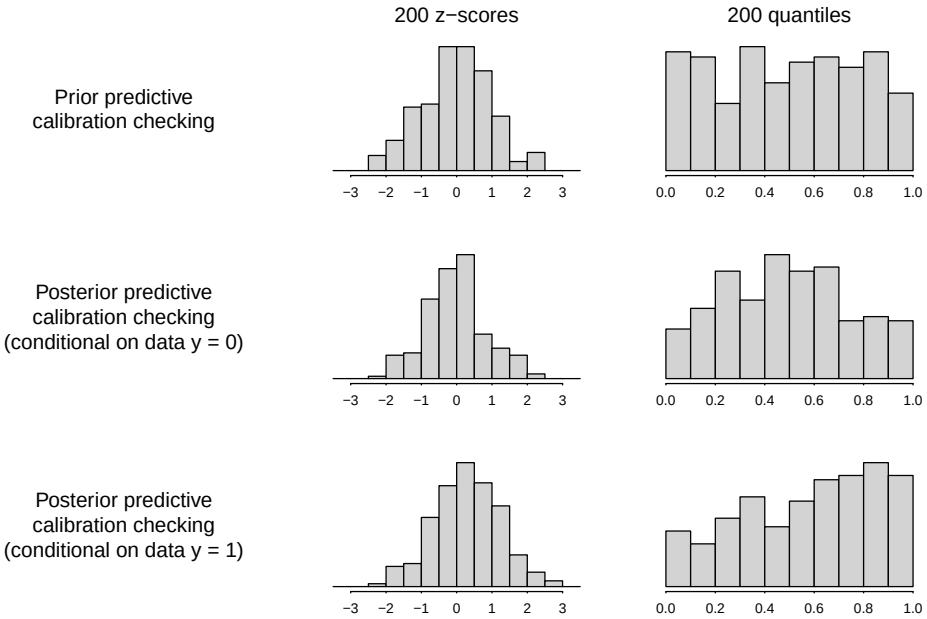

Figure 2: *Simulation-based calibration checking using 200 replications of 1000 simulation draws for each.* Top row: *Simulating the parameter θ^l from the prior distribution.* Center and bottom rows: *Simulating θ^l from the posterior distribution, in which case the distributions depend on the data, y. As predicted by theory, the prior predictive simulations show calibration and the posterior predictive simulations do not.*

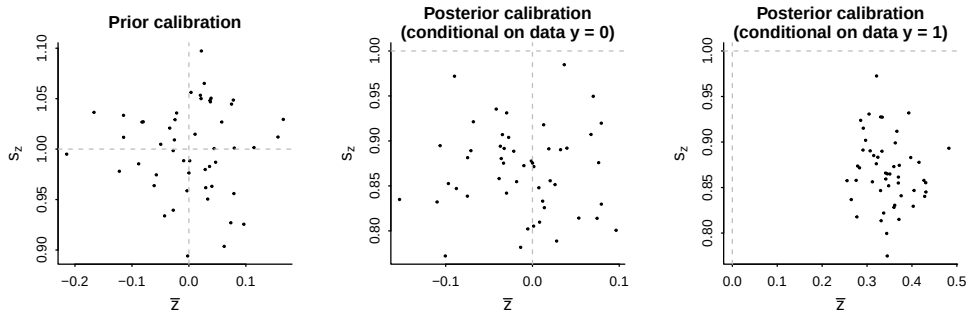

Figure 3: *Estimated mean and scale shifts from simulation-based calibration using 50 replications of 1000 simulation draws. In each scatterplot, the dots correspond to 100 independent simulations of the calibration process.* Left: *Simulating the parameter θ^l from the prior distribution; adjustments 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.*

Figure 2 shows corresponding results from a simulation with $L = 200$ 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 the fact that there is no approximation in the simulations.

2.3. Prior calibration

The idea of simulation-based calibration is to use any miscalibration found in the checking to adjust the simulation draws. Here we work with one of the methods of Gershunskaya et al. (2023) using a location-scale shift. The first step is to perform calibration checking, obtaining z-scores $z_l, l = 1, \ldots, L$, summarizing 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 θ^s_{post} by the factor s_z and shifting them by the relative value $\bar{z}$. In this affine transformation, the S simulation draws are replaced by these adjusted draws:

$$\theta^{\text{adj},s}_{\text{post}} = \bar{\theta}_{\text{post}} + s_z(\theta^s_{\text{post}} - \bar{\theta}_{\text{post}}) + \bar{z}s_{\text{post}}\,, \tag{7}$$

where $\bar{\theta}_{\text{post}}$ and s_{post} are the mean and standard deviation of the S values, θ^s_{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.

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 3 shows the results; most of the time the mean shift is less than 0.1 and the scale shift is less than 5%.

2.4. Posterior calibration

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 3 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 (6). 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 calibration 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.

3. Discussion

3.1. Prior or posterior calibration?

Our motivation for simulation-based calibration 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 calibration 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 very 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 calibration to adjust for systematic errors in the calibration procedure itself. Similar issues arise with the bootstrap, another class of procedures that uses simulation to correct for inferential biases (Efron, 1982).

The fundamental problem with posterior calibration—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).

3.2. Why posterior calibration can approximately work for multilevel models

We conjecture that posterior recalibration works so well in that example 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. In that case, posterior draws of α for J new groups are essentially the same as prior draws. 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 calibration checking should look a lot like prior calibration checking for a hierarchical model.

3.3. A halfhearted defense of posterior calibration in our problem

This seems 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, 2023).

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 calibration could serve some robustness goal. Alternatively, perhaps posterior calibration could itself be recalibrated in some way.

3.4. 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 calibration 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 calibration. 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 calibration could be a useful tool in hierarchical models, as suggested by Gershunskaya et al. (2023).

References

Cockayne, J., Graham, M. M., Oates, C. J., Sullivan, T. J., and Teymur, O. (2022). Testing whether a learning procedure is calibrated. *Journal of Machine Learning Research* **23**, 303.

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

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

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

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733–807.

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

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

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

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research* **18**, 14.

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2023). Simulation-based calibration checking for Bayesian computation: The choice of test quantities shapes sensitivity. *Bayesian Analysis*.

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

Sinharay, S., and Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference* **111**, 209–221.

Yao, Y. (2023). Meta-Bayes. Technical report.

Yao, Y., Pirš, G., Vehtari, A., and Gelman, A. (2018). Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis* **17**, 1043–1071.

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