Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity[*]

Martin Modrák[†], Angie H. Moon[‡], Shinyoung Kim[§], Paul Bürkner[¶],
Niko Huurre[‖], Andrew Gelman[**], Aki Vehtari[††]

14 Oct 2022

Abstract

Simulation-based calibration checking (SBC) is a practical method to validate computationally-derived posterior distributions or their approximations. In this paper, we introduce a new variant of SBC to alleviate several known problems. Our variant allows the user to in principle detect any possible issue with the posterior, while previously reported implementations could never detect large classes of problems including when the posterior is equal to the prior. This is made possible by including additional data-dependent test quantities when running SBC. We argue and demonstrate that the joint likelihood of the data is an especially useful test quantity. Some other types of test quantities and their theoretical and practical benefits are also investigated. We support our recommendations with numerical experiments on a multivariate normal example and theoretical analysis of SBC, thereby providing a more complete understanding of the underlying statistical mechanisms. From the theoretical side, we also bring attention to a relatively common mistake in the literature and clarify the difference between SBC and checks based on the data-averaged posterior. The SBC variant introduced in this paper is implemented in the `SBC` R package.

1. Introduction

Simulation-based calibration checking (SBC; Talts et al. 2020) is a method to validate Bayesian computation, extending ideas from Cook, Gelman, and Rubin (2006).[1] While it is primarily intended for validating sampling algorithms such as MCMC, it can be used for validating any method implementing or approximating Bayesian inference. Published applications include variational inference (Yao et al. 2018) and neural posterior approximations (**radev2020bayesflow**).

Throughout this paper we assume an implicit and fixed (Bayesian) statistical model π with data space Y and parameter space Θ. For $y \in Y, \theta \in \Theta$ the model implies the following joint, marginal,

---

[†]Institute of Microbiology of the Czech Academy of Sciences

[‡]Sloan School of Management, MIT

[§]Department of Computer Science, Kookmin University

[¶]Cluster of Excellence SimTech, University of Stuttgart

[‖]Unaffiliated

[**]Department of Statistics and Political Science, Columbia University

[††]Department of Computer Science, Aalto University

[1]The term in the literature is "simulation-based calibration"; here we have added the word "checking" to emphasize that these methods do not themselves produce calibration; rather, they measure departure from calibration.

and posterior distributions:

$$\pi_{\text{joint}}(y, \theta) = \pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)$$

$$\pi_{\text{marg}}(y) = \int_{\Theta} d\theta \, \pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)$$

$$\pi_{\text{post}}(\theta|y) = \frac{\pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)}{\pi_{\text{marg}}(y)}.$$

Typically, the posterior distribution (π_{post}) is the target of inference but is impossible to evaluate directly. While many computational approaches exist for sampling from the posterior or its approximations, they may fail to provide a correct answer. Problems can arise from errors in how the algorithm or the statistical model are encoded or from inherent inability of the computational method to handle a given model with a given dataset.

1.1. Self-consistency of Bayesian models

To discover problems with computation, several classes of checks can be derived from self-consistency properties of statistical models. One such property is used by the data-averaged posterior (Geweke 2004):

$$\pi_{\text{prior}}(\theta) = \int_Y dy \int_{\Theta} d\tilde{\theta} \, \pi_{\text{post}}(\theta|y)\pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}). \tag{1}$$

Where confusion is possible, we use $\tilde{\theta}$ to denote a parameter vector drawn from the prior distribution.

SBC relies on a different property that involves the joint distribution of prior and posterior samples from the same model (Cook, Gelman, and Rubin 2006):

$$\pi_{\text{SBC}}(y, \theta, \tilde{\theta}) = \pi_{\text{prior}}(\tilde{\theta})\pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{post}}(\theta|y). \tag{2}$$

Since $\pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}) = \pi_{\text{marg}}(y)\,\pi_{\text{post}}(\tilde{\theta}|y)$, this implies,

$$\pi_{\text{SBC}}(y, \theta, \tilde{\theta}) = \pi_{\text{marg}}(y)\pi_{\text{post}}(\theta|y)\pi_{\text{post}}(\tilde{\theta}|y). \tag{3}$$

Equation (3) immediately shows that conditional on a specific data $y \in Y$, the distributions of θ and $\tilde{\theta}$ in Equations (2) and (3) are identical.

SBC and related methods employ two different implementations of the same statistical model and check if the results have the same distribution conditional on data. The first step is to define a *generator* capable of directly simulating draws from $\pi_{\text{prior}}(\tilde{\theta})$ and $\pi_{\text{obs}}(y|\tilde{\theta})$, and the second step is to define a *probabilistic program* that, in combination with a given *posterior approximation algorithm*, samples from the posterior distribution $\pi_{\text{post}}(\theta|y)$. Each simulation from the generator yields,

$$\tilde{\theta} \sim \pi_{\text{prior}}(\tilde{\theta})$$

$$y \sim \pi_{\text{obs}}(y|\tilde{\theta})$$

$$\theta_1, \ldots \theta_M \sim \pi_{\text{post}}(\theta|y), \tag{4}$$

where M is the number of posterior draws sampled. We run many such simulations and then inspect the realized distributions of θ and $\tilde{\theta}$ conditional on y. Specific calibration methods differ in how exactly they check for conditional equality of the two distributions.

1.2. Proposed SBC variant

We argue for the following variant of the SBC check: First, project the potentially high-dimensional parameter and data space into a scalar *test quantity* $f : \Theta \times Y \to \mathbb{R}$. Second, compute the ranks of the prior draws in the posterior conditional on y. Specifically, we take the number of posterior sample draws where the test quantity is lower than in the prior sample, and, if there are any ties, choosing the rank randomly among the tied positions:

$$N_{\texttt{less}} := \sum_{m=1}^{M} \mathbb{I}\left(f(\theta_m, y) < f(\tilde{\theta}, y)\right)$$

$$N_{\texttt{equals}} := \sum_{m=1}^{M} \mathbb{I}\left(f(\theta_m, y) = f(\tilde{\theta}, y)\right)$$

$$K \sim \text{uniform}(0, N_{\texttt{equals}})$$

$$N_{\texttt{total}} := N_{\texttt{less}} + K.$$

The procedure simplifies if there are no ties, which will be true for most practical test quantities over models with continuous parameter space as in this case $N_{\texttt{total}} = N_{\texttt{less}}$. Then, if the probabilistic program and the generator implement the same probabilistic model, we have

$$N_{\texttt{total}} \sim \text{uniform}(0, M). \tag{5}$$

See Theorems 2 and 4 in Appendix A for a formal statement and proof. As a result, once the empirical distribution of $N_{\texttt{total}}$ is computed from the simulations, we can perform a test for uniformity. The process is then repeated for all test quantities we want to consider. If we are using MCMC to sample from π_{post}, the posterior sample typically needs to be thinned to ensure that $\theta_1, \ldots, \theta_M$ are approximately independent (Talts et al. 2020). The overall SBC process is illustrated in Figure 1.

While it is possible to use numerical tests for uniformity with SBC, we generally prefer to use visualisations of the rank distribution as they are more informative than numerical summaries and discourage dichotomous thinking. Most prominent are rank histograms and plots of empirical cumulative distribution functions (Säilynoja, Bürkner, and Vehtari 2021).

Our proposed SBC variant improves upon the way SBC has been previously reported and used in two major ways:

- We let test quantities depend on both data and parameters, while previous work only considered quantities that depend on the parameters. In practice, these test quantities were almost exclusively just the individual parameters themselves.

- Previous formulations of SBC required uniformity of $N_{\texttt{less}}$. However, even if the probabilistic program is exactly correct, $N_{\texttt{less}}$ will not be uniform if $\Pr(N_{\texttt{equals}} > 0) > 0$, that is, if ties can occur. With our improved SBC procedure, we can handle test quantities that have distributions with point masses and thus ties between $f(\tilde{\theta}, y)$ and $(f(\theta_1, y), \ldots, f(\theta_M, y))$. Resolving ties lets us use SBC for models with discrete parameters as well as in some other special cases, such as when a theoretically strictly positive test quantity suffers underflow and some prior/posterior sample draws become zeroes numerically.

1.3. Practical considerations

SBC will be satisfied if three components are in harmony: the generator and the probabilistic program should correspond to the same data-generating process, and the posterior approximation
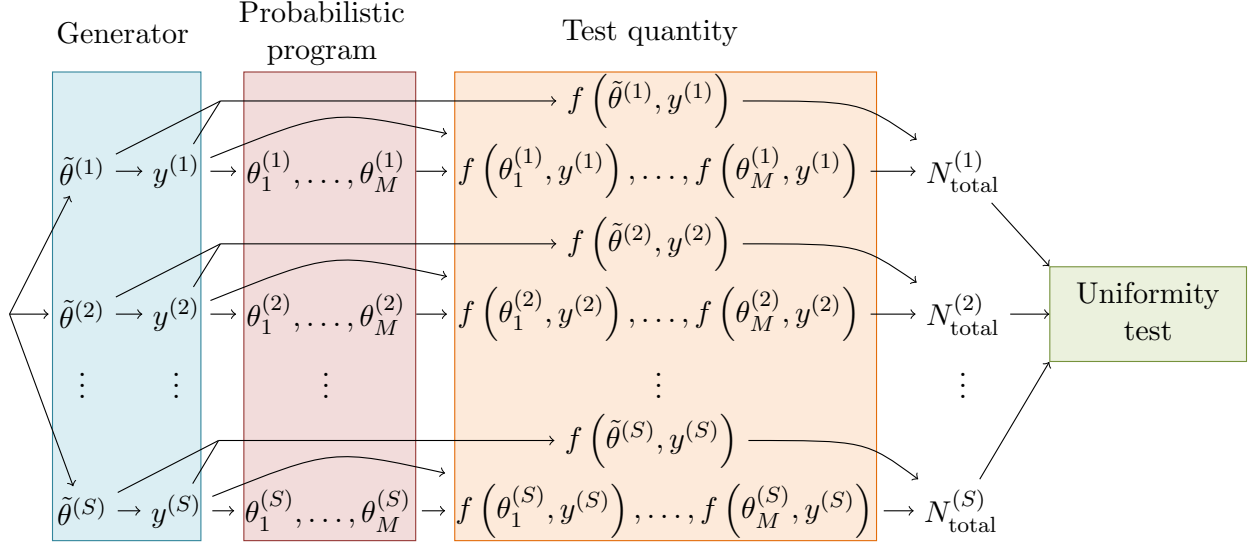
Figure 1: *Schematic representation of SBC with S simulations. The generator is responsible for generating draws from the prior distribution $\tilde{\theta} \sim \pi_{prior}(\tilde{\theta})$ and from the observational model $y \sim \pi_{obs}(y \mid \tilde{\theta})$. The draws from the observational model are then treated as input for the probabilistic program which takes M posterior draws $\theta_1, \ldots \theta_M \sim \pi_{post}(\theta|y)$ or its approximation. Each test quantity projects both the prior draw and the posterior draws (potentially using data) onto the real line, letting us compute a single rank (N_{total}). Finally, deviations from discrete uniform distribution are assessed statistically or visually.*

algorithm (including the associated tuning parameters) should sample from the correct posterior (or with negligible bias) for the probabilistic program, given the data simulated from the prior. Failure indicates that at least one of the components is mismatched to the others. However, by itself, SBC cannot determine where exactly the problem lies. As a result, two broad classes of SBC use cases arise:

- We have code is available to simulate data and a probabilistic program we trust, and the goal is to check that the program correctly computes the posterior.

- We have an algorithm that we trust is correct and trustworthy code to simulate data, and the goal is to check that we correctly implemented our probabilistic program.

In practice, those classes overlap and mix: we are rarely completely certain of the correctness of any algorithm, generator, or probabilistic program. SBC as a simulation method has no way to inform us about a discrepancy between the process that generated real data and the assumptions of our statistical model. For reliable inference, SBC thus needs to be combined with other elements of Bayesian workflow that can detect model misspecification, such as posterior predictive checks or analysis of residuals (Gabry et al. 2019; Gelman et al. 2020; Kay 2021).

1.4. Importance of test quantities

It has been generally believed that methods based on Equation (2), including SBC, are never sensitive to some classes of mismatches between the generator and the probabilistic program—most notably that it is impossible to detect if the probabilistic program samples from the prior

distribution and ignores the information in the data (e.g., Equation (1.3) of Lee, Nicholls, and Ryder 2019, although with incorrect reasoning; Appendix M.2 of Lueckmann et al. 2021; Schad et al. 2022; Zhao et al. 2021; Ramesh et al. 2022; Cockayne et al. 2020).

In this paper we show that the choice of test quantities greatly influences the usefulness and sensitivity of SBC. We show that using test quantities that depend on data makes it possible to detect any conceivable mismatch between the generator and the probabilistic program. Thus, we demonstrate that the belief in inherent limitations of SBC has relied on overly restrictive and sometimes plainly incorrect assumptions. We discuss useful classes of test quantities that have not been used so far and provide characterization of possible remaining undetected failures. We provide simulation studies as well as theoretical analysis of SBC to support our findings. We hope that our theoretical framework can serve as a basis for a better understanding of the properties of SBC and related methods. All of the techniques discussed are implemented in the `SBC` R package (Kim et al. 2022).

The rest of the paper is structured as follows: Section 2 discusses related work, Section 3 summarizes the theoretical results we derived, Section 4 shows results of simulation experiments and Section 5 discusses the results and our recommendations for practical use of SBC.

2.   Related work

Prior contributions to validation of Bayesian computation can be roughly split into works that focus on the data-averaged posterior, those that focus on the SBC property, and other relevant works that do not directly invoke any self-consistency property.

2.1.   Data-averaged posterior

The idea of using simulations via a generator to verify Bayesian computation can be traced back to Geweke (2004) who compared the moments of the prior and the data-averaged posterior distributions for multiple test quantities. The paper proposed to integrate a transition kernel for an MCMC sampler targeting π_{post} into a scheme that samples π_{joint} directly. This lets us obtain the data-averaged posterior from a single run of this sampler, potentially reducing the computational cost but increasing implementation burden. Geweke's formalism allows the test quantities to depend on data, although all the examples actually shown only depend on parameters. Comparing the mean vector and covariance matrix of the prior distribution and the data-averaged posterior distribution is also discussed by Yu et al. (2021), who use repeated fits to build the data-averaged posterior.

2.2.   SBC-like checks

The identity of prior and posterior distributions conditional on a specific dataset as a tool to check computation was proposed by Cook, Gelman, and Rubin (2006), and further refined by Talts et al. (2020) who introduced SBC as it is currently used. Specific variants of SBC have been proposed for variational inference (Yao et al. 2018), Bayes factors (Schad et al. 2022), neural posterior approximators (Radev, D'Alessandro, et al. 2021; Radev, Mertens, et al. 2022), and Gaussian processes (Mcleod and Simpson 2021).

Gandy and Scott (2020) proposed a procedure similar to SBC that can work with shorter sequences of Markov transitions than a full fit, reducing computational cost. This is, however, less relevant for algorithms that need a nontrivial warmup phase to adapt to the specific posterior (e.g., the adaptive Hamiltonian Monte Carlo sampler implemented in Stan; Carpenter et al. 2017). This

is because warmup is a fixed cost that occurs during every model fit even if fewer post-warmup draws are needed.

Prangle et al. (2014) proposed an SBC-like procedure for approximate Bayesian computation (ABC). They note that the possibility that the probabilistic program simply samples from the prior distribution cannot be ignored in this context and resolve this issue by separately inspecting ranks for some subsets of the simulated datasets. SBC is closely related to the *coverage property* discussed by Prangle et al.: when using M posterior sample draws, SBC can be understood as checking for all posterior intervals of width $\alpha \in \{\frac{1}{M}, \ldots, \frac{M-1}{M}\}$ that the probability the interval contains the original simulated value of the test quantity is α.

A broader framework for calibration of learning procedures has been proposed by Cockayne et al. (2020). There, Bayesian inference is just one example of procedures where calibration can be empirically verified with an SBC-like check. They distinguish between *strong calibration* which corresponds to passing SBC (specifically continuous SBC as defined in Appendix A) for all measurable functions and *weak calibration* which corresponds to having a correct data-averaged posterior (Equation 1). They however only consider test quantities that depend only on parameters.

2.3. Miscellaneous

The problem of diagnosing and understanding computational issues is transformed by Rendsburg et al. (2022). Their approach tries to find a prior distribution that would make the probabilistic program and algorithm exactly match the generator.

Domke (2021) propose to use fits to multiple generated datasets to estimate the symmetrized KL-divergence between a distributional approximation to the correct posterior (e.g., Laplace or variational inference) and the true posterior.

While almost all of the related works focus on one-dimensional test quantities, it is also possible to use multivariate test quantities. To make this practical, Saad et al. (2019) discusses a family of goodness of fit tests for multivariate uniformity. We believe most of the results presented in this paper generalize relatively straightforwardly to multivariate test quantities.

3.   Summary of theoretical results

Due to necessity to build a relatively large theoretical machinery we have moved precise statements of theorems and proofs to Appendix A and here only summarise the main results in plain English.

3.1. Correctness

We show that passing our SBC variant is a necessary condition for the correctness of posterior estimation (Theorems 2 and 4). That is, the correct posterior will always produce uniformly distributed ranks, including for test quantities that may have ties.

3.2. Characterization of SBC failures

Still, many incorrect posteriors will also pass SBC for any given test quantity, so in Theorem 3 we characterize those situations. Not only does the correct posterior result in uniform distribution of ranks when averaging over the whole data space Y, but the ranks are uniformly distributed even when we only consider simulations that yielded data in some $\bar{Y} \subset Y$. The reverse implication also holds: when the ranks are uniformly distributed for all subsets of the data space $\bar{Y} \subset Y$, then the implied posterior distribution of the test quantity under investigation has to be exactly correct. In other words, whenever SBC "fails" and the implied posterior distribution of a given test quantity is

incorrect although the rank distribution is uniform, we can find a subset of the data space, where the ranks are non-uniform. It just so happens that all the deviations in various subsets cancel each other out perfectly.

An obvious application of Theorem 3 is that we could partition our simulations based on some features of the data space and investigate uniformity separately for each part, similarly to the procedure suggested by Prangle et al. (2014). This however quickly runs into issues of multiple testing due to the lower number of simulations in each part. It is thus in our experience not practical except for the special case that interest lies only in some subset of the data space, so that the SBC checks can focus only on that data space of interest. This is a form of rejection sampling and can be practically useful if it is easy to formulate a criterion that constrains plausible real data sets but hard to construct a defensible prior distribution that would enforce this criterion implicitly. For example, prior information can be available on the plausible variance of an outcome across the whole population, which may be hard to express as a prior on coefficients associated with predictors.

3.3. Data-dependent test quantities

The characterization of SBC failures discussed above an intuition why test quantities that depend on data are useful: If SBC passes for a test quantity f, but the posterior is in fact incorrect, we can always pick a test quantity g that combines f with some aspect of the data and ensures that the discrepancies in various parts of data space add up instead of canceling out. For example, we could have over-abundance of low ranks and under-abundance of high ranks in $Y_1 \subset Y$ and a matching under-abundance of low ranks and over-abundance of high ranks in $Y_2 \subset Y$. Setting

$$g(\theta, y) = \begin{cases} -f(\theta, y) & y \in Y_1 \\ f(\theta, y) & \text{otherwise} \end{cases}$$

will ensure over-abundance of high ranks in both Y_1 and Y_2. Since such a test quantity uses all the simulations, we do not lose power from reduced number of simulations.

An even stronger reason to use data-dependent test quantities is that they make SBC in some sense complete: If there is any difference between the correct posterior and the posterior implemented by the probabilistic program, there will exist a data-dependent test quantity that fails SBC (Theorem 6). The proof is constructive and provides a specific test quantity that detects failures, which is the ratio of the correct posterior density to the posterior density actually implemented by the probabilistic program. This is not a practical test quantity, as it (a) depends on the specific probabilistic program we implemented and (b) requires that we already have the correct posterior density. However our empirical results in this paper, and our experience with using SBC in model development more generally, shows that the model likelihood $\pi_{\text{obs}}(y|\theta)$ is frequently useful as a general-purpose test quantity. This makes sense intuitively, as the likelihood is an important contributor to the above density ratio.

3.4. Ignoring data

We generalize the result that probabilistic program sampling from the prior distribution will pass SBC against all test quantities that do not depend on data (Theorem 7). Specifically, a probabilistic program that ignores some aspect of the data—for example, does not take some specific data points into account, rounds all data to integers, etc.—but handles the reduced dataset correctly will pass SBC of all test quantities that do not depend on the ignored aspect of the data.

3.5. Detailed analysis of simple models and test quantities

Appendix B provides full theoretical analysis of SBC for simple models and test quantities where we can actually characterize all possible posterior distributions that will satisfy SBC. This is aimed at providing intuition on what SBC actually does and also serve as counterexamples to some claims. In some literature (e.g., Lee, Nicholls, and Ryder 2019, Lueckmann et al. 2021, Schad et al. 2022, Grinsztajn et al. 2021, Ramesh et al. 2022), it is assumed that SBC is based on the data-averaged posterior (1). We show that this is incorrect: Example 2 not only explicitly constructs posterior distributions that will satisfy (1) for some test quantity while not passing SBC, but also posterior distributions that pass SBC while not satisfying (1). One possibly more general lesson is that SBC is most naturally understood as enforcing constraints on the quantile function of the test quantity while having a correct data-averaged posterior is most naturally seen as constraint on the density of the test quantity.

This implies that there might be some gains in using both the data-averaged posterior and SBC when verifying the correctness of Bayesian computation. We however suspect that the additional practical benefit of using the data-averaged posterior is small in the sense that the incorrect posteriors that pass SBC but are ruled out by (1) are mostly contrived and unlikely to be a result of an inadvertent mistake. Lemma 2.19 of Cockayne et al. (2020) proves that if a posterior passes SBC for *all possible* test quantities that do not depend on data, it will have correct data averaged posterior for all test quantities that do not depend on data, so SBC is stronger at least in the limit of using infinitely many test quantities. We leave a more thorough examination of the relationship between data-averaged posterior and SBC as future work.

3.6. Monotonic transformations of test quantities

Finally, transforming a test quantity by a strictly monotonic function results in equivalent SBC results; see Theorem 8. The result cannot be easily strengthened as many non-monotonic transformations lead to different, non-equivalent SBC checks. Example 3 shows that flipping the ordering of values only for some subset of the data space yields a different SBC check. Example 4 shows that we can also obtain a different check if we combine a test quantity with a non-monotonic bijection, and Example 5 shows the same for the case when a whole range of values is projected onto a single point. In all those examples, the transformed test quantities rule out some sets of posteriors that pass SBC for the original quantity, but there are also sets of posteriors not passing SBC for the original quantity but pass SBC for the transformed quantity.

4. Numerical experiments

The theoretical analysis in previous section primarily deals with behavior of SBC in the limit of both infinitely many posterior draws per fit and infinitely many simulations. Here, we further support the results by numerical experiments which let us understand not only whether a certain problem is detectable at all but also how much computational effort is required for SBC to detect the problem. Code for simulations and all the figures in this section can be found at `https://github.com/martinmodrak/sbc_test_quantities_paper`.

4.1. Setup

To illustrate some of the properties of various types of test quantities, we will use a simple multi-variate normal model,

$$\mu \sim \text{MVN}(0, \boldsymbol{\Sigma})$$
$$\mathbf{y}_1, \ldots, \mathbf{y}_n \sim \text{MVN}(\mu, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \tag{6}$$

where the two-element vector μ is the target of inference, while $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are observed. Introducing $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i$, the correct analytic posterior is MVN $\left(\frac{N\bar{\mathbf{y}}}{n+1}, \frac{1}{n+1} \boldsymbol{\Sigma} \right)$. Unless mentioned otherwise we will use $n = 3$. In most previous use-cases of SBC, the only test quantities used would have been the parameters themselves, that is, the elements of μ in the above example. Below, we also check a host of derived quantities: the sum, difference and product of the μ elements, the joint likelihood of all the data, and pointwise likelihoods for the first two data points.

To quantify the discrepancy between an observed distribution of posterior ranks and the uniform distribution, we take the likelihood of observing the most extreme point on the empirical CDF if the rank distribution was indeed uniform:

$$\gamma = 2 \min_{i \in \{1, \ldots, M+1\}} (\min\{\text{Bin}(R_i|S, z_i), 1 - \text{Bin}(R_i - 1|S, z_i)\}). \tag{7}$$

Here, M is the number of draws in the sample obtained from the posterior, S is the number of simulations (and thus the number of observed ranks), $z_i = \frac{i}{M+1}$ is the expected proportion of observed ranks smaller than i, R_i is the observed count of ranks smaller than i, and $\text{Bin}(R|S, p)$ is the CDF of the binomial distribution with S trials and probability of success p evaluated at R. This metric was introduced in a paper by Säilynoja, Bürkner, and Vehtari (2021), where we can also find computational methods to evaluate the distribution of γ under uniform distribution of ranks for given M and S. If you are working within a hypothesis-testing framework, your primary metric of interest could then be $\log \frac{\gamma}{\bar{\gamma}}$, where $\bar{\gamma}$ is the 5th percentile of the null distribution. That is, $\log \frac{\gamma}{\bar{\gamma}} < 0$ implies a rejection of the hypothesis of uniform distribution at the 5% level so that visual checks of the ECDF plots would show problems (for a single test quantity). This diagnostic is typically more sensitive than the Kolmogorov-Smirnoff or χ^2 test.

Figure 2 shows how this statistic evolves in a fairly typical SBC run as we add more simulations using a probabilistic program that samples from the correct posterior. There is some variability, but most of the time all quantities would indicate uniformity and if they indicate some non-uniformity, the discrepancies tend to be small so we are unlikely to reject this model as incorrect.

4.2. Ignoring data

For comparison, Figure 3 shows the evolution of the same quantities for a typical run with an incorrect posterior that is completely equal to the prior. All quantities that do not depend on data pass SBC, barring small short-term deviations as seen for the correct posterior. But all the likelihood-based quantities start showing big discrepancies after just a handful of simulations. While the overall distribution of ranks for the parameters themselves is uniform, when we look separately at data with large average y and low average y, the ranks are strongly non-uniform in both regions (Figure 4).

We observe similar behaviour for the posterior that ignores only the first data point; see Figure 5. The biggest difference is that that now the pointwise likelihood for the second data point—which
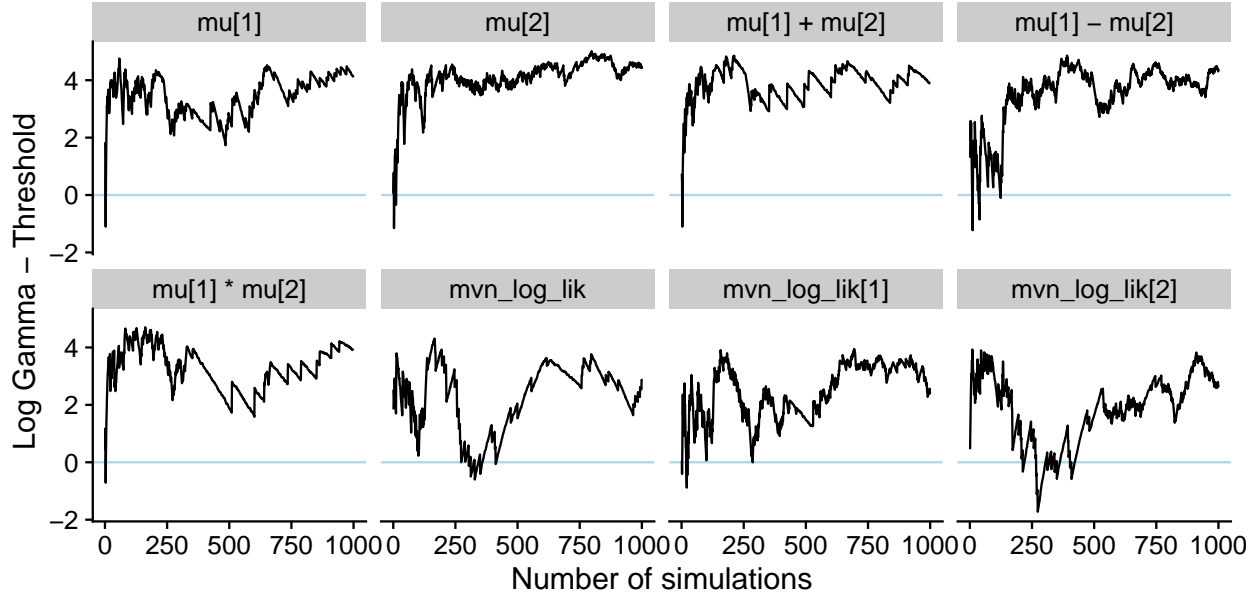
Figure 2: *Evolution of the difference between the gamma statistic and threshold ($\log \bar{\gamma}$) for rejecting uniformity at 5% for the correct posterior.* `mvn_log_lik[1]` *and* `mvn_log_lik[2]` *are the pointwise likelihoods* $\pi(\mathbf{y}_1|\mu)$ *and* $\pi(\mathbf{y}_2|\mu)$ *respectively, while* `mvn_log_lik` *is the joint likelihood. As expected for using a 5% level for rejection, false positives (values below the threshold) do happen, but they tend to produce only small discrepancies.*
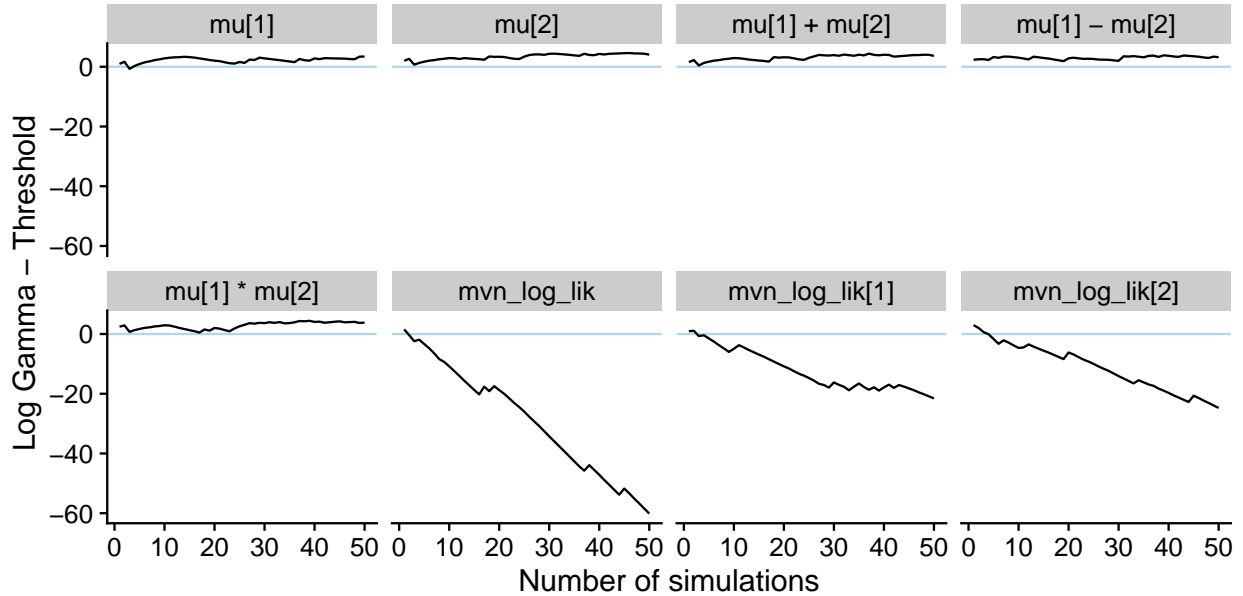

Figure 3: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for an incorrect posterior that equals the prior.*

10
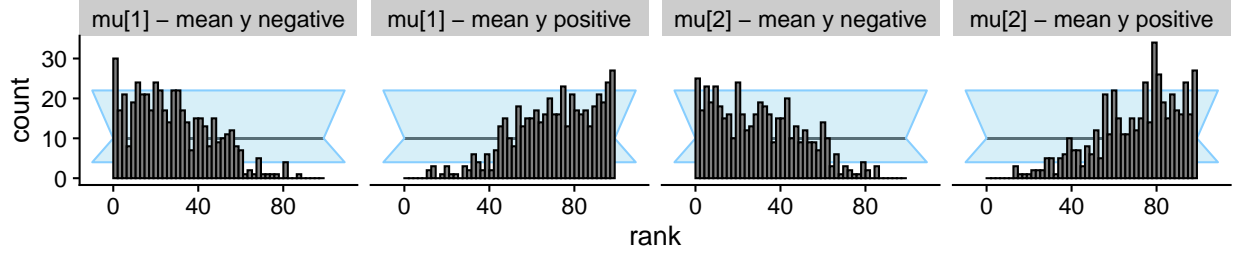
Figure 4: *Rank distribution for the elements of μ split by the average value of the corresponding y elements for the incorrect posterior that is completely equal to the prior. The distributions for the two cases exactly compensate to make the overall distribution uniform.*

Figure 5: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for an incorrect posterior that ignores a single datapoint among a small data set (n = 3).*

Figure 6: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for an incorrect posterior that ignores a single datapoint among a larger dataset ($n = 20$).*
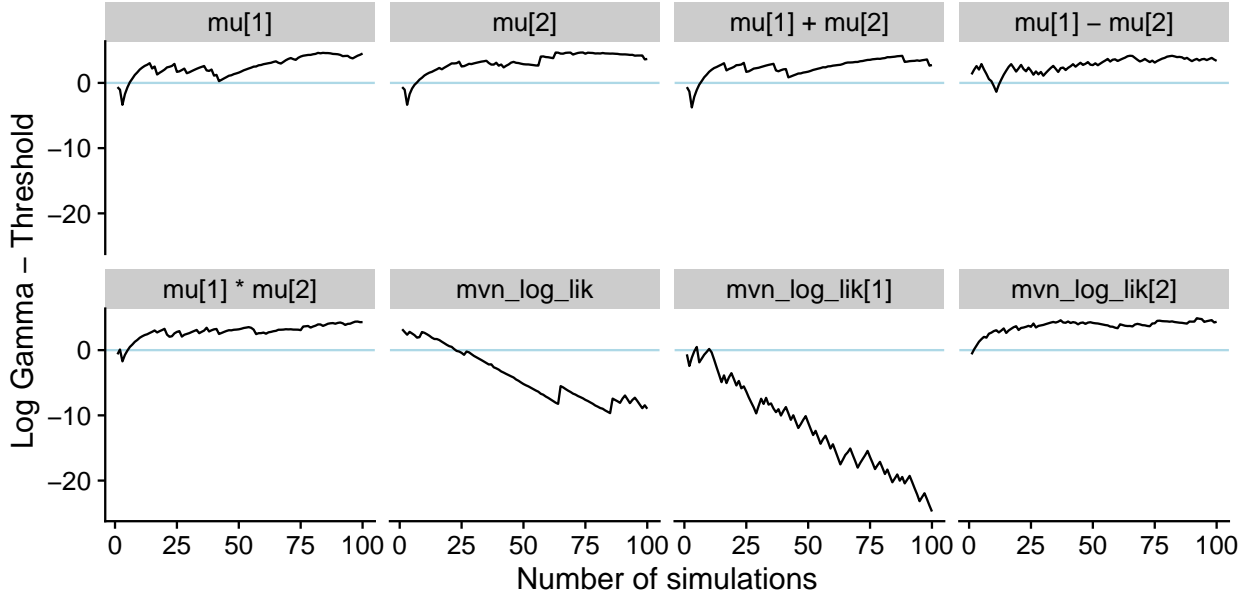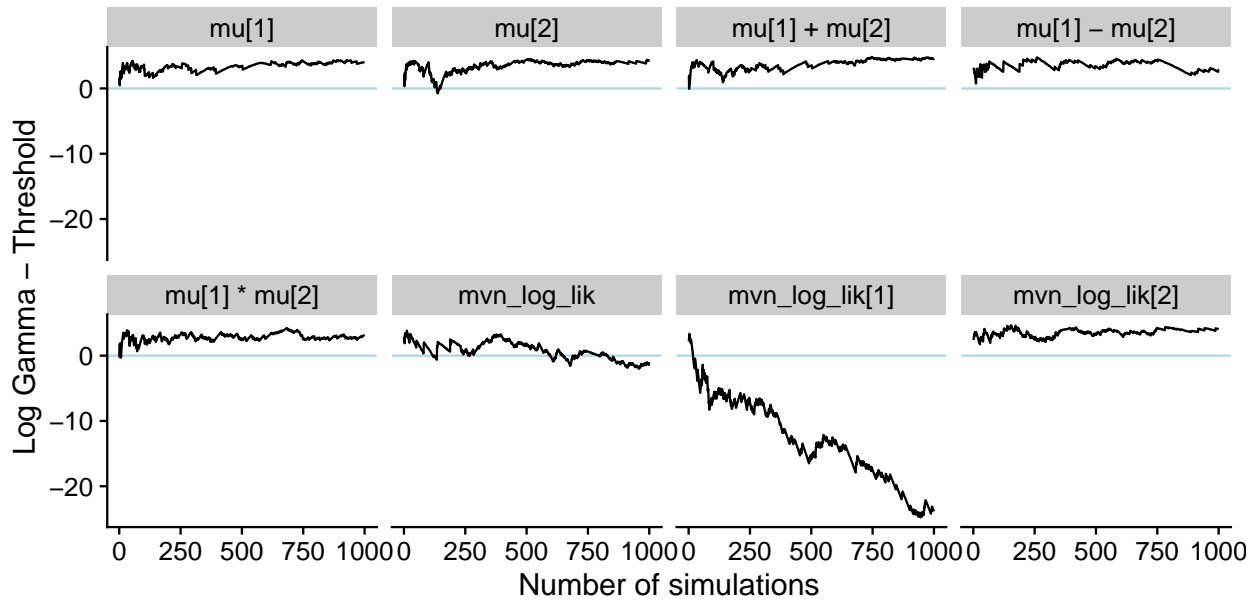
was not ignored—passes SBC while the joint likelihood as well as the pointwise likelihood for the first ignored data point show problems. Additionally, the pointwise likelihood for the ignored data point now shows bigger discrepancy than the joint likelihood. For both quantities, the discrepancy is smaller and requires about $S = 20$ simulations to reliably uncover, because ignoring a single data point produces a posterior that is closer to the correct one than when ignoring all the data. If we increase the number of data points to $n = 20$ (Figure 6), ignoring just a single data point produces a posterior that is close to correct and even after 1000 simulations, the discrepancy for the joint likelihood is small. The pointwise likelihood for the first (ignored) data point still detects the problem relatively quickly.

More generally, if the model (partially) ignores data, then adding a function that involves both data and parameters can detect this failure. Specifically adding the joint log-likelihood of the data as a derived quantity seems to be a useful default. If only a small part of the data is missing, using the joint-likelihood in SBC will turn it into a problem of precision. Missing just a single datapoint in a large dataset (e.g., an off-by-one error in the probabilistic program) may change the posterior only slightly and be undetectable with realistic computational effort.

4.3. Incorrect correlations

Suppose we have an incorrect posterior that has the correct marginal distributions for both parameters, and sampling is done independent normal distributions, $\mu_i \mid y_1, \ldots, y_n \sim N\left(\frac{n\bar{y}_i}{n+1}, \frac{1}{n+1}\Sigma_{i,i}\right)$. The evolution of the discrepancy as simulations are added is shown in Figure 7. If the test quantities are the univariate parameters, SBC passes without any indication of problems, while the likelihood-based quantities as well as the difference, product, and sum of the variables show problems relatively quickly. The joint likelihood is the first to show serious issues.

If the inference does not represent correlations in the posterior correctly, this should as well
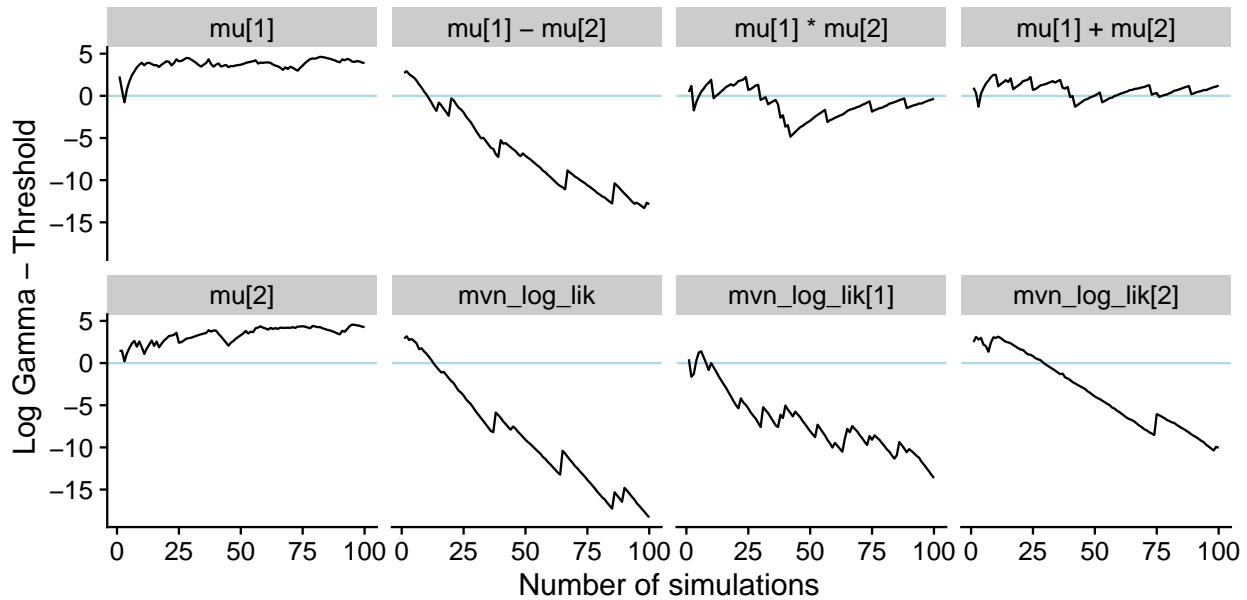
12

Figure 7: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for incorrect posterior that has wrong correlation structure.*

manifest in an SBC failure for some function of the parameters. This can be directly targeted by using products ("interactions") of model parameters, but the log-likelihood once again seems to be generally useful as a highly nonlinear function of all model parameters.

4.4. Less plausible problems

In this subsection our results get less practical and more theoretical. The (partially) unused data case may easily arise in practice due to a bug in the probabilistic program such as an indexing bug or a deficient overall approach. For example, an approximate Bayesian computation algorithm may not learn from the data at all and just stick to the prior (Prangle et al. 2014). Incorrect correlations or more general higher-order structure of the posterior may also easily arise due to a problem with an approximate inference algorithm. For example, mean-field variational inference will never recover any correlations by design. Beyond those examples, we have found it hard to find incorrect probabilistic programs that would satisfy the SBC identity and could plausibly arise from unintentional mistakes in program code or problems with an algorithm. We see this as anecdotal evidence that SBC augmented with a few well-chosen test quantities that probe usage of data and higher order posterior structure such as the likelihood can rather robustly detect these kinds of mistakes. That said, for specific models, wide sets of artificial counterexamples that incorrectly pass SBC can be constructed.

A specific case of a more general class of setups is that we can create an incorrect posterior approximation that produces overabundance of low ranks for datasets with average of $\mathbf{y}$ positive and compensates by producing overabudnance of high ranks for other datasets. If this is done right, the test quantity will pass SBC. The distribution of the ranks conditional on the average of $\mathbf{y}$ for one such setup is shown in Figure 8. As seen in Figure 9, when averaging over all datasets, SBC indeed passes for the univariate parameter test quantities, but if we instead look at absolute value of μ (as well as some other non-monotonic transformations of μ), we immediately see problems as now some
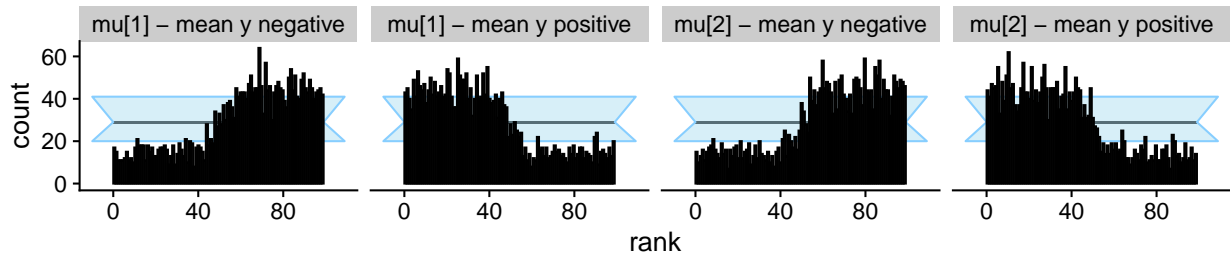
13

Figure 8: *Rank distribution for the elements of μ split by the average value of the corresponding y elements for the incorrect posterior that satisfies SBC for individual parameters. The distributions for the two cases exactly compensate to make the overall distribution uniform.*

of the previously low ranks flip to high ranks and the discrepancies accumulate instead of canceling each other. In this particular case, the problem is also eventually picked up by the product of the μ values and with enough simulations even by the joint likelihood, but there is no guarantee this will always happen. In general, non-monotonic transformations can discover incorrect posteriors that would be otherwise hidden when looking at the original variables. Still, the practical relevance of non-monotonic transforms in SBC is, in our view, likely limited, as it required careful work to construct posteriors that manifested this behaviour. We were unable to find even remotely plausible scenarios where an issue with Bayesian computation was best discovered by using a non-monotonic transformation of another test quantity.

4.5. Small discrepancies

A final case study considers small discrepancies in the posterior. To be specific, we introduce a small bias in the posterior drawn from $\text{normal}(0, 0.3)$ independently for each simulation and element of μ. The resulting SBC history is shown in Figure 10. While all of the monitored quantities will eventually show the problem, the likelihood-based quantities and the difference of μ do that noticeably sooner than others. This demonstrates that derived quantities can somewhat improve precision of SBC: small changes in the univariate marginals can result in big (and thus easy to detect) changes for some test quantities combining the univariate marginals with data and other parameters.

5. Conclusions

5.1. Choosing test quantities for SBC

We have found that enriching the repertoire of test quantities used in SBC provides both qualitative and quantitative improvements to the ability of SBC to detect problems in Bayesian computation. For practical use of SBC in everyday model and algorithm development, we recommend to use by default the individual model parameters as test quantities as well as the joint likelihood of the data and potentially a small number of other quantities.

Individual parameters are recommended as they are always immediately available and are able to diagnose a large number of problems with a posterior approximation. Also, the parameters are themselves often of primary interest for inference, so it is desirable to check that their uncertainty is correctly calibrated.

The joint likelihood is a highly useful quantity to detect the types of problems discussed in

Figure 9: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for incorrect posterior that satisfies SBC for individual parameters but not their absolute values. Note the different horizontal axis between top row (showing quantities that detect the problem slowly or not at all) and bottom row (quantities that detect the problem quickly). The vertical red dashed line marks 500 simulations in all subplots. For brevity we only show quantities derived from the first element of μ; the situation is analogous for the second element. The* `drop(mu[1])` *quantity is defined as μ_1 if $\mu_1 < 1$ and as $\mu_1 - 5$ otherwise.*
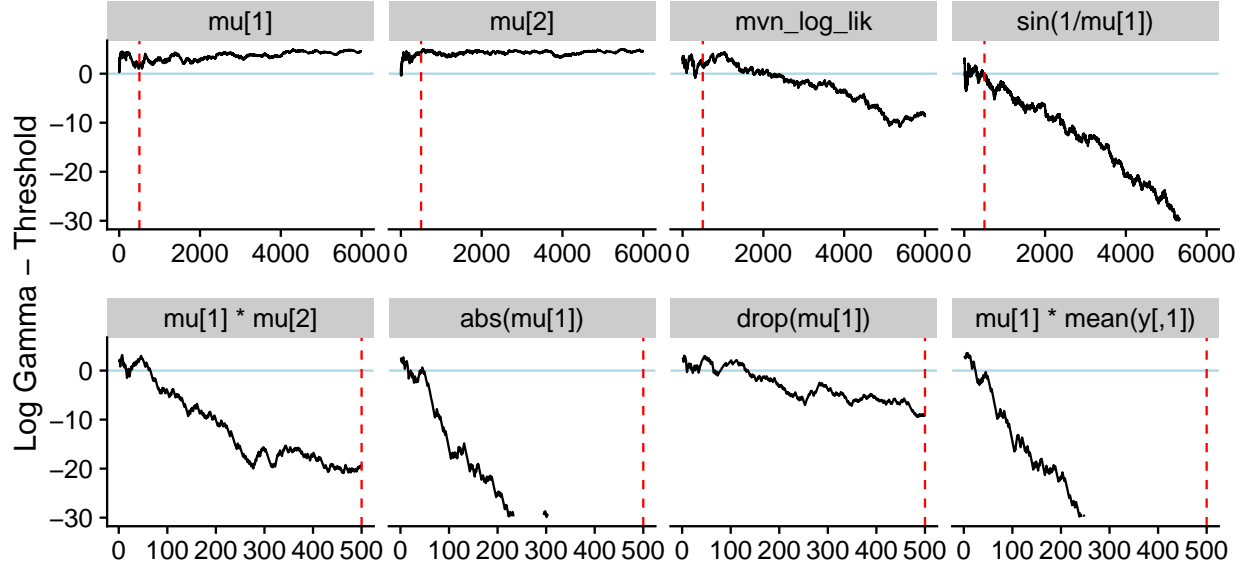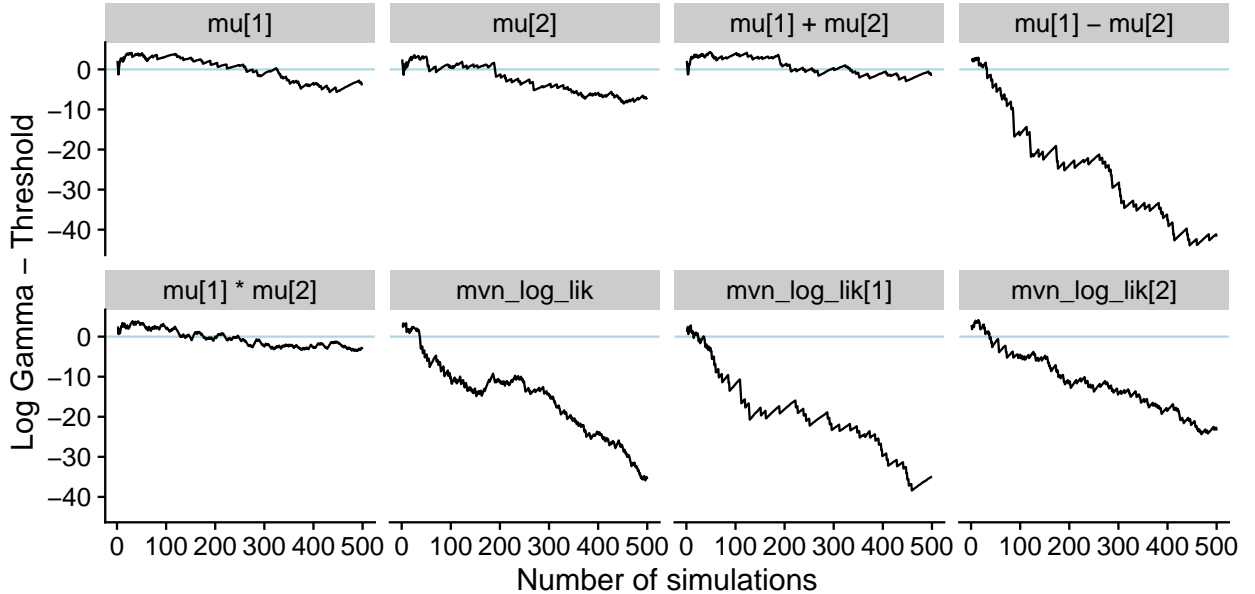

Figure 10: *Evolution of the difference between the gamma statistic and threshold for rejecting uniformity at 5% for incorrect posterior that introduces a small bias for each parameter.*

15

Section 4 (especially ignoring data and incorrect correlations). In all of the cases presented in our simulations, the joint likelihood was able to detect the discrepancies and in many cases it was even able to detect them with the fewest simulations among all considered quantities. While, for some specific problems, we could find quantities that are more sensitive than the joint likelihood, none other was useful in all cases. Section 3.3 provides theoretical justification for why we could expect this to hold frequently and not only in the examples we discussed. We think this generality makes the joint likelihood a good default quantity to monitor in SBC. If not using all the data correctly is a potential issue (e.g., because the code handling the data is particularly complex), then adding selected likelihoods for subsets of the data might also be sensible.

It also makes sense to add test quantities tailored to the specific inferential goals we have built the model for (e.g., some specific model predictions). These quantities often let us implicitly check the correctness of parameter correlations or other dependency structures and safeguard the user against problems that they care about the most. If correlations or other dependencies in the posterior are directly of interest, then pairwise products or differences of the model parameters can also be sensible test quantities.

5.2. Limitations

Although we have shown that SBC can in principle diagnose any problem, limitations for practical use remain. For nontrivial models, adding a finite number of test quantities cannot guard against all possible ways the SBC identity may be satisfied by an incorrect posterior. However, as we check more quantities, the potential counterexamples become contrived, hard to construct, and unlikely to be the result of an inadvertent bug in model code. At the same time, adding more test quantities increases the risk of false SBC failures simply due to the number of tests performed (if no corrections for multiple comparisons are made for the SBC checks) or it may reduce the overall power of the check (if corrections for multiple comparisons are made), so choosing test quantities carefully remains important.

This problem could potentially be alleviated by improving our understanding of the expected dependency structure of different test quantities' uniformity checks, as then we could correct for multiple comparisons without loosing that much power. We leave that as future work. In practice, we have seen similarity in the degree of uniformity violation between different test quantities using the same inputs, making the need for multiple comparison correction less urgent.

Moreover, there are practical limitations imposed by the fact that we always have only limited computational resources for SBC: We can produce only a limited number of simulated datasets to fit the model on and only a limited number of posterior draws per fitted model. Both contribute to the stringency and precision of the uniformity test we can perform.

Additional test quantities do not help much with precision problems—if the posterior is close to correct, the test quantities will also be close to correct. Although in some cases, some test quantities can slightly increase the sensitivity of the check by combining multiple parameters, so small imprecisions in each of the parameters can get compounded (once again the nonlinearity of the likelihood seems to be at least sometimes useful in this regard).

5.3. Implications for non-SBC checks

As a contribution to the broader discussion about validation of Bayesian computation, we show that SBC and data-averaged posterior provide different checks, despite being repeatedly conflated in the literature (see Section 3.5). We leave a more detailed comparison of SBC and data-averaged posterior as future work, although there are some tentative arguments to believe that SBC provides

stricter checks.

SBC is not the only approach to validating Bayesian computation that relies on choosing specific test quantities—test quantities are fundamental to the methods of Geweke (2004), Cook, Gelman, and Rubin (2006), Prangle et al. (2014), Gandy and Scott (2020), and Cockayne et al. (2020). We suspect that many of the considerations regarding their choice for SBC are applicable more generally also in these other approaches.

References

Carpenter, Bob et al. (2017). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1. URL: `https://www.jstatsoft.org/index.php/jss/article/view/v076i01`.

Cockayne, Jon et al. (2020). "Testing whether a learning procedure is calibrated". In: DOI: `10.48550/ARXIV.2012.12670`. URL: `https://arxiv.org/abs/2012.12670`.

Cook, Samantha R., Andrew Gelman, and Donald B. Rubin (2006). "Validation of software for Bayesian models using posterior quantiles". In: *Journal of Computational and Graphical Statistics* 15.3, pp. 675–692. URL: `http://www.jstor.org/stable/27594203`.

Domke, Justin (2021). "An easy to interpret diagnostic for approximate inference: Symmetric divergence over simulations". In: DOI: `10.48550/ARXIV.2103.01030`. URL: `https://arxiv.org/abs/2103.01030`.

Gabry, Jonah et al. (2019). "Visualization in Bayesian workflow". In: *Journal of the Royal Statistical Society: Series A* 182, pp. 389–402. DOI: `10/ftqws`.

Gandy, Axel and James Scott (2020). "Unit testing for MCMC and other Monte Carlo methods". In: DOI: `10.48550/ARXIV.2001.06465`. URL: `https://arxiv.org/abs/2001.06465`.

Gelman, Andrew et al. (2020). "Bayesian workflow". In: URL: `https://arxiv.org/abs/2011.01808`.

Geweke, John (2004). "Getting it right". In: *Journal of the American Statistical Association* 99, pp. 799–804.

Grinsztajn, Léo et al. (2021). "Bayesian workflow for disease transmission modeling in Stan". In: *Statistics in Medicine* 40, pp. 6209–6234.

Kay, Matthew (Dec. 30, 2021). "Extracting and visualizing tidy residuals from Bayesian models". In: *Documentation for R package tidybayes*. URL: `http://mjskay.github.io/tidybayes/articles/tidybayes-residuals.html`.

Kim, Shinyoung et al. (2022). "SBC: Simulation based calibration for rstan/cmdstanr models". In: URL: `https://github.com/hyunjimoon/SBC/`.

Lee, Jeong Eun, Geoff K. Nicholls, and Robin J. Ryder (2019). "Calibration procedures for approximate Bayesian credible sets". In: *Bayesian Analysis* 14, pp. 1245–1269. URL: `https://projecteuclid.org/journals/bayesian-analysis/volume-14/issue-4/Calibration-Procedures-for-Approximate-Bayesian-Credible-Sets/10.1214/19-BA1175.full`.

Lueckmann, Jan-Matthis et al. (2021). "Benchmarking simulation-based inference". In: *Proceedings of Machine Learning Research* 130, pp. 343–351. URL: `https://proceedings.mlr.press/v130/lueckmann21a.html`.

Mcleod, John and Fergus Simpson (2021). "Validating Gaussian process models with simulation-based calibration". In: *2021 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pp. 101–102.

Prangle, D. et al. (2014). "Diagnostic tools for approximate Bayesian computation using the coverage property". In: *Australian & New Zealand Journal of Statistics* 56, pp. 309–329. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12087`.

Radev, Stefan T., Marco D'Alessandro, et al. (2021). "Amortized Bayesian model comparison with evidential deep learning". In: *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15. DOI: `10.1109/TNNLS.2021.3124052`.

Radev, Stefan T., Ulf K. Mertens, et al. (2022). "BayesFlow: Learning complex stochastic models with invertible neural networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.4, pp. 1452–1466. DOI: `10.1109/TNNLS.2020.3042395`.

Ramesh, Poornima et al. (2022). "GATSBI: Generative adversarial training for simulation-based inference". In: URL: `http://arxiv.org/abs/2203.06481`.

Rendsburg, Luca et al. (2022). "Discovering inductive bias with Gibbs priors: A diagnostic tool for approximate Bayesian inference". In: *Proceedings of Machine Learning Research* 151, pp. 1503–1526. URL: `https://proceedings.mlr.press/v151/rendsburg22a.html`.

Saad, Feras A. et al. (2019). "A family of exact goodness-of-fit tests for high-dimensional discrete distributions". In: *Proceedings of Machine Learning Research* 89, pp. 1640–1649. URL: `https://proceedings.mlr.press/v89/saad19a.html`.

Säilynoja, Teemu, Paul-Christian Bürkner, and Aki Vehtari (2021). "Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison". In: URL: `http://arxiv.org/abs/2103.10522`.

Schad, Daniel J. et al. (2022). "Workflow techniques for the robust use of Bayes factors". en. In: *Psychological Methods*. DOI: `10.1037/met0000472`. URL: `http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000472`.

Talts, Sean et al. (2020). "Validating Bayesian inference algorithms with simulation-based calibration". In: URL: `http://www.stat.columbia.edu/~gelman/research/unpublished/sbc.pdf`.

Yao, Yuling et al. (2018). "Yes, but did it work?: Evaluating variational inference". In: *Proceedings of Machine Learning Research* 80, pp. 5581–5590. URL: `https://proceedings.mlr.press/v80/yao18a.html`.

Yu, Xuejun et al. (2021). "Assessment and adjustment of approximate inference algorithms using the law of total variance". In: *Journal of Computational and Graphical Statistics* 30, pp. 977–990. URL: `https://doi.org/10.1080/10618600.2021.1880921`.

Zhao, David et al. (2021). "Diagnostics for conditional density models and Bayesian inference algorithms". In: *Proceedings of Machine Learning Research* 161, pp. 1830–1840. URL: `https://proceedings.mlr.press/v161/zhao21b.html`.

Appendix A: Formalized theory and proofs

We will denote the integral of $f(x)$ w.r.t. x over domain X as $\int_X \mathrm{d}x\, f(x)$. $\mathbb{I}(P)$ is the indicator function for a given predicate P. When a function can be understood as describing a conditional probability distribution, we will use | to separate the function arguments we condition on. This is only to assist comprehension and has the same semantic meaning as using a comma.

In all cases, we assume an underlying statistical model π which decomposes into a prior and observational model. Given a data space Y and a parameter space Θ, then for $y \in Y, \theta \in \Theta$ the model implies the following joint, marginal and posterior distributions:

$$\pi_{\text{joint}}(y, \theta) = \pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)$$

$$\pi_{\text{marg}}(y) = \int_\Theta \mathrm{d}\theta\, \pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)$$

$$\pi_{\text{post}}(\theta|y) = \frac{\pi_{\text{obs}}(y|\theta)\pi_{\text{prior}}(\theta)}{\pi_{\text{marg}}(y)}.$$

Unless noted otherwise, all definitions and proofs implicitly assume a single model π is given.

Definition 1 (Posterior family). Given a data space Y and a parameter space Θ, the posterior family ϕ assigns a normalized posterior density to each possible $y \in Y$. I.e. posterior family is a function $\phi : \Theta \times Y \to \mathbb{R}^+$ such that

$$\forall y : \int \mathrm{d}\theta \, \phi(\theta|y) = 1.$$

For each y, we will denote the implied distribution over Θ as ϕ_y.

Definition 2 (Test quantity). Given a data space Y and a parameter space Θ a *test quantity* is any measurable function $f : \Theta \times Y \to \mathbb{R}$.

Definition 3 (Sample rank CDF, sample Q, sample SBC). Given a data space Y, a parameter space Θ, a test quantity f, $M \in \mathbf{N}$ and a posterior family ϕ. For any $y \in Y$, if $\theta_1, \ldots, \theta_M \sim \phi_y$ we can define the following random variables:

$$N_{\mathtt{less}} := \sum_{m=1}^{M} \mathbb{I}\left(f(\theta_m, y) < f(\tilde{\theta}, y)\right)$$

$$N_{\mathtt{equals}} := \sum_{m=1}^{M} \mathbb{I}\left(f(\theta_m, y) = f(\tilde{\theta}, y)\right)$$

$$K \sim \mathrm{uniform}(0, N_{\mathtt{equals}})$$

$$N_{\mathtt{total}} := N_{\mathtt{less}} + K.$$

For a fixed $\tilde{\theta}$, we define the *M-sample rank CDF* as:

$$R_{\phi,f}(i|\tilde{\theta}, y) := \Pr\left(N_{\mathtt{total}} \leq i\right).$$

Averaging over the correct posterior, we can define the *M-sample Q* as:

$$Q_{\phi,f}(i|y) := \int_{\Theta} \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y) R_{\phi,f}(i|\tilde{\theta}, y)$$

We then say that ϕ *passes* M-sample SBC w.r.t. f if $\forall i \in 0, \ldots, M-1$ we have

$$\int_Y \mathrm{d}y \, Q_{\phi,f}(i|y)\pi_{\mathrm{marg}}(y) = \frac{i+1}{M+1}.$$

This definition does not match immediately with the procedure we actually use to run SBC in practice but is more convenient for further analysis. We now prove the equivalence with the SBC procedure:

Theorem 1 (Procedural definition of sample SBC). Given a data space Y, a parameter space Θ, a test quantity f, $M \in \mathbf{N}$ and a posterior family ϕ. Assuming $\tilde{\theta} \sim \pi_{\mathrm{prior}}(\theta)$, $y \sim \pi_{obs}(y|\tilde{\theta})$ and $\theta_1, \ldots, \theta_M \sim \phi_y$, ϕ passes M-sample SBC w.r.t. f if and only if $N_{\mathtt{total}} \sim \mathrm{uniform}(0, M)$.

Proof. We show the equivalence of CDF for $N_{\mathtt{total}}$ with the formula in the definition of sample SBC. For all $i \in \{0, \ldots, M\}$ we obtain

$$\Pr(N_{\mathtt{total}} \leq i) = \int_{\Theta} \mathrm{d}\tilde{\theta} \int_Y \mathrm{d}y \, \pi_{\mathrm{prior}}(\tilde{\theta})\pi_{\mathrm{obs}}(y|\tilde{\theta})R_{\phi,f}(i|\tilde{\theta}, y) =$$

$$= \int_Y \mathrm{d}y \int_{\Theta} \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y)\pi_{\mathrm{marg}}(y)R_{\phi,f}(i|\tilde{\theta}, y) = \int_Y \mathrm{d}y \, Q_{\phi,f}(i|y)\pi_{\mathrm{marg}}(y)$$

$\square$

However, the sample SBC is not particularly amenable to direct analysis as the choice of M can matter. We will thus focus on a continuous case, which can be understood as the limit of the sample SBC as $M \to \infty$.

Definition 4 (Predictive CDF, tie probability, quantile functions). Given a model π, a data space Y, a parameter space Θ, a posterior family ϕ, and a test quantity f, we define

- fitted CDF: $C_{\phi,f}^\pi : \bar{\mathbb{R}} \times Y \to [0,1]$,
  $C_{\phi,f}^\pi(s|y) := \int_\Theta \mathrm{d}\theta\, \mathbb{I}\left(f\left(\theta, y\right) \leq s\right) \phi\left(\theta, y\right)$

- true CDF: $C_f^\pi : \bar{\mathbb{R}} \times Y \to [0,1]$,
  $C_f^\pi(s|y) := \int_\Theta \mathrm{d}\theta\, \mathbb{I}\left(f\left(\theta, y\right) \leq s\right) \pi_{\mathrm{post}}(\theta|y)$

- fitted quantile function: $C_{\phi,f}^{\pi,-1} : [0,1] \times Y \to \bar{\mathbb{R}}$,
  $C_{\phi,f}^{\pi,-1}(x|y) := \inf\{s : x \leq C_{\phi,f}(s|y)\}$

- true quantile function: $C_f^{\pi,-1} : [0,1] \times Y \to \bar{\mathbb{R}}$,
  $C_f^{\pi,-1}(x|y) := \inf\{s : x \leq C_f(s|y)\}$

- fitted tie probability: $D_{\phi,f}^\pi : \bar{\mathbb{R}} \times Y \to [0,1]$,
  $D_{\phi,f}^\pi(s|y) := \int_\Theta \mathrm{d}\theta\, \phi(\theta|y)\mathbb{I}\left(f(\theta, y) = s\right)$

- true tie probability: $D_f^\pi : \bar{\mathbb{R}} \times Y \to [0,1]$,
  $D_f^\pi(s|y) := \int_\Theta \mathrm{d}\theta\, \pi_{\mathrm{post}}(\theta|y)\mathbb{I}\left(f(\theta, y) = s\right)$,

where $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ is the extended real number line. When no confusion arises, the model superscript will be omitted.

Definition 5 (Continuous rank CDF, continuous q, continuous SBC). Assume a data space Y, a parameter space Θ, a test quantity f, and a posterior family ϕ. For fixed $\tilde{\theta} \in \Theta$ and $y \in Y$ we define the *continuous rank CDF* $r_{\phi,f} : [0,1] \times \Theta \times Y \to [0,1]$ as

$$r_{\phi,f}(x|\tilde{\theta}, y) := \Pr\left(C_{\phi,f}\left(f\left(\tilde{\theta}, y\right)\Big| y\right) - uD_{\phi,f}\left(f\left(\tilde{\theta}, y\right)\Big| y\right) \leq x\right),$$

assuming u is a random variable distributed uniformly over the $[0,1]$ interval. If the fitted tie probability is 0, then r is a step function and the implied rank distribution is degenerate.

Averaging over the correct posterior, we can define the *continuous* $q : [0,1] \times Y \to [0,1]$ as

$$q_{\phi,f}(x|y) := \int_\Theta \mathrm{d}\tilde{\theta}\, \pi_{\mathrm{post}}(\tilde{\theta}|y)r_{\phi,f}(x|\tilde{\theta}, y).$$

We then say that ϕ *passes* continuous SBC w.r.t. f if

$$\forall x \in [0,1] : \int_Y \mathrm{d}y\, q_{\phi,f}(x|y)\pi_{\mathrm{marg}}(y) = x. \tag{8}$$

In both the sample and the continuous case, the presence of ties introduces analytical difficulties. We thus start by a useful lemma that lets us avoid dealing with ties for many purposes.

Definition 6 (Ties)**.** Given a test quantity $f : \Theta \times Y \to \mathbb{R}$ a model π *has ties* w.r.t. f when there exists $y \in Y, \tilde{\theta} \in \Theta$ such that

$$D_f(f(\tilde{\theta}, y)|y) > 0.$$

A posterior family ϕ *has ties* w.r.t. f when there exists $y \in Y, \tilde{\theta} \in \Theta$ such that

$$D_{\phi,f}(f(\tilde{\theta}, y)|y) > 0.$$

Lemma 1 (Tie removal)**.** Given a model π, posterior family ϕ, and test quantity function $f : \Theta \times Y \to \mathbb{R}$, where either π has prior ties w.r.t. f or ϕ has ties w.r.t. f, we can construct a model π', with parameter space Θ' and data space Y, posterior family ϕ' and test quantity f' such that (1) π' has no ties w.r.t. f' and ϕ' has no ties w.r.t. f', (2) $\forall \theta \in \Theta, y \in Y : \phi(\theta|y) = \pi_{\text{post}}(\theta|y)$ if and only if $\forall \theta \in \Theta', y \in Y : \phi'(\theta, y) = \pi'_{\text{post}}(\theta|y)$ and (3) both sample Q and continuous q are maintained. That is,

$$\forall y \in Y, x \in [0, 1] : q^\pi_{\phi,f}(x|y) = q^{\pi'}_{\phi,f'}(x|y)$$

$$\forall y \in Y, i \in 0, \ldots, M - 1 : Q^\pi_{\phi,f}(i|y) = Q^{\pi'}_{\phi,f'}(i|y).$$

Proof. The overall idea is that we can always smooth the implied ties in the predictive distributions by introducing a gap at each tied point and adding a uniformly random value. Specifically, for each $y \in Y, s \in \mathbb{R}$ we define the set $V_{\phi,f}(y) \subset \mathbb{R}$ and number $w_{\phi,f}(s, y) \in [0, 1]$:

$$V_{\phi,f}(y) := \{v : D^\pi_{\phi,f}(v|y) > 0 \text{ or } D^\pi_f(v|y) > 0\}$$

$$w_{\phi,f}(s, y) := \sum_{v \in V_{\phi,f}(y), v < s} D^\pi_{\phi,f}(v|y).$$

For each y there can be at most a countable number of point masses in the implied distributions and the total mass is at most 1, so $w_{\phi,f}(s, y)$ will always be defined and smaller than 1.

We now construct the new model, posterior, and test quantity by adding a new parameter p uniformly distributed over the $[0, 1]$ interval, so $\Theta' := \Theta \times [0, 1]$ and for all $y \in Y, p \in [0, 1]$ we set:

$$\theta' := (\theta, p)$$
$$\pi'_{\text{joint}}(y, \theta') := \pi_{\text{joint}}(y, \theta)$$
$$\phi'(\theta'|y) := \phi(\theta|y)$$
$$f'(\theta', y) := \begin{cases} f(\theta, y) + w_{\phi,f}(f(\theta, y), y) + p D^\pi_{\phi,f}(f(\theta, y)|y) & \text{if } f(\theta, y) \in V_{\phi,f}(y) \\ f(\theta, y) + w_{\phi,f}(f(\theta, y), y) & \text{otherwise.} \end{cases}$$

Here, w provides gaps as it increases by $D^\pi_{\phi,f}(f(\theta, y))$ for each tied value. Those gaps are then filled uniformly randomly by adding scaled p. By inspection, there are no ties in f'. The ordering of previously non-tied elements for both continuous and sample ranks has not changed and the order among those previously tied is uniformly random so both sample Q and continuous q the rank distributions do not change. Since $\forall y \in Y, \theta \in \Theta, p \in [0, 1] : \pi'_{\text{post}}(\theta'|y) = \pi_{\text{post}}(\theta|y)$, the construction also directly ensures that $\phi = \pi_{\text{post}}$ if and only if $\phi' = \pi'_{\text{post}}$. $\qquad\square$

Lemma 2 (No ties and inverting CDFs)**.** For all models π, posterior families ϕ, and test quantities $f : \Theta \times Y$:

1. If model π has no ties w.r.t. f then for all y the quantile function $C_f^{-1}(s|y)$ is the inverse of $C_f(s|y)$. Additionally, $C_f(s|y)$ is a surjection on the $[0, 1]$ interval.

21

2. If ϕ has no ties w.r.t. f then for all y the quantile function $C_{\phi,f}^{-1}(s|y)$ is the inverse of $C_{\phi,f}(s|y)$. Additionally, $C_{\phi,f}(s|y)$ is a surjection on the $[0,1]$ interval.

3. If both π and ϕ have no ties w.r.t. f then $q_{\phi,f}(x|y) = C_f\left(C_{\phi,f}^{-1}(x|y)\Big|y\right)$ and $q_{\phi,f}(x|y)$ is a bijection from $[0,1]$ to $[0,1]$.

Proof. Items 1 and 2 follow directly from the basic properties of CDFs. For item 3, when there are no ties, the definition of continuous rank simplifies to $r_{\phi,f}(x|\tilde{\theta},y) = \mathbb{I}\left(x \geq C_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right)\right)$ and thus:

$$q_{\phi,f}(x|y) = \int d\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y)\mathbb{I}\left(x \geq C_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right)\right) =$$

$$= \int d\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y)\mathbb{I}\left(f\left(\tilde{\theta},y\right) \leq C_{\phi,f}^{-1}(x|y)\right) = C_f\left(C_{\phi,f}^{-1}(x|y)\Big|y\right),$$

where the second equality holds because $C_{\phi,f}^{-1}$ is the inverse of $C_{\phi,f}$ for all y. $\square$

Theorem 2 (Correct posterior and q). For any $y \in Y$, if $\forall \theta \in \Theta : \phi(\theta|y) = \pi_{\text{post}}(\theta|y)$ then for any test quantity f we have $\forall x \in [0,1] : q_{\phi,f}(x|y) = x$.

Proof. When π has ties w.r.t. f, we can use Lemma 1 to construct a model with no ties with the same continuous q. So we only need to prove the case where there are no ties.

Without ties, we have $\forall s \in \bar{\mathbb{R}}, y \in Y : C_{\phi,f}(s|y) = C_f(s|y)$ and $\forall x \in [0,1], y \in Y : C_f\left(C_f^{-1}(x|y)\right) = x$ (Lemma 2) and thus

$$q_{\phi,f}(x|y) = \int_Y dy\, C_f\left(C_{\phi,f}^{-1}(x|y)\Big|y\right) \pi_{\text{marg}}(y) = \int_Y dy\, x \pi_{\text{marg}}(y) =$$

$$= x \int_Y dy\, \pi_{\text{marg}}(y) = x.$$

$\square$

Corollary. The correct posterior passes continuous SBC.

Passing SBC is a necessary condition to have a correct posterior distribution, but it is not a sufficient condition. The following theorem characterizes a sufficient condition for having a correct predictive distribution for a given test quantity.

Theorem 3 (Characterization of SBC failures). Given a posterior family ϕ over Y, Θ and test quantity f, then for all $y \in Y$ and $s \in \mathbb{R} : C_{\phi,f}(s|y) = C_f(s|y)$ if and only if $\forall x \in [0,1] : q_{\phi,f}(x|y) = x$.

Proof. Given a model π and a test quantity f, we build a model π' over the same data space and a univariate parameter space $\Theta' := \mathbb{R}$ by first setting:

$$\pi'_{\text{marg}}(y) := \pi_{\text{marg}}(y).$$

We can then define the true posterior and the posterior family ϕ' via their univariate CDFs, for all $s \in \mathbb{R}, y \in Y$:

$$\int_{-\infty}^{s} d\theta'\, \pi'_{\text{post}}(\theta'|y) := C_f^{\pi}(s|y)$$

$$\int_{-\infty}^{s} d\theta'\, \phi'(\theta'|y) := C_{\phi,f}(s|y).$$

22

This always defines a valid joint distribution $\pi'_{\text{joint}}(\theta', y) = \pi'_{\text{post}}(\theta'|y)\pi'_{\text{marg}}(y)$ and thus a valid model.

We now consider the projection function $p : \mathbb{R} \times Y \to \mathbb{R}$, $p(\theta', y) = \theta'$. By construction of π' and ϕ' we have for all $s \in \mathbb{R}, x \in [0, 1], y \in Y$:

$$C_p^{\pi'}(s|y) = C_f^\pi(s|y)$$
$$C_{\phi',p}^{\pi'}(s|y) = C_{\phi,f}^\pi(s|y).$$

This then implies $\forall x \in [0, 1], y \in Y$:

$$q_{\phi',p}^{\pi'}(x|y) = q_{\phi,f}^\pi(x|y).$$

It therefore suffices to prove that for π',

$$\forall \theta' \in \mathbb{R} : \phi'(\theta', y) = \pi'_{\text{post}}(\theta'|y) \iff \forall x \in [0, 1] : q_{\phi',p}^{\pi'}(x|y) = x.$$

The forward implication follows from Theorem 2 because the correct posterior always produces correct q. According to Lemma 1 it suffices to prove the reverse implication only for posteriors without ties. In this case using Lemma 2 we have for all $x \in [0, 1]$:

$$x = q_{\phi',p}^{\pi'}(x|y) = C_p^{\pi'}(C_{\phi,p}^{\pi',-1}(x|y)|y).$$

So for all $y \in Y$ we see that $C^{\pi',-1}$ is the inverse of $C_p^{\pi'}$ and thus the CDFs have to be equal:

$$\forall s \in \mathbb{R} : C_p^{\pi'}(s|y) = C_{\phi,p}^{\pi'}(s|y).$$

Since there is only a single parameter θ' in π', the equality of the CDFs for the projection function implies the equality of the densities, and

$$\forall \theta' \in \mathbb{R} : \phi'(\theta'|y) = \pi'_{\text{post}}(\theta'|y).$$

This completes the proof. $\qquad\square$

This characterizes SBC failures in the sense that SBC investigates the average over Y of the $q_{\phi,f}$ functions, which can appear correct even if for some subset of Y with non-zero marginal mass the $q_{\phi,f}$ functions do not yield expected results. The consequence is that whenever we have an incorrect predictive distribution of a given function, the failure can be isolated to some $\bar{Y} \subset Y$ and detected by separately considering $\int_{\bar{Y}} \mathrm{d}y \, q_{\phi,f}(x|y)$. While in practical uses of SBC inspecting rank distributions for subset of the simulations based on the realized $y \in Y$ leads to loss of power, we can often get similar results by using a suitable test quantity g that depends on f and y in a way that causes the problems in individual y to accumulate instead of canceling. This is one of the reasons why allowing test quantities to depend on data is useful in general.

We now move to establishing a close correspondence between the sample SBC and continuous SBC.

Theorem 4 (Continuous SBC implies sample SBC)**.** Let ϕ be any posterior family over Y, Θ then for all $M \in \mathbf{N}$ and any test quantity f:

1. For any $y \in Y$, if $\forall x \in [0, 1] : q_{\phi,f}(x|y) = x$ then $\forall i \in \{0, \ldots, M - 1\} : Q_{\phi,f}(i|y) = \frac{i+1}{M+1}$.

2. If ϕ passes continuous SBC w.r.t. f, then ϕ passes M-sample SBC w.r.t. f.

Proof. If either ϕ or π has ties w.r.t. f, we can use Lemma 1 to construct a model and posterior family with no ties but the same continuous q and sample Q; therefore it suffices to prove the statement when there are no ties.

When there are no ties, then given y, $q_{\phi,f}(x|y)$ is a bijection and we can thus also define q^{-1}. Additionally, $\tilde{\theta} \sim \pi_{\text{post}}(\tilde{\theta}|y)$ and $\theta_1, \ldots \theta_M \sim \phi_y$ are all conditionally independent given y.

This allows us to determine the probability of the first r draws being smaller than the prior sample $\tilde{\theta} \sim \pi(\theta)$ and the remaining $M - r$ draws being larger conditional on a specific y:

$$\Pr\left(\left(f(\theta_1, y) < f(\tilde{\theta}, y)\right) \wedge \ldots \wedge \left(f(\theta_r, y) < f(\tilde{\theta}, y)\right)\right.$$
$$\left.\wedge \left(f(\theta_{r+1}, y) \geq f(\tilde{\theta}, y)\right) \wedge \ldots \wedge \left(f(\theta_M, y) \geq f(\tilde{\theta}, y)\right) \middle| y\right) =$$

$$\int_0^1 d\tilde{x} \int_{\mathbf{x} \in [0,1]^M} d\mathbf{x} \prod_{i=1}^r \mathbb{I}\left(C_{\phi,f}^{-1}(x_i|y) < C_f^{-1}(\tilde{x}|y)\right) \prod_{i=r+1}^M \mathbb{I}\left(C_{\phi,f}^{-1}(x_i|y) \geq C_f^{-1}(\tilde{x}|y)\right) =$$

$$\int_0^1 d\tilde{x} \int_{\mathbf{x} \in [0,1]^M} d\mathbf{x} \prod_{i=1}^r \mathbb{I}\left(x_i < C_{\phi,f}(C_f^{-1}(\tilde{x}|y)|y)\right) \prod_{i=r+1}^M \mathbb{I}\left(x_i \geq C_{\phi,f}(C_f^{-1}(\tilde{x}|y)|y)\right) =$$

$$= \int_0^1 d\tilde{x} \int_{0 < x_1, \ldots, x_k < q_{\phi,f}^{-1}(\tilde{x}|y) \leq x_{r+1} \ldots x_L < 1} dx_1 \, 1 =$$

$$= \int_0^1 dx \left(q_{\phi,f}^{-1}(x|y)\right)^r \left(1 - q_{\phi,f}^{-1}(x|y)\right)^{M-r}.$$

Since $\theta_1, \ldots, \theta_L$ are independent given y, we can easily extend to all possible orderings:

$$\forall r \in 0, 1, \ldots, M : \Pr\left(\sum_{i=1}^M \mathbb{I}(f(\theta_i, y) < f(\tilde{\theta}, y)) = r \middle| y\right) =$$

$$\int_0^1 dx \binom{M}{r} \left(q_{\phi,f}^{-1}(x|y)\right)^r \left(1 - q_{\phi,f}^{-1}(x|y)\right)^{M-r}$$

$$= \int_0^1 dz \binom{M}{r} z^r (1-z)^{M-r} \frac{\partial}{\partial z} q_{\phi,f}(z|y)$$

From the precondition in claim 1, $q_{\phi,f}(x|y) = x$ and thus $\frac{\partial}{\partial z} q_{\phi,f}(z|y) = 1$. Then $\int_0^1 dz \, z^r (1-z)^{M-r} = B(r+1, M-r+1)$ is the beta integral. Thus,

$$\forall r \in 0, 1, \ldots, M : \Pr\left(\sum_{i=1}^M \mathbb{I}(f(\theta_i, y) < f(\tilde{\theta}, y)) = r \middle| y\right) =$$

$$\binom{M}{r} B(r+1, M-r+1) = \frac{1}{M+1}.$$

Therefore,

$$Q_{\phi,f}(j|y) = \sum_{r=0}^j \Pr\left(\sum_{i=1}^M \mathbb{I}(f(\theta_i, y) < f(\tilde{\theta}, y)) = r \middle| y\right) = \frac{j+1}{M+1},$$

which proves claim 1. For claim 2 we investigate the unconditional probability,

$$\forall r \in 0, 1, \ldots, M : \Pr\left(\sum_{i=1}^{M} \mathbb{I}\left(f(\theta_i, y) < f(\tilde{\theta}, y)\right) = r\right) =$$

$$\int_Y dy\, \pi_{\mathrm{marg}}(y) \Pr\left(\sum_{i=1}^{M} \mathbb{I}\left(f(\theta_i, y) < f(\tilde{\theta}, y)\right) = r \,\middle|\, y\right) =$$

$$= \int_Y dy\, \pi_{\mathrm{marg}}(y) \int_0^1 dz \binom{M}{r} z^r (1-z)^{M-r} \frac{\partial}{\partial z} q_{\phi,f}(z|y) =$$

$$= \int_0^1 dz \binom{M}{r} z^r (1-z)^{M-r} \frac{\partial}{\partial z} \int_Y dy\, \pi_{\mathrm{marg}}(y)\, q_{\phi,f}(z|y).$$

Since claim 2 assumes ϕ satisfies SBC w.r.t. f, we have $\frac{\partial}{\partial z} \int_Y dy\, \pi_{\mathrm{marg}}(y)\, q_{\phi,f}(z|y) = 1$ and thus we can proceed as in claim 1:

$$\int_Y dy\, Q_{\phi,f}(j|y)\pi_{\mathrm{marg}}(y) = \sum_{r=0}^{j} \Pr\left(\sum_{i=1}^{M} \mathbb{I}\left(f(\theta_i, y) < f(\tilde{\theta}, y)\right) = r\right) =$$

$$= \sum_{r=0}^{j} \frac{1}{M+1} = \frac{j+1}{M+1}.$$

$\square$

Corollary. Combining the result with Theorem 2, the correct posterior will produce uniform Q and pass M-sample SBC for any M and f.

Theorem 5 (Sample SBC implies continuous SBC)**.** Let ϕ be any posterior family over Y and Θ, and let f be a test quantity:

1. For any fixed $y \in Y$ if as the number of sample draws $M \to \infty$ we have $\forall i \in \{0, \ldots, M\}$: $Q_{\phi,f}(i|y) \to \frac{i+1}{M+1}$ then $\forall x \in [0,1] : q_{\phi,f}(x|y) = x$.

2. If as $M \to \infty$ we have $\forall i \in \{0, \ldots, M\}$: $\int_Y dy\, Q_{\phi,f}(i|y)\pi_{\mathrm{marg}}(y) \to \frac{i+1}{M+1}$ then ϕ passes continuous SBC for f.

Proof. Using Lemma 1 it once again suffices to prove the case with no ties. We start by showing that as $M \to \infty$ normalized sample ranks converge almost everywhere to the continuous ranks.

Assuming no ties, $R_{\phi,f}(\lfloor xM \rfloor | \tilde{\theta}, y) = \mathrm{Bin}(\lfloor xM \rfloor | M, C_{\phi,f}(f(\tilde{\theta}, y)))$ where $\mathrm{Bin}(K|M, p)$ is the CDF of a binomial distribution with M trials and success probability p evaluated at K. Using the central limit theorem, we have

$$\lim_{M \to \infty} \left(\mathrm{Bin}\left(\lfloor xM \rfloor \,\middle|\, M, p\right) - \mathrm{Nor}\left(\lfloor xM \rfloor \,\middle|\, Mp, \sqrt{Mp(1-p)}\right)\right) = 0,$$

where $\mathrm{Nor}(y|\mu, \sigma)$ is the CDF of a normal distribution with mean μ and standard deviation σ evaluated at y. We can now inspect the limiting behaviour of the z-score:

$$\lim_{M\to\infty} \frac{\lfloor xM \rfloor - Mp}{\sqrt{Mp(1-p)}} = \lim_{M\to\infty} \frac{\lfloor xM \rfloor - xM}{\sqrt{Mp(1-p)}} + \lim_{M\to\infty} \frac{xM - Mp}{\sqrt{Mp(1-p)}} =$$

$$= 0 + \lim_{M\to\infty} \frac{(x-p)\sqrt{M}}{\sqrt{p(1-p)}} = \begin{cases} +\infty & \text{for } x > p \\ 0 & \text{for } x = p \\ -\infty & \text{for } x < p. \end{cases}$$

Thus,

$$\lim_{M\to\infty} R_{\phi,f}(\lfloor xM \rfloor \,|\, \tilde{\theta}, y) = \begin{cases} 1 & \text{for } x > C_{\phi,f}(f(\tilde{\theta}, y)) \\ \frac{1}{2} & \text{for } x = C_{\phi,f}(f(\tilde{\theta}, y)) \\ 0 & \text{for } x < C_{\phi,f}(f(\tilde{\theta}, y)) \end{cases}$$

$$= \begin{cases} r_{\phi,f}(x|\tilde{\theta}, y) & \text{for } x \neq C_{\phi,f}(f(\tilde{\theta}, y)) \\ \frac{1}{2} & \text{for } x = C_{\phi,f}(f(\tilde{\theta}, y)). \end{cases}$$

We have thus established pointwise convergence of $R_{\phi,f}(\lfloor xM \rfloor \,|\, \tilde{\theta}, y)$ to $r_{\phi,f}(x|\tilde{\theta}, y)$ almost everywhere w.r.t. x, because we assume there are no ties. This means that we can satisfy the conditions of the dominated convergence theorem.

We can now prove both claims in an analogous way. We start with claim 2, which is more complex. For all $x \in [0,1]$:

$$\lim_{M\to\infty} \int_Y \mathrm{d}y \, Q_{\phi,f}(\lfloor xM \rfloor \,|\, y)\pi_{\mathrm{marg}}(y) =$$

$$= \lim_{M\to\infty} \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y)R_{\phi,f}(\lfloor xM \rfloor \,|\, \tilde{\theta}, y)\pi_{\mathrm{marg}}(y) =$$

$$= \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y)r_{\phi,f}(x|\tilde{\theta}|y)\pi_{\mathrm{marg}}(y) =$$

$$= \int_Y \mathrm{d}y \, q_{\phi,f}(x|y)\pi_{\mathrm{marg}}(y). \tag{9}$$

Equation (9) holds regardless of the actual rank distributions. We now use the assumption that as $M \to \infty$ the data-averaged rank distribution becomes uniform, which implies $\forall x \in [0,1]$:

$$\lim_{M\to\infty} \int_Y \mathrm{d}y \, Q_{\phi,f}(\lfloor xM \rfloor \,|\, y)\pi_{\mathrm{marg}}(y) = \lim_{M\to\infty} \frac{\lfloor xM \rfloor + 1}{M+1} = x. \tag{10}$$

Combining (9) and (10), we get that $\int_Y \mathrm{d}y \, q_{\phi,f}(x|y)\pi_{\mathrm{marg}}(y) = x$ and therefore ϕ passes continuous SBC w.r.t. f. The reasoning for claim 1 is analogous, only omitting the integration over Y. $\qquad\square$

We will now show that our SBC variant is in some sense complete. Previous work was correct in noting that with test quantities that are only a function of the parameters some problematic posteriors (e.g., the unaltered prior distribution) can never fail SBC. However, when test quantities are allowed to depend on observed data we can in principle discover any problematic ϕ. For this we will need a technical lemma.

Lemma 3 (Value of q). Let ϕ be any posterior family over Y, Θ and f a test quantity and let $s = C_{\phi,f}^{-1}(x|y)$. Then for all $y \in Y, x \in [0,1]$,

$$q_{\phi,f}(x \mid y) = \begin{cases} C_f(s \mid y) + \frac{D_f(s|y)}{D_{\phi,f}(s|y)}(x - C_{\phi,f}(s \mid y)) & \text{for } D_{\phi,f}(s \mid y) > 0 \\ C_f(s \mid y) & \text{otherwise.} \end{cases}$$

26

Proof. For a given x, we split Θ into three disjoint sets:

$$\Theta_1^x := \{\theta \in \Theta : x > C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right)\}$$
$$\Theta_2^x := \{\theta \in \Theta : C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right) - D_{\phi,f}\left(f\left(\theta, y\right) \mid y\right) < x \le C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right)\}$$
$$\Theta_3^x := \{\theta \in \Theta : x \le C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right) - D_{\phi,f}\left(f\left(\theta, y\right) \mid y\right)\}.$$

Now

$$q_{\phi,f}(x|y) = \sum_{i=1}^{3} \int_{\Theta_i^x} \mathrm{d}\theta\, \pi_{\text{post}}(\theta|y) r_{\phi,f}(x|\theta, y).$$

The definition of $r_{\phi,f}$ implies that $\theta \in \Theta_1^x \implies r_{\phi,f}(x|\theta, y) = 1$ and thus:

$$\int_{\Theta_1^x} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) r_{\phi,f}(x|\tilde{\theta}, y) = \int_{\Theta_1^x} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) =$$
$$= \int_{\Theta} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) \mathbb{I}\left(x > C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right)\right) =$$
$$= 1 - \int_{\Theta} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) \mathbb{I}\left(x \le C_{\phi,f}\left(f\left(\theta, y\right) \mid y\right)\right) =$$
$$= 1 - \int_{\Theta} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) \mathbb{I}\left(C_{\phi,f}^{-1}(x|y) \le f\left(\theta, y\right)\right) =$$
$$= \int_{\Theta} \mathrm{d}\tilde{\theta}\, \pi_{\text{post}}(\tilde{\theta}|y) \mathbb{I}\left(f\left(\theta, y\right) < s\right) = C_f(s|y) - D_f(s|y),$$

using the Galois connection between quantile and CDF functions, so that $x \le C_{\phi,f}(s|y) \iff C_{\phi,f}^{-1}(x|y) \le s$ even when $C_{\phi,f}$ is not invertible.

The definition also directly implies $\theta \in \Theta_3^x \implies r_{\phi,f}(x|\theta, y) = 0$, and thus

$$\int_{\Theta_3^x} \mathrm{d}\theta\, \pi_{\text{post}}(\theta|y) r_{\phi,f}(x|\theta, y) = 0.$$

If $D_{\phi,f}\left(s \mid y\right) = 0$ then Θ_2^x is empty, and have $q_{\phi,f}(x|y) = C_f(s|y)$.

If $D_{\phi,f}\left(s \mid y\right) > 0$, then the rank distribution is uniform between $C_{\phi,f}(s|y) - D_{\phi,f}(s|y)$ and $C_{\phi,f}(s|y)$, and the corresponding CDF thus has a linear segment, specifically:

$$\theta \in \Theta_2^x \implies r_{\phi,f}(x|\theta, y) = \frac{x - (C_{\phi,f}(s|y) - D_{\phi,f}(s|y))}{D_{\phi,f}(s|y)} = \frac{x - C_{\phi,f}(s|y)}{D_{\phi,f}(s|y)} + 1.$$

Additionally, $\theta \in \Theta_2^x \iff f(\theta, y) = s$, so

$$\int_{\Theta_2^x} \mathrm{d}\theta\, \pi_{\text{post}}(\theta|y) r_{\phi,f}(x|\theta, y) = \int_{\Theta_2^x} \mathrm{d}\theta\, \pi_{\text{post}}(\theta|y) \left(\frac{x - C_{\phi,f}(s|y)}{D_{\phi,f}(s|y)} + 1\right) =$$
$$= \left(\frac{x - C_{\phi,f}(s|y)}{D_{\phi,f}(s|y)} + 1\right) \int_{\Theta} \mathrm{d}\theta\, \pi_{\text{post}}(\theta|y) \mathbb{I}\left(f(\theta, y) = s\right) =$$
$$= \left(\frac{x - C_{\phi,f}(s|y)}{D_{\phi,f}(s|y)} + 1\right) D_f(s|y).$$

Combining with the previous results yields,

$$q_{\phi,f}\left(x \mid y\right) = C_f\left(s \mid y\right) + \frac{D_f\left(s \mid y\right)}{D_{\phi,f}\left(s \mid y\right)}\left(x - C_{\phi,f}\left(s \mid y\right)\right) \text{for } D_{\phi,f}\left(s \mid y\right) > 0.$$

$\square$

27

Corollary. Whenever $D_{\phi,f}(s \mid y) > 0$, $q_{\phi,f}(x|y)$ linearly interpolates between the two nearest values $x_1 < x < x_2$ such that $x_1 - C_{\phi,f}(C_{\phi,f}^{-1}(x_1|y)|y) = x_2 - C_{\phi,f}(C_{\phi,f}^{-1}(x_2|y)|y) = 0$.

Theorem 6 (Density ratio). For any posterior family ϕ, take $g(\theta, y) = \frac{\pi_{\text{post}}(\theta|y)}{\phi(\theta|y)}$. Then ϕ passes continuous SBC w.r.t. g if and only if π_{post} and ϕ are equal except for a set of measure 0:

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\theta \, \pi_{\text{joint}}(y, \theta) \mathbb{I}\left[\pi_{\text{post}}(\theta|y) \neq \phi(\theta|y)\right] = 0. \tag{11}$$

Proof. First we establish that $\forall y \in Y, s \in \mathbb{R} : C_{\phi,g}(s|y) \geq C_g(s|y)$:

$$C_{\phi,g}(s \mid y) = \int_\Theta \mathrm{d}\theta \, \phi(\theta|y) \, \mathbb{I}(g(\theta, y) \leq s)$$

$$C_g(s \mid y) = \int_\Theta \mathrm{d}\theta \, \pi_{\text{post}}(\theta \mid y) \mathbb{I}(g(\theta, y) \leq s) = \int \mathrm{d}\theta \, \phi(\theta|y) \, g(\theta, y) \mathbb{I}(g(\theta, y) \leq s).$$

First consider $s \leq 1$. We can rearrange:

$$C_{\phi,g}(s|y) - C_g(s|y) = \int_{\theta \in \Theta : g(\theta,y) \leq s} \mathrm{d}\theta \, \phi(\theta|y) \, (1 - g(\theta, y)).$$

Since $\forall \theta \in \Theta, y \in Y : \phi(\theta|y) \geq 0$, the integral only covers non-negative values, and so the overall difference has to be non-negative.

Now consider $s > 1$. Because $\int_\Theta \mathrm{d}\theta \phi(\theta|y) g(\theta, y) = \int_\Theta \mathrm{d}\theta \, \pi_{\text{post}}(\theta|y) = 1$, we can again rearrange:

$$C_{\phi,g}(s|y) - C_g(s|y) = \int_{\theta \in \Theta : g(\theta,y) \leq s} \mathrm{d}\theta \, \phi(\theta|y) - \int_{\theta \in \Theta : g(\theta,y) \leq s} \mathrm{d}\theta \, \phi(\theta|y) g(\theta, y) =$$

$$= 1 - \int_{\theta \in \Theta : g(\theta,y) > s} \mathrm{d}\theta \, \phi(\theta|y) - 1 + \int_{\theta \in \Theta : g(\theta,y) > s} \mathrm{d}\theta \, \phi(\theta|y) g(\theta, y) =$$

$$= \int_{\theta \in \Theta : g(\theta,y) > s} \mathrm{d}\theta \, \phi(\theta|y) \, (g(\theta, y) - 1).$$

where we once again integrate over non-negative values.

So the difference is non-negative for any s. Also the difference cannot be zero for all s once there is region of nonzero mass where $g(\theta, y) \neq 1$. In other words $\forall s : C_{\phi,g}(s|y) = C_g(s|y)$ if and only if the condition in (11) holds.

Since $C_{\phi,g}(s|y) \geq C_g(s|y)$ we have $C_{\phi,g}^{-1}(s|y) \leq C_g^{-1}(s|y)$. And so due to the monotonicity of CDF and quantile functions, $C_g(C_{\phi,g}^{-1}(x|y)|y) \leq C_g(C_g^{-1}(x|y)|y)$. If there are no ties, we can apply Lemma 2 and have for all $y \in Y, x \in [0, 1]$:

$$q_{\phi,g}(x|y) = C_g(C_{\phi,g}^{-1}(x|y)|y) \leq C_g(C_g^{-1}(x|y)|y) = x.$$

Thus $\int_Y \mathrm{d}y \, q_{\phi,g}(x|y) \pi_{\text{marg}}(y) \leq x$, where equality holds only if Equation (11) is satisfied.

When ties are present, we use Lemma 3 to show that $q_{\phi,g}(x|y) = C_g(C_{\phi,g}^{-1}(x|y)|y)$ whenever $C_{\phi_g}(C_{\phi,g}^{-1}) = x$ and for other x, the value is linearly interpolated. This then means that even with ties $\int_Y \mathrm{d}y \, q_{\phi,g}(x|y) \pi_{\text{marg}}(y) \leq x$ and equality holds only if (11) is satisfied.

$\square$

Corollary. For any posterior family ϕ that would provide different posterior inferences than the true posterior π_{post}, there is a test quantity f such that ϕ does not pass SBC w.r.t. f.

The density ratio is typically not available in practice—and if it were, it would likely make more sense to directly check whether $g(\theta, y) = 1$. The above theorem however provides some intuition why using the likelihood $f_{\text{lik}} : f_{\text{lik}}(\theta, y) = \pi_{\text{obs}}(y|\theta)$ is useful as it is related to the density ratio, but not dependent on particular ϕ and usually easily available in practice.

We now generalize the result that ignoring data in test quantities implies that some problems in the posterior family can never be detected. We use a technical lemma, a continuous analog of Theorem 1.

Lemma 4 (Integral representation of SBC). Given a test quantity f and a posterior family ϕ such that both π and ϕ have no ties w.r.t. f, ϕ satisfies continuous SBC w.r.t. f if and only if for all $x \in [0,1]$:

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I} \left(\int_\Theta \mathrm{d}\theta \, \Big(\mathbb{I}(f(\theta, y) < f(\tilde{\theta}, y))\phi(\theta|y) \Big) \le x \right) \pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}) = x. \qquad (12)$$

Proof. We directly apply the definition of $C_{\phi,f}$ and use the absence of ties to get the invertibility of $C_{\phi,f}$ results from Lemma 2.

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I} \left(\int_\Theta \mathrm{d}\theta \, \Big(\mathbb{I}(f(\theta, y) < f(\tilde{\theta}, y))\phi(\theta|y) \Big) \le x \right) \pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}) =$$

$$= \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I} \left(C_{\phi,f}(f(\tilde{\theta}, y)|y) \le x \right) \pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}) =$$

$$= \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I} \left[f(\tilde{\theta}, y) \le C_{\phi,f}^{-1}(x|y) \right] \pi_{\text{obs}}(y|\tilde{\theta})\pi_{\text{prior}}(\tilde{\theta}) =$$

$$= \int_Y \mathrm{d}y\, C_f \left(C_{\phi,f}^{-1}(x|y)|y \right) \pi_{\text{marg}}(y) = \int_Y \mathrm{d}y\, q_{\phi,f}(x|y)\pi_{\text{marg}}(y).$$

$\square$

Theorem 7 (Incomplete use of data). Assume a model π with observation space Y and parameter space Θ, a space Y', and a measurable function $t : Y \to Y'$. Denote the set $t^{-1}(y') = \{y \in Y : t(y) = y'\}$. Consider the model π' with parameter space Θ and observation space Y' such that for all $\theta \in \Theta, y' \in Y'$:

$$\pi'_{\text{prior}}(\theta) = \pi_{\text{prior}}(\theta)$$

$$\pi'_{\text{obs}}(y'|\theta) = \int_{t^{-1}(y')} \mathrm{d}y\, \pi_{\text{obs}}(y|\theta).$$

Assume a test quantity $f' : Y' \times \Theta \to \mathbb{R}$. If we have a posterior family ϕ' on Y', Θ such that ϕ' passes continuous SBC w.r.t. f' and set test quantity $f : Y \times \Theta \to \mathbb{R}, f(\theta, y) = f'(\theta, t(y))$ and posterior family ϕ on Θ, Y such that $\phi(\theta|y) = \phi'(\theta|t(y))$ then ϕ passes continuous SBC w.r.t. f.

Here, the choice of t lets us choose which aspects of the data are ignored, if $\forall y \in Y : t(y) = 1$, we recover the case where all data are ignored: $\pi'_{\text{post}}(\theta|y) = \pi_{\text{prior}}(\theta)$ and thus $\phi(\theta|y) = \pi_{\text{prior}}(\theta)$ will pass SBC w.r.t. f. If t is a bijection, no information is lost. Other choices of t then let us interpolate between those two extremes, for example ignoring just a subset of the data points or treating some data points as censored.

Proof. We start with the integral representation of SBC for ϕ w.r.t. f from Lemma 4.

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(\int_\Theta \mathrm{d}\theta\, \mathbb{I}(f(\theta,y) < f(\tilde{\theta},y))\phi(\theta|y) \leq x\right) \pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}) =$$

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(\int_\Theta \mathrm{d}\theta\, \mathbb{I}(f'(\theta,t(y)) < f'(\tilde{\theta},t(y)))\phi'(\theta,t(y)) \leq x\right) \pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}) =$$

$$\int_{Y'} \mathrm{d}y' \int_{t^{-1}(\bar{y})} \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(\int_\Theta \mathrm{d}\theta\, \mathbb{I}(f'(\theta,y') < f'(\tilde{\theta},y'))\phi'(\theta,y') \leq x\right) \pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}) =$$

$$\int_{Y'} \mathrm{d}y' \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(\int_\Theta \mathrm{d}\theta\, \mathbb{I}(f'(\theta,y') < f'(\tilde{\theta},y'))\phi'(\theta,y') \leq x\right) \int_{t^{-1}(y')} \mathrm{d}y\, \pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}) =$$

$$\int_{Y'} \mathrm{d}y' \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(\int_\Theta \mathrm{d}\theta\, \mathbb{I}(f'(\theta,y') < f'(\tilde{\theta},y'))\phi'(\theta,y') \leq x\right) \pi'_{\mathrm{obs}}(y|\tilde{\theta})\pi'_{\mathrm{prior}}(\tilde{\theta}).$$

The last step is an integral representation of SBC for ϕ' with respect to f' from Lemma 4. $\qquad\square$

Corollary. If we perform SBC with $f_1, \ldots, f_n$ such that there exist $P_1, P_2 \subset Y$, $\int_{P_k} \mathrm{d}y\, \pi_{\mathrm{marg}}(y) > 0$ where $\forall i, \forall y_1 \in P_1, y_2 \in P_2, \forall \theta \in \Theta : f_i(y_1, \theta) = f_i(y_2, \theta)$ and $\int_{P_1} \mathrm{d}y_1 \int_{P_2} \mathrm{d}y_2 \int_\Theta \mathrm{d}\theta\, \mathbb{I}(\pi_{\mathrm{obs}}(y_1|\theta) \neq \pi_{\mathrm{obs}}(y_2|\theta))$, then there will exist ϕ passing SBC w.r.t. $f_1, \ldots, f_n$ that is distinct from the correct posterior $\pi_{\mathrm{post}}(\theta|y)$.

Proof. If the conditions hold, then one can construct t that merges P_1 and P_2 while preserving the values of f_i. The correct posterior for the implied model π' will imply an incorrect posterior for π that passes SBC. $\qquad\square$

Lemma 5 (Order-preserving transformations). Given test quantities $f, g : Y \times \Theta \to \mathbb{R}$ such that $\forall y \in Y, \theta_1, \theta_2 \in \Theta : g(\theta_1, y) < g(\theta_2, y) \iff f(\theta_1, y) < f(\theta_2, y)$ and $g(\theta_1, y) = g(\theta_2, y) \iff f(\theta_1, y) = f(\theta_2, y)$ and a posterior family ϕ then for all $y \in Y, \tilde{\theta} \in \Theta, x \in [0,1], M \in \mathbf{N}, i \in \{0, \ldots, M\}$:

$$r_{\phi,f}(x|\tilde{\theta}, y) = r_{\phi,g}(x|\tilde{\theta}, y)$$
$$R_{\phi,f}(i|\tilde{\theta}, y) = R_{\phi,g}(i|\tilde{\theta}, y).$$

Proof. The result follows directly from definitions of r and R as they depend only on ordering of the values. $\qquad\square$

Lemma 6 (Reverse transformation). Given test quantities $f, g : Y \times \Theta \to \mathbb{R}$ such that $\forall y \in Y, \theta \in \Theta : g(\theta, y) = -f(\theta, y)$, then for all $y \in Y, \tilde{\theta} \in \Theta, x \in [0,1], M \in \mathbf{N}, i \in \{0, \ldots, M\}$:

$$r_{\phi,g}(x|\tilde{\theta}, y) = 1 - r_{\phi,f}(1 - x|\tilde{\theta}, y)$$
$$R_{\phi,g}(i|\tilde{\theta}, y) = 1 - R_{\phi,f}(M - i - 1|\tilde{\theta}, y).$$

Proof. First consider the continuous rank CDF r. We first establish relations between the CDFs and tie probabilities:

$$C_{\phi,g}(s|y) = \int_\Theta \mathrm{d}\theta\, \mathbb{I}(g(\theta, y) \leq s)\, \phi(\theta|y) = \int_\Theta \mathrm{d}\theta\, \mathbb{I}(-f(\theta, y) \leq s)\, \phi(\theta|y) =$$

$$= \int_\Theta \mathrm{d}\theta\, \mathbb{I}(f(\theta, y) \geq -s)\, \phi(\theta|y) = 1 - C_{\phi,f}(-s|y) + D_{\phi,f}(-s|y)$$

$$D_{\phi,g}(s|y) = D_{\phi,f}(-s|y).$$

30

We can then proceed to directly evaluate $r_{\phi,g}$:

$$
\begin{aligned}
r_{\phi,g}(x|\tilde{\theta},y) &= \Pr\left(C_{\phi,g}\left(g\left(\tilde{\theta},y\right)\Big|y\right) - uD_{\phi,g}\left(g\left(\tilde{\theta},y\right)\Big|y\right) \le x\right) = \\
&\Pr\left(C_{\phi,g}\left(-f\left(\tilde{\theta},y\right)\Big|y\right) - uD_{\phi,g}\left(-f\left(\tilde{\theta},y\right)\Big|y\right) \le x\right) = \\
&\Pr\left(1 - C_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) + (1-u)D_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) \le x\right) = \\
&\Pr\left(C_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) - (1-u)D_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) \ge 1 - x\right) = \\
&1 - \Pr\left(C_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) - uD_{\phi,f}\left(f\left(\tilde{\theta},y\right)\Big|y\right) \le 1 - x\right) = \\
&1 - r_{\phi,f}(1 - x|\tilde{\theta},y),
\end{aligned}
$$

where u is uniformly distributed on $[0,1]$ and the second-to-last identity holds, because the probability of exact equality is 0 and $1 - u$ has the same distribution as u.

Now we can focus on the sample rank CDF. We first distinguish how the behaviour of the N_{equals} and N_{less} random variables (from Definition 4) relates between f and g:

$$
N_{\text{equals}}^g = N_{\text{equals}}^f
$$
$$
N_{\text{less}}^g = M - N_{\text{less}}^f - N_{\text{equals}}^f.
$$

We can then directly evaluate $R_{\phi,g}$:

$$
\begin{aligned}
R_{\phi,g}(i|\tilde{\theta},y) &= \Pr\left(N_{\text{less}}^g + K \le i\right) = \Pr\left(M - N_{\text{less}}^f - N_{\text{equals}}^f + K \le i\right) = \\
&\Pr\left(N_{\text{less}}^f + N_{\text{equals}}^f - K \ge M - i\right) = \\
&1 - \Pr\left(N_{\text{less}}^f + K < M - i\right) = \\
&1 - \Pr\left(N_{\text{less}}^f + K \le M - i - 1\right) = 1 - R_{\phi,f}(M - i - 1|\tilde{\theta},y).
\end{aligned}
$$

$\square$

Theorem 8 (Monotonic transformations). Assume test quantities f, g and a set of measurable functions $h_y : \mathbb{R} \to \mathbb{R}$ such that $\forall y \in Y, \theta \in \Theta : f(\theta,y) = h_y(g(\theta_1,y))$ and a posterior family ϕ. If either for all $y \in Y : h_y$ is strictly increasing or for all $y \in Y : h_y$ is strictly decreasing then 1) ϕ passes continuous SBC w.r.t. f if and only if ϕ passes continuous SBC w.r.t. g and 2) ϕ passes M-sample SBC w.r.t. f if and only if ϕ passes M-sample SBC w.r.t. g.

Proof. This is a direct consequence of Lemmas 5 and 6. The test quantity g has to be either an order-preserving transform or a sequence of an order-preserving transform and a reverse transform. Since order-preserving transforms do not change the rank CDFs at all, SBC results will be exactly identical. The only remaining thing to prove is that a reverse transform maintains uniform rank distribution in both the sample and continuous case.

Assuming $\forall y \in Y : h_y$ is strictly decreasing, we get for the sample case:

$$
\begin{aligned}
Q_{\phi,g}(i|y) &= \int_\Theta \mathrm{d}\tilde{\theta}\,\pi_{\text{post}}(\tilde{\theta}|y)R_{\phi,g}(i|\tilde{\theta},y) = \\
\int_\Theta \mathrm{d}\tilde{\theta}\,\pi_{\text{post}}(\tilde{\theta}|y) &- \int_\Theta \mathrm{d}\tilde{\theta}\,\pi_{\text{post}}(\tilde{\theta}|y)R_{\phi,f}(M - i - 1|\tilde{\theta},y) = \\
&1 - Q_{\phi,f}(M - i - 1|y),
\end{aligned}
$$

31

which then implies,

$$\int_Y \mathrm{d}y \, Q_{\phi,g}(i|y)\pi_{\mathrm{marg}}(y) = \int_Y \mathrm{d}y \, \pi_{\mathrm{marg}}(y) - \int_Y \mathrm{d}y \, Q_{\phi,f}(M - i - 1|y)\pi_{\mathrm{marg}}(y) =$$
$$1 - \int_Y \mathrm{d}y \, Q_{\phi,f}(M - i - 1|y).$$

So if ϕ passes SBC w.r.t. f, then,

$$\int_Y \mathrm{d}y \, Q_{\phi,g}(i|y)\pi_{\mathrm{marg}}(y) = 1 - \frac{M - i}{M + 1} = \frac{i + 1}{M + 1},$$

and thus ϕ passes SBC w.r.t. g. If instead we first assume ϕ passes SBC w.r.t. g, then,

$$\int_Y \mathrm{d}y \, Q_{\phi,f}(M - i - 1|y) = 1 - \int_Y \mathrm{d}y \, Q_{\phi,g}(i|y)\pi_{\mathrm{marg}}(y) =$$
$$= 1 - \frac{i + 1}{M + 1} = \frac{(M - i - 1) + 1}{M + 1}.$$

After substituting $j = M - i - 1$ this shows ϕ passes SBC w.r.t. f.

For the continuous case,

$$q_{\phi,g}(x|y) = \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y)r_{\phi,g}(x|\tilde{\theta}, y) =$$
$$= \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y) - \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\tilde{\theta}|y)r_{\phi,f}(1 - x|\tilde{\theta}, y) = 1 - q_{\phi,f}(1 - x|y),$$

which then implies,

$$\int_Y \mathrm{d}y \, q_{\phi,g}(x|y)\pi_{\mathrm{marg}}(y) = \int_Y \mathrm{d}y \, \pi_{\mathrm{marg}}(y) - \int_Y \mathrm{d}y \, q_{\phi,f}(1 - x|y)\pi_{\mathrm{marg}}(y) =$$
$$1 - \int_Y \mathrm{d}y \, q_{\phi,f}(1 - x|y)\pi_{\mathrm{marg}}(y),$$

where both directions of implication for passing SBC w.r.t. f and g follow immediately. □

Theorem 8 shows that certain simple transformations of test quantities lead to equivalent behaviour for SBC. The result cannot be easily strengthened as many transformations in fact can lead to different behaviours for SBC. Appendix B provides several counterexamples: Example 3 shows that if we allow h_y to be strictly increasing for some $y \subset Y$ while other y will have h_y strictly decreasing, we obtain a different check. Example 4 shows that we can obtain different behaviour if we combine a test quantity with a discontinuous non-monotonic bijection, and Example 5 shows the same when a whole range of values is projected onto a single point. In all those examples, the transformed test quantities rule out some posterior families that pass SBC for the original, but there also are posterior families not passing SBC for the original that pass SBC for the transformed quantity.

In some discussions, SBC is linked to the concept of data-averaged posterior, which equals the prior:

$$\forall \theta \in \Theta : \pi_{\mathrm{prior}}(\theta) = \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta} \, \pi_{\mathrm{post}}(\theta|y)\pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}). \tag{13}$$

As with SBC, one can use this condition to compare the prior and the data-averaged posterior for some test quantity derived from the parameters and (optionally) data.

Definition 7 (Data-averaged posterior)**.** A posterior family ϕ has the correct data-averaged posterior for a test quantity f if for all $s \in \mathbb{R}$

$$\int_Y \mathrm{d}y \int_\Theta \mathrm{d}\theta\, \mathbb{I}\left(f(\theta, y) < s\right) \pi_{\mathrm{obs}}(y|\theta)\pi_{\mathrm{prior}}(\theta) =$$
$$= \int_\Theta \mathrm{d}\theta \int_Y \mathrm{d}y \int_\Theta \mathrm{d}\tilde{\theta}\, \mathbb{I}\left(f(\theta, y) < s\right) \phi(\theta|y)\pi_{\mathrm{obs}}(y|\tilde{\theta})\pi_{\mathrm{prior}}(\tilde{\theta}).$$

Even restricting to test quantities that do not depend on data, there will be posterior families with correct data-averaged posterior for a test quantity but not passing SBC as well as posterior families with incorrect data-averaged posterior that pass SBC. Example 2 in Appendix B shows both cases.

The space of posterior families with correct data-averaged posteriors is closed under convex combinations: if ϕ_1 and ϕ_2 have the correct data-averaged posterior for f, then for any $0 \leq a \leq 1, \bar{\phi} : \bar{\phi}(\theta|y) = a\phi_1(\theta|y) + (1-a)\phi_2(\theta|y)$ will also have correct data averaged posterior for f. However a convex combination of two posterior families passing SBC w.r.t. f will not in general pass SBC; again, see Example 2 for a counterexample.

One intuitive way to understand the difference between SBC and data-averaged posterior is that following Lemma 3 and examples in Appendix B we see that passing SBC is naturally understood as a constraint on the quantile function of the test quantity while a correct data-averaged posterior is naturally expressed as a constraint on the density of the test quantity.

This difference is relevant to broader discussions of SBC and related diagnostics, as other works sometimes characterize SBC as comparing prior to data-averaged posterior (e.g., Lee, Nicholls, and Ryder 2019, Appendix M.2 of Lueckmann et al. 2021, Equation 17 of Schad et al. 2022, Grinsztajn et al. 2021, Ramesh et al. 2022).

There is however a sense in which SBC is a stronger condition than data-averaged posterior—as proved in Lemma 2.19 of Cockayne et al. (2020), a posterior family that passes SBC for *all* test quantities not depending on data will have a correct data-averaged posterior for all test quantities not depending on data.

Appendix B: Formal analysis of SBC for a simple model

Here we take a simple model and characterize the posterior families that pass SBC against several test quantities. This is intended to help build intuition about what SBC is actually doing. Additionally some of the solutions serve as counterexamples for some discussions in Appendix A. We will work with the model,

$$\begin{aligned}
\Theta &:= \mathbb{R} \\
Y &:= \{0, 1\} \\
\theta &\sim \mathrm{uniform}(0, 1) \\
y &\sim \mathrm{Bernoulli}(\theta).
\end{aligned} \tag{14}$$

We know the correct posterior and marginal distributions are

$$\begin{aligned}
\pi_{\mathrm{marg}}(0) = \pi_{\mathrm{marg}}(1) &= \frac{1}{2} \\
\pi_{\mathrm{post}}(\theta|0) = \mathrm{beta}(\theta|1, 2) &= 2(1 - \theta) \\
\pi_{\mathrm{post}}(\theta|1) = \mathrm{beta}(\theta|2, 1) &= 2\theta.
\end{aligned} \tag{15}$$

We denote $\Phi(p|y) := \int_0^p \mathrm{d}\theta \phi(\theta|y)$, the CDF of the posterior given by ϕ, and $\Phi^{-1}(x|y)$ is the associated inverse CDF.

Since Y has just two elements, passing SBC for this model translates to:

$$\forall x \in [0,1] : \frac{1}{2}(q_{\phi,f}(x|0) + q_{\phi,f}(x|1)) = x. \tag{16}$$

Further, since for continuous models there are typically no ties we have $q_{\phi,f}(x|y) = C_f(C_{\phi,f}^{-1}(x|y)|y)$ (Lemma 2). All the following examples thus have the same structure:

1. Determine C_f;

2. Express $C_{\phi,f}^{-1}$ in terms of Φ^{-1};

3. Express $q_{\phi,f}$ in terms of Φ^{-1};

4. Solve (16) for Φ^{-1}.

We focus on Φ^{-1} instead of ϕ to make the computation easier as Φ^{-1} occurs naturally in expressions for $C_{\phi,f}^{-1}$. It however has the side effect that the inverse CDF can be immediately used to sample from the posterior family and thus compare the theoretical results to simulations of SBC.

All the examples are accompanied by simulations that show how specific incorrect posteriors manifest with regards to a set of test quantities the simulation results can be found at `https://martinmodrak.github.io/sbc_test_quantities_paper`, and the underlying code at `https://github.com/martinmodrak/sbc_test_quantities_paper`.

Example 1 (Projection). First consider the projection $f_1(\theta, y) := \theta$, so that,

$$C_{f_1}(s|0) = \int_0^s \mathrm{d}\theta \, 2(1-\theta) = 2s - s^2$$

$$C_{f_1}(s|1) = \int_0^s \mathrm{d}\theta \, 2\theta = s^2$$

$$C_{\phi,f_1}^{-1}(x|y) = \Phi^{-1}(x|y). \tag{17}$$

Plugging this into (16) yields, for all $x \in [0,1]$,

$$\Phi^{-1}(x|0) - \frac{(\Phi^{-1}(x|0))^2}{2} + \frac{(\Phi^{-1}(x|1))^2}{2} = x. \tag{18}$$

This means we are relatively free to choose one of the inverse CDFs and solve for the other, for all $x \in [0,1]$:

$$\Phi^{-1}(x|1) = \sqrt{2x + (\Phi^{-1}(x|0) - 1)^2 - 1}. \tag{19}$$

Given a valid inverse CDF $\Phi^{-1}(x|1)$ on $[0,1]$, Equation (19) defines a valid ϕ as long as:

$$\forall x, 0 \le x \le \frac{1}{2} : \Phi^{-1}(x|0) \le 1 - \sqrt{1 - 2x} \tag{20}$$

$$\forall x, \frac{1}{2} \le x \le 1 : \Phi^{-1}(x|0) \ge 1 - \sqrt{2 - 2x} \tag{21}$$

$$\frac{\partial \Phi^{-1}(x|0)}{\partial x}(\Phi^{-1}(x|0) - 1) \ge -1, \tag{22}$$

where (20) and (21) ensure a non-negative value inside the square root in (19), and (22) ensures $\Phi^{-1}(x|1)$ is nondecreasing.

While this allows for a wide array of incorrect solutions, passing SBC for this condition already avoids some pathological behavior, e.g. the low quantiles of the posterior when 0 is observed cannot be too high. If we instead solve (18) for $\Phi^{-1}(x|1)$ we obtain that when 1 is observed, high quantiles cannot be to low; specifically, $\forall \frac{1}{2} < x < 1 : \Phi^{-1}(x|1) > \sqrt{2x-1}$).

Example 2 (Projection and data-averaged posterior). We can further build upon Example 1 to illustrate the differences between passing SBC and having correct data-averaged posterior. Correctness of the data-averaged posterior for the projection $f_1(\theta, y) = \theta$ simply entails for all $\theta \in [0,1]$:

$$\phi(\theta|0) + \phi(\theta|1) = 2. \tag{23}$$

For example, flipping the correct posterior to

$$\phi_A(\theta, 0) := \pi_{\text{post}}(\theta|1) = 2\theta$$
$$\phi_A(\theta, 1) := \pi_{\text{post}}(\theta|0) = 2 - 2\theta$$

satisfies the data-averaged posterior criterion. This however does not pass SBC w.r.t. f_1 as the implied inverse CDF $\Phi_A^{-1}(x|0) = \sqrt{x}$ violates all the conditions in (20)–(22).

On the other hand, we can take a simple posterior family satisfying SBC for f_1 (via Equation 19):

$$\Phi_B^{-1}(x|0) := \begin{cases} \frac{2}{3}x & x < \frac{3}{4} \\ \frac{1}{2} + 2(x - \frac{3}{4}) & x \geq \frac{3}{4} \end{cases}$$

$$\Phi_B^{-1}(x|1) := \begin{cases} \frac{1}{3}\sqrt{6x + 4x^2} & x < \frac{3}{4} \\ \sqrt{3 - 6x + 4x^2} & x \geq \frac{3}{4}. \end{cases}$$

The implied density has an incorrect data-averaged posterior:

$$\phi_B(\theta|0) = \begin{cases} \frac{3}{2} & \theta \leq \frac{1}{2} \\ \frac{1}{2} & \theta > \frac{1}{2} \end{cases}$$

$$\phi_B(\theta|1) = \begin{cases} \frac{3s}{\sqrt{1+4s^2}} & \theta \leq \frac{\sqrt{3}}{2} \\ \frac{s}{\sqrt{4s^2-3}} & \theta > \frac{\sqrt{3}}{2}. \end{cases}$$

Finally, Lee, Nicholls, and Ryder (2019) in Equation 1.3 claim that several methods including SBC cannot notice when ϕ is a convex combination of the prior and the correct posterior. This is not true; the authors derive it from the incorrect assumption that SBC checks for the correct data-averaged posterior. We can readily build a counterexample: consider $\phi_C(\theta|y) := \frac{1}{2}\left(\pi_{\text{post}}(\theta|y) + \pi_{\text{prior}}(\theta)\right)$. We then have:

$$\phi_C(\theta, 0) = \frac{3}{2} - \theta$$
$$\phi_C(\theta, 1) = \frac{1}{2} + \theta$$
$$\Phi_C(s|0) = \frac{1}{2}(3s - s^2)$$
$$\Phi_C(s|1) = \frac{1}{2}(s + s^2)$$
$$\Phi_C^{-1}(x|0) = \frac{3}{2} - \frac{1}{2}\sqrt{9 - 8x}$$
$$\Phi_C^{-1}(x|1) = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + 8x}. \tag{24}$$

However, plugging $\Phi_C^{-1}(x|0)$ into (19) which is a necessary condition for passing SBC w.r.t f_1 yields,

$$\Phi_{C'}^{-1}(x|1) = \frac{\sqrt{3 - \sqrt{9 - 8x}}}{\sqrt{2}},$$

which is in conflict with what we derived in (24) and ϕ_C thus cannot pass SBC w.r.t. f_1 (although it has correct data-averaged posterior).

Example 3 (Likelihood)**.** Here we evaluate SBC for the likelihood,

$$f_2(\theta, y) := \text{Bernoulli}(y|\theta) = \begin{cases} 1 - \theta & y = 0 \\ \theta & y = 1. \end{cases}$$

In this case,

$$C_{f_1}(s|0) = \int_0^1 d\theta \, \mathbb{I}(1 - \theta < s) 2(1 - \theta) = 2 \int_{1-s}^1 d\theta \, (1 - \theta) = 1 - 2s + s^2$$

$$C_{f_2}(s|1) = C_{f_1}(s|1) = \int_0^s d\theta \, 2\theta = s^2$$

$$C_{\phi,f_2}(s|0) = \int_0^1 d\theta \, \mathbb{I}(1 - \theta < s)\phi(\theta|0) = \int_{1-s}^1 d\theta \, \phi(\theta|0) = 1 - \Phi(\theta|0)$$

$$C_{\phi,f_2}^{-1}(x|0) = \Phi^{-1}(1 - x|0)$$

$$C_{\phi,f_2}^{-1}(x|1) = C_{\phi,f_1}^{-1}(x|1) = \Phi^{-1}(x|1).$$

Plugging this into (16) gives, $\forall x \in [0, 1]$,

$$\frac{1}{2} - \Phi^{-1}(1 - x|0) + \frac{(\Phi^{-1}(1 - x|0))^2}{2} + \frac{(\Phi^{-1}(x|1))^2}{2} = x, \tag{25}$$

which we can solve for $\Phi^{-1}(1 - x|0)$:

$$\Phi^{-1}(1 - x|0) = 1 - \sqrt{2x - (\Phi^{-1}(x|1)^2)}. \tag{26}$$

This defines a valid ϕ under the same conditions as SBC for the projection f_1; see (20)–(22).

To combine this with the solution in 19, we additionally need to ensure continuity of $\Phi^{-1}(\frac{1}{2}|0)$, so that computing the number via 19 followed by 26 matches the original value. This translates to

$$\Phi^{-1}\left(\frac{1}{2}\bigg|0\right) = 1 - \frac{\sqrt{2}}{2}. \tag{27}$$

Combining the projection and likelihood therefore fixes the midpoint of the quantile function to its true value for both observations. However, we are still able to choose $\Phi^{-1}(0, x)$ for $x \in [0, \frac{1}{2}]$ freely as long as the conditions 20 - 22 and 27 are met, and this then defines a ϕ that passes SBC w.r.t. both f_1 and f_2.

Example 4 (Non-monotonous bijection)**.** Consider the test quantity

$$f_3(y, \theta) := \begin{cases} \theta & \text{for } \theta < \frac{1}{2} \\ \theta - 1, & \text{otherwise.} \end{cases}$$

36

It might seem intuitively plausible that since f_3 captures the same information as f_1 that passing SBC w.r.t. f_1 will entail passing SBC w.r.t. f_3. This intuition proves to be incorrect.

Assuming $-\frac{1}{2} \leq s \leq \frac{1}{2}$, then the true CDF is,

$$C_{f_3}(s|0) =$$

$$= \int_0^1 d\theta\, \mathbb{I}\left[f_3(\theta) \leq s\right](2 - 2\theta) = \begin{cases} \int_{\frac{1}{2}}^{s+1} d\theta\,(2 - 2\theta) & -\frac{1}{2} < s < 0 \\ \int_{\frac{1}{2}}^1 d\theta(2 - 2\theta) + \int_0^s d\theta(2 - 2\theta) & 0 \leq s < \frac{1}{2} \end{cases} =$$

$$= \begin{cases} \frac{1}{4} - s^2 & -\frac{1}{2} < s < 0 \\ \frac{1}{4} + 2s - s^2 & 0 \leq s < \frac{1}{2} \end{cases}$$

$$C_{f_3}(s|1) = \begin{cases} \int_{\frac{1}{2}}^{s+1} d\theta\, 2\theta & -\frac{1}{2} < s < 0 \\ \int_{\frac{1}{2}}^1 d\theta\, 2\theta + \int_0^s d\theta\, 2\theta & 0 \leq s < \frac{1}{2} \end{cases} = \begin{cases} \frac{3}{4} + 2s + s^2 & -\frac{1}{2} < s < 0 \\ \frac{3}{4} + s^2 & 0 \leq s < \frac{1}{2}. \end{cases}$$

Let us further set $h_y := \Phi\left(\frac{1}{2}|y\right)$; then we can then express the fitted CDF as:

$$C_{\phi,f_3}(s|y) = \begin{cases} \int_{\frac{1}{2}}^{s+1} d\theta\, \phi(\theta, y) & -\frac{1}{2} < s < 0 \\ \int_{\frac{1}{2}}^1 d\theta\, \phi(\theta, y) + \int_0^s d\theta\, \phi(\theta, y) & 0 \leq s < \frac{1}{2} \end{cases} =$$

$$= \begin{cases} \Phi(s + 1|y) - h_y & \text{for } -\frac{1}{2} < s < 0 \\ 1 - h_y + \Phi(s|y) & 0 \leq s < \frac{1}{2} \end{cases}$$

and invert it to obtain:

$$C_{\phi,f_3}^{-1}(x|y) = \begin{cases} \Phi^{-1}(x + h_y|y) - 1 & \text{for } x < 1 - h_y \\ \Phi^{-1}(x - 1 + h_y|y) & \text{otherwise.} \end{cases}$$

We can now evaluate q_{ϕ,f_3}:

$$q_{\phi,f_3}(x|0) = \begin{cases} \frac{1}{4} - \left(\Phi^{-1}(x + h_0|0) - 1\right)^2 & \text{for } x < 1 - h_0 \\ \frac{1}{4} + 2\Phi^{-1}(x - 1 + h_0|0) - \left(\Phi^{-1}(x - 1 + h_0|0)\right)^2 & \text{otherwise} \end{cases}$$

$$q_{\phi,f_3}(x|1) = \begin{cases} -\frac{1}{4} + \left(\Phi^{-1}(x + h_1|1)\right)^2 & \text{for } x < 1 - h_1 \\ \frac{3}{4} + \left(\Phi^{-1}(x - 1 + h_1|1)\right)^2 & \text{otherwise.} \end{cases}$$

So the SBC condition $x = \frac{1}{2}(q_{\phi,f_3}(x|0) + q_{\phi,f_3}(x|1))$ resolves to four cases. First, when $x \leq \min\{1 - h_0, 1 - h_1\}$,

$$2x = -\left(\Phi^{-1}(x + h_0|0) - 1\right)^2 + \left(\Phi^{-1}(x + h_1|1)\right)^2$$

$$\Phi^{-1}(x + h_1|1) = \sqrt{\left(\Phi^{-1}(x + h_0|0) - 1\right)^2 + 2x}.$$

Second, when $1 - h_0 \leq x \leq 1 - h_1$,

$$2x = -\left(\Phi^{-1}(x - 1 + h_0|0) - 1\right)^2 + 1 + \left(\Phi^{-1}(x + h_1|1)\right)^2$$

$$\Phi^{-1}(x + h_1|1) = \sqrt{\left(\Phi^{-1}(x - 1 + h_0|0) - 1\right)^2 + 2x - 1}.$$

Third, when $1 - h_1 \le x \le 1 - h_0$.

$$2x = -\left(\Phi^{-1}(x + h_0|0) - 1\right)^2 + 1 + \left(\Phi^{-1}(x - 1 + h_1|1)\right)^2$$

$$\Phi^{-1}(x - 1 + h_1|1) = \sqrt{\left(\Phi^{-1}(x + h_0|0) - 1\right)^2 + 2x - 1}.$$

Fourth, when $x \ge \max\{1 - h_0, 1 - h_1\}$,

$$2x = 1 - \left(\Phi^{-1}(x - 1 + h_0|0) - 1\right)^2 + 1 + \left(\Phi^{-1}(x - 1 + h_1|1)\right)^2$$

$$\Phi^{-1}(x - 1 + h_1|1) = \sqrt{\left(\Phi^{-1}(x - 1 + h_0|0) - 1\right)^2 + 2x - 2}.$$

Substituting $\bar{h} := h_1 - h_0$ and $y := x + \bar{h} + h_0$ in the first two cases and $y := x - 1 + \bar{h} + h_0$ into the latter two cases, we obtain:

$$\Phi^{-1}(y|1) = \begin{cases} \sqrt{\left(\Phi^{-1}(y - \bar{h}|0) - 1\right)^2 + 2(y - \bar{h} - h_0)} & \text{for } \bar{h} < y \le 1 + \bar{h} \\ \sqrt{\left(\Phi^{-1}(y - 1 - \bar{h}|0) - 1\right)^2 + 2(y - \bar{h} - h_0) - 1} & \text{for } 1 + \bar{h} \le y \\ \sqrt{\left(\Phi^{-1}(y + 1 - \bar{h}|0) - 1\right)^2 + 2(y - \bar{h} - h_0) + 1} & \text{for } y \le \bar{h}. \end{cases} \quad (28)$$

We have one less case, after the substitution the first and fourth cases are identical.

There are a few conditions to make Equation 28 define a valid posterior family. For brevity we do not evaluate all, but we show the relation between h_0 and h_1. We need to ensure $\Phi^{-1}(0|1) = 0$, and so:

$$\begin{cases} 0 = \left(\Phi^{-1}(1 - h_1 + h_0|0) - 1\right)^2 - h_1 + 1 & \text{for } 0 \le \bar{h} \\ 0 = \left(\Phi^{-1}(h_0 - h_1|0) - 1\right)^2 - 2h_1 & \text{for } \bar{h} < 0. \end{cases}$$

This can be further rearranged to:

$$\begin{cases} \Phi^{-1}(1 - h_1 + h_0|0) = 1 - \sqrt{2h_1 - 1} & \text{for } h_0 \le h_1 \\ \Phi^{-1}(h_0 - h_1|0) = 1 - \sqrt{2h_1} & \text{for } h_0 > h_1. \end{cases} \quad (29)$$

So to construct a posterior family satisfying SBC w.r.t. f_3, we can choose almost any $\Phi^{-1}(x|0)$, calculate h_0, then use Equation 29 to solve for h_1 and finally compute $\Phi^{-1}(x|1)$ via (28).

Now we can show that this is indeed a different condition than the one derived in Example 1. Let us take ϕ_B from Example 2—as discussed there, it passes SBC w.r.t. f_1:

$$\Phi_B^{-1}(x|0) = \begin{cases} \frac{2}{3}x & x < \frac{3}{4} \\ \frac{1}{2} + 2(x - \frac{3}{4}) & x \ge \frac{3}{4} \end{cases}$$

$$\Phi_B^{-1}(x|1) = \begin{cases} \frac{1}{3}\sqrt{6x + 4x^2} & x < \frac{3}{4} \\ \sqrt{3 - 6x + 4x^2} & x \ge \frac{3}{4}. \end{cases}$$

From Φ_B, we obtain $h_0 = \frac{3}{4} > \frac{3}{4}(\sqrt{2} - 1) = h_1$, while

$$\Phi_B^{-1}(h_0 - h_1|0) = \Phi_B^{-1}\left(\frac{3}{4}(2 - \sqrt{2})|0\right) = \frac{1}{2}(2 - \sqrt{2}) = 1 - \frac{1}{\sqrt{2}}$$

$$1 - \sqrt{2h_1} = 1 - \sqrt{\frac{3}{2}(\sqrt{2} - 1)} = 1 - \frac{\sqrt{3(\sqrt{2} - 1)}}{\sqrt{2}},$$

so the condition in (29) does not hold and ϕ_B cannot pass SBC w.r.t. f_3.

Example 5 (Ties - continuous)**.** We obtain a yet different check when we deliberately choose to ignore some information. Take:

$$f_4(\theta, y) := \begin{cases} \theta & \theta < \frac{1}{2} \\ \frac{1}{2} & \theta \geq \frac{1}{2}. \end{cases}$$

We now have ties, and so we must evaluate the tie probabilities:

$$C_{f_4}(s|0) = \begin{cases} \int_0^s d\theta \, 2(1-\theta) = 2s - s^2 & s < \frac{1}{2} \\ 1 & s \geq \frac{1}{2} \end{cases}$$

$$D_{f_4}(s|0) = \begin{cases} 0 & s \neq \frac{1}{2} \\ \int_{\frac{1}{2}}^1 d\theta \, 2(1-\theta) = \frac{1}{4} & s = \frac{1}{2} \end{cases}$$

$$C_{f_4}(s|1) = \begin{cases} \int_0^s d\theta \, 2\theta = s^2 & s < \frac{1}{2} \\ 1 & s \geq \frac{1}{2} \end{cases}$$

$$D_{f_4}(s|1) = \begin{cases} 0 & s \neq \frac{1}{2} \\ \int_{\frac{1}{2}}^1 d\theta \, 2\theta = \frac{3}{4} & s = \frac{1}{2}. \end{cases}$$

Reusing the notation $h_y := \Phi\left(\frac{1}{2}|y\right)$ from previous example, we get:

$$C_{\phi, f_4}(s|y) = \begin{cases} \Phi(s|y) & s < \frac{1}{2} \\ 1 & s \geq \frac{1}{2} \end{cases}$$

$$C_{\phi, f_4}^{-1}(x|y) = \begin{cases} \Phi^{-1}(x|y) & x < h_y \\ \frac{1}{2} & x \geq h_y \end{cases}$$

$$D_{\phi, f_4}(s|y) = \begin{cases} \int_{\frac{1}{2}}^1 d\theta \, \phi(\theta|y) = 1 - h_y & s = \frac{1}{2} \\ 0 & s \neq \frac{1}{2}. \end{cases}$$

We can now use Lemma 3:

$$q_{\phi, f_4}(x|y) = \begin{cases} C_{f_4}(\Phi^{-1}(x|y)|y) & x < h_y \\ C_{f_4}\left(\frac{1}{2}|y\right) + \frac{D_{f_4}\left(\frac{1}{2}|y\right)}{D_{\phi, f_4}\left(\frac{1}{2}|y\right)}\left(x - C_{\phi, f_4}\left(\frac{1}{2}|y\right)\right) & x \geq h_y \end{cases} =$$

$$= \begin{cases} C_{f_4}(\Phi^{-1}(x|y)|y) & x < h_y \\ 1 + \frac{D_{f_4}\left(\frac{1}{2}|y\right)}{1 - h_y}(x - 1) & x \geq h_y \end{cases}$$

$$q_{\phi, f_4}(x|0) = \begin{cases} 1 - \left(\Phi^{-1}(x|0) - 1\right)^2 & x < h_0 \\ 1 + \frac{x-1}{4(1-h_0)} & x \geq h_0 \end{cases}$$

$$q_{\phi, f_4}(x|1) = \begin{cases} \left(\Phi^{-1}(x|1)\right)^2 & x < h_1 \\ 1 + \frac{3(x-1)}{4(1-h_1)} & x \geq h_1. \end{cases}$$

And the SBC condition becomes:

$$2x = \begin{cases} -\left(\Phi^{-1}(x|0) - 1\right)^2 + \left(\Phi^{-1}(x|1)\right)^2 + 1 & x < \min\{h_0, h_1\} \\ -\left(\Phi^{-1}(x|0) - 1\right)^2 + 2 + \frac{3(x-1)}{4(1-h_1)} & h_1 \leq x < h_0 \\ 1 + \frac{x-1}{4(1-h_0)} + \left(\Phi^{-1}(x|1)\right)^2 & h_0 \leq x < h_1 \\ 2 + \frac{x-1}{4(1-h_0)} + \frac{3(x-1)}{4(1-h_1)} & \max\{h_0, h_1\} \leq x. \end{cases} \tag{30}$$

If $\max\{h_0, h_1\} = 1$, then SBC fails as values of $f_4(\theta, y) = \frac{1}{2}$ are never generated in the posterior while appearing in the prior. If not, then the last branch implies,

$$h_1 = \frac{5h_0 - 4}{8h_0 - 7}, h_0 < \frac{4}{5}. \tag{31}$$

Substituting this into the second branch in 30, we obtain a set of constraints on Φ^{-1}:

$$x < \min\{h_0, h_1\} \implies \Phi^{-1}(x|1) = \sqrt{2x + (\Phi^{-1}(x|0) - 1)^2 - 1} \tag{32}$$

$$h_1 \le x < h_0 \implies \Phi^{-1}(x|0) = 1 - \frac{1}{2}\sqrt{\frac{x-1}{h_0 - 1}} \tag{33}$$

$$h_0 \le x < h_1 \implies \Phi^{-1}(x|1) = \frac{1}{2}\sqrt{\frac{3 - 4h_0 + x(8h_0 - 7)}{h_0 - 1}} \tag{34}$$

Take $\Phi_D^{-1}(x|0) := x^2, \Phi_D^{-1}(x|1) := \sqrt{2x - 2x^2 + x^4}$. This will pass SBC w.r.t. f_1, as it is derived directly via (19). However, it will not pass SBC w.r.t. f_4 for multiple reasons. Probably the easiest to see is that in this case we have $h_0 = \sqrt{\frac{1}{2}}$ and $h_1 \approx 0.15$, so over a large range we would need $\Phi_D^{-1}(x|0)$ proportional to square root of x due to (33). Additionally, condition (31) entails $h_0 = \frac{1}{2} \implies h_1 = \frac{1}{2}$ and is therefore also violated.

Passing SBC for f_4 restricts the functional form of a potentially big segment of the inverse CDFs via (33) or (34), which makes it strict in this area while providing no constraints on the distribution of values above $\frac{1}{2}$.

Constructing a posterior family satisfying SBC for f_4 can then proceed as follows: Pick $0 < h_0 < \frac{4}{5}$, calculate h_1 via (31). For $x > h_y$ the values of $\Phi(x|y)$ can be arbitrary (as long as they define valid quantile functions). We can also freely choose $\Phi^{-1}(x|0)$ for $x < \min\{h_0, h_1\}$, but we need to ensure that (a) $\Phi^{-1}(x|1)$ is a valid quantile function—the conditions derived for f_1 in (20)–(22) restricted to $x < \min\{h_0, h_1\}$ are sufficient—and (b) $\Phi^{-1}(h_y|y) = \frac{1}{2}$. The latter condition implies that if $h_0 < h_1$ then $\Phi(h_0|1) < \frac{1}{2}$ which we can combine with (32) to get

$$\sqrt{2h_0 + (\Phi^{-1}(h_0|0) - 1)^2 - 1} < \frac{1}{2},$$

which reduces to $\frac{3}{8} \le h_0 < \frac{1}{2}$. The condition $\Phi(h_1|1) = \frac{1}{2}$ is then already ensured by (34). If on the other hand $h_0 \ge h_1$ then substituting $x = h_1$ into (32) implies

$$\sqrt{2h_1 + (\Phi^{-1}(h_1|0) - 1)^2 - 1} = \frac{1}{2},$$

which reduces to $\Phi^{-1}(h_1|0) = 1 - \frac{1}{2}\sqrt{\frac{9}{7-8h_0}}$ and $\frac{1}{2} \le h_0 < \frac{25}{32}$. The condition $\Phi(h_0|0) = \frac{1}{2}$ is then already ensured by (33).

Example 6 (Ties—discrete)**.** To get some intuition on the behaviour of SBC for discrete parameters, we further simplify the model into,

$$\Theta := \left\{\frac{1}{3}, \frac{2}{3}\right\}$$
$$Y := \{0, 1\}$$
$$\forall \theta \in \Theta : \pi_{\text{prior}}(\theta) := \frac{1}{2}$$
$$y \sim \text{Bernoulli}(\theta). \tag{35}$$

We know the correct posterior and marginal distributions are

$$\pi_{\text{marg}}(0) = \pi_{\text{marg}}(1) = \frac{1}{2}$$

$$\pi_{\text{post}}\left(\frac{1}{3}\middle|0\right) = \pi_{\text{post}}\left(\frac{2}{3}\middle|1\right) = \frac{2}{3}$$

$$\pi_{\text{post}}\left(\frac{1}{3}\middle|1\right) = \pi_{\text{post}}\left(\frac{2}{3}\middle|0\right) = \frac{1}{3}. \tag{36}$$

The posterior family is fully defined by just two numbers: $a = \phi\left(\frac{1}{3}|0\right)$ and $b = \phi\left(\frac{1}{3}|1\right)$. We can directly use the definitions to obtain for f_1 (i.e., just θ):

$$q_{\phi,f_1}(x|0) = \begin{cases} \frac{2x}{3a} & x \leq a \\ \frac{2}{3} + \frac{x-a}{3(1-a)} & \text{otherwise} \end{cases}$$

$$q_{\phi,f_1}(x|1) = \begin{cases} \frac{x}{3b} & x \leq b \\ \frac{1}{3} + \frac{2(x-b)}{3(1-b)} & \text{otherwise}. \end{cases}$$

For the case $x \leq \min\{a,b\}$ the SBC condition becomes:

$$2x = \frac{2x}{3a} + \frac{x}{3b}.$$

this can hold for all x in the range only if

$$2 = \frac{2}{3a} + \frac{1}{3b}$$

$$ab - \frac{1}{3}b - \frac{1}{6}a = 0. \tag{37}$$

Now we can inspect the case $x > \max\{a,b\}$ where the SBC condition becomes.

$$2x = 1 + \frac{x-a}{1-a} + \frac{2(x-b)}{3(1-b)}.$$

Focusing on the coefficient for x in the above equation yields

$$1 - \frac{1}{6(1-a)} - \frac{1}{3(1-b)} = 0$$

$$6ab - 4a - 5b + 3. \tag{38}$$

Combining (37) and (38) already leaves only two solution: either the posterior equals the prior ($\phi\left(\frac{1}{3}|0\right) = \phi\left(\frac{1}{3}|1\right) = \frac{1}{2}$) or $\phi = \pi_{\text{post}}$. Those solutions also satisfy all the other cases.

This example shows that when the underlying parameter space is discrete, there is less flexibility to craft ϕ to pass SBC and the space of posterior families passing SBC can have more structure than when no ties are present.