Assessing Inference Quality for Probabilistic Programs using Multivariate Simulation Based Calibration

Sharan Yalburgi*, Jameson Quinn†, Veronica Weiner†§, Sam Witty‡, Vikash Mansinghka† and Cameron Freer†

*BITS Pilani, India; †MIT, USA; ‡UMass Amherst, USA; §Probabilistic Computing Associates, USA

Introduction

Problem:

- To have a principled approach that can test **arbitrary inference procedures** that are specified as code.

Solution:

- This paper presents **Multivariate Simulation Based Calibration (mSBC)**, a multivariate extension to SBC [1] , which assesses inference quality using the approximate posterior averaged across data simulated from the prior.
- For each iteration, mSBC only requires a bag of samples from the inferred parameter posterior and a sample from parameter prior.
- This paper addresses how to run mSBC on models with a mixture of continuous and discrete parameters; open-universe models where the number of variables is uncertain; and models where existence of one or more parameters in a given trace is uncertain.

Simulation Based Calibration (SBC)

$$\underbrace{\pi(\theta)}_{\text{prior}} = \int \overbrace{\pi(\theta \mid \tilde{y})}^{\text{computed posterior}} \underbrace{\pi(\tilde{y} \mid \tilde{\theta})}_{\text{data simulation}} \overbrace{\pi(\tilde{\theta})}^{\text{parameter sampled from prior}} d\tilde{y}\, d\tilde{\theta}$$

If parameters are sampled from the prior

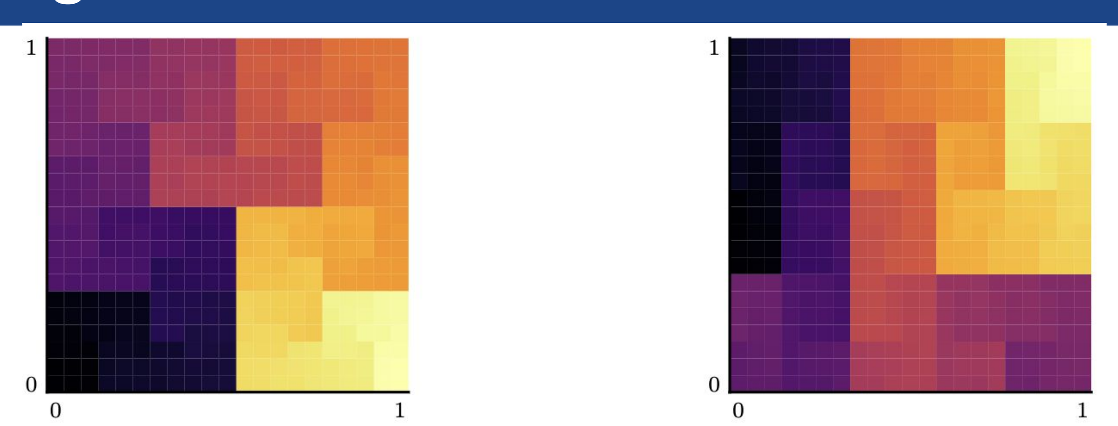
AND

Data is simulated using the sampled parameters

THEN

The inferred parameter posteriors should in expectation (w.r.t. data) match the parameter priors.

Space Filling Curves



(a) Hilbert curve ranking (b) Peano curve ranking

Multivariate Ranks

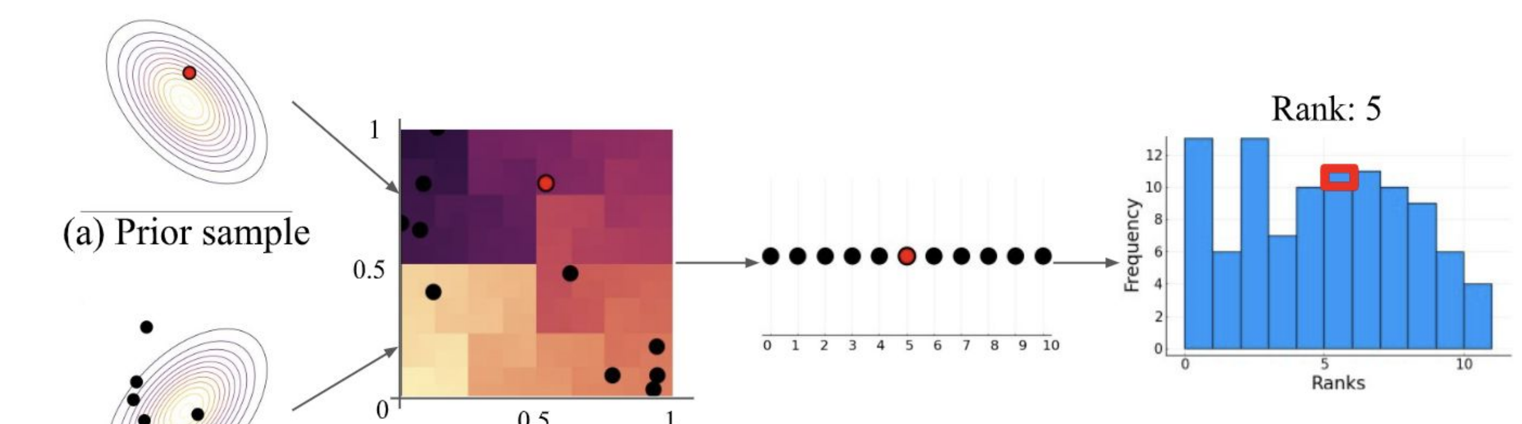
A *multivariate ranking function* is defined to be a function

$r_N : \mathbb{R}^{d\times N} \to S_N$ for some $d, N \geq 1$, which maps *N* points of *d*-dimensional real space to a permutation of [*N*], with the natural symmetry condition.

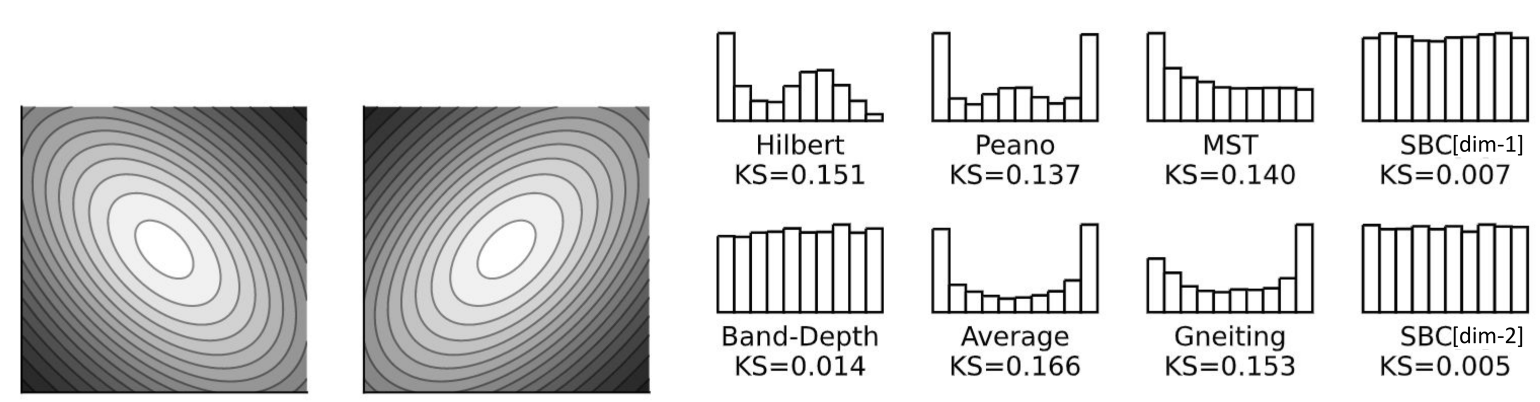
Existing multivariate ranking functions.

Name	Ranking function ($r_N(x_0, \ldots, x_{N-1})$ for $x_i \in \mathbb{R}^d$)	Properties
MST [3]	$\arg\mathrm{sort}_{i\in[N]} \left(\text{length of Euclidean MST on the set } \boldsymbol{x}\backslash x_i\right)$	Center-outward
Band-Depth [3]	$\arg\mathrm{sort}_{i\in[N]} \frac{1}{d} \sum_{1\leq k\leq d, 1\leq j_1<j_2\leq N} \mathbb{1}\left\{\min\{x^k_{j_1}, x^k_{j_2}\} \leq x^k_i \leq \max\{x^k_{j_1}, x^k_{j_2}\}\right\}$	Center-outward
Gneiting [2]	$\arg\mathrm{sort}_{i\in[N]} \sum_{j=1}^{N} \mathbb{1}\{\forall k\ x^k_j \leq x^k_i\}$	Reduces to SBC for $d = 1$.
Average [3]	$\arg\mathrm{sort}_{i\in[N]} \frac{1}{d} \sum_{k=1}^{d} \sum_{j=1}^{N} \mathbb{1}\{x^k_j < x^k_i\}$	Reduces to SBC for $d = 1$.

Multivariate Simulation Based Calibration (mSBC)



(a) Prior sample (b) Inferred posterior samples (c) Impose rank using a multivariate ranking procedure. (d) Imposed rank viewed on ℕ (e) Rank contributes to the mSBC rank histogram



(a) Exact (b) Approximate (c) Univariate (marginal) SBC versus mSBC

Validity of mSBC

Theorem 1 *Suppose* $(R_n)_{n\in}$ *is a ranking procedure. Sample* $\theta \sim \pi(\theta)$ *and* $y \mid \theta \sim \pi(y \mid \theta)$. *Let* $L \geq 1$, *and for each* $\ell = 1, \ldots, L$, *sample* $\theta_\ell \mid y \sim \pi(\theta \mid y)$. *Assume that* $\pi(\theta \mid y)$ *is a continuous measure with probability 1. Define* $\sigma = R_{L+1}(\theta, \theta_1, \ldots, \theta_L)$, *so that* $\sigma(0)$ *is the rank of* θ *with respect to* $\{\theta_1, \ldots, \theta_L\}$. *Then* $\sigma(0)$ *is distributed uniformly on* $[L + 1]$.

Non-uniform rank histogram ⇒ Incorrect inference

Case Studies

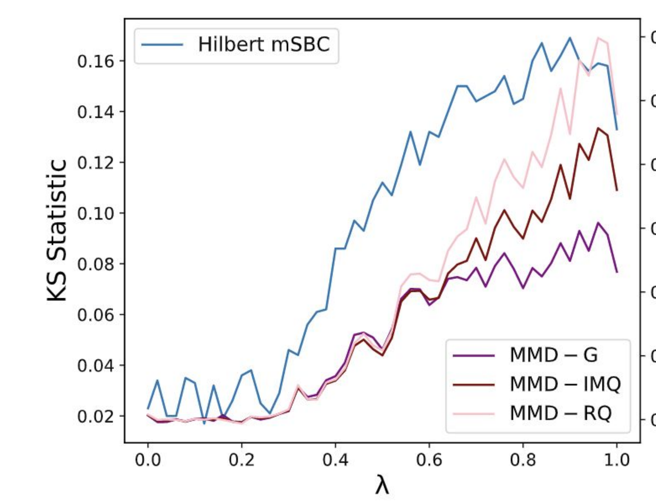
Neal's Funnel

$$y \sim \mathrm{Normal}(0, 3)$$
$$x \mid y \sim \mathrm{Normal}(0, e^{y/2})$$

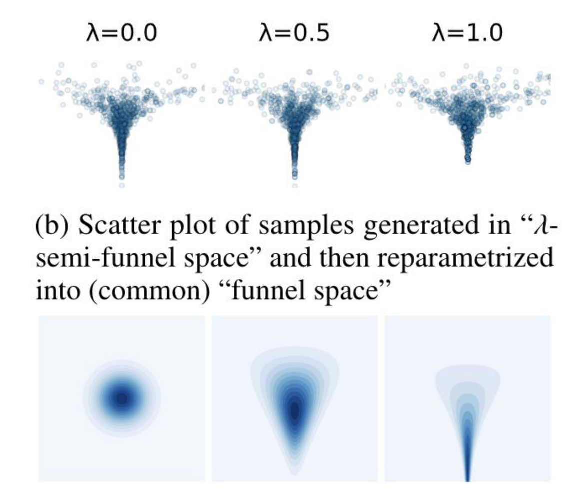
(a) Generative model

$$z_{\mathrm{raw}} \sim \mathrm{Normal}(0, 3^{\lambda})$$
$$z \mid z_{\mathrm{raw}} = 3^{1-\lambda} z_{\mathrm{raw}}$$
$$x_{\mathrm{raw}} \mid z \sim \mathrm{Normal}(0, e^{\frac{z}{2}})$$
$$x \mid z, x_{\mathrm{raw}} = e^{\frac{z*(1-\lambda)}{2}} x_{\mathrm{raw}}$$

(b) Reparametrized generative model



(a) Non-uniformity measures for Hilbert-curve rank histograms, as a function of reparametrization parameter



(b) Scatter plot of samples generated in "λ-semi-funnel space" and then reparametrized into (common) "funnel space"

(c) Marginal densities of "λ semi-funnel space" for λ = 0.0, 0.5, and 1.0.

Linear regression with outliers

```
function block_resimulation_kernel(tr)
    line_params = Gen.select(:noise, :slope, :
        ↪ intercept)
    (tr, _) = mh(tr, line_params)
    (xs,) = get_args(tr)
    n = length(xs)
    for i=1:n
        (tr, _) = mh(tr, Gen.select((:
            ↪ is_outlier, i)))
    end
    (tr, _) = mh(tr, Gen.select(:prob_outlier))
    tr
end
```

Inference	Hilbert	Hilbert KS	MST	MST KS	Band-Depth	Band-Depth KS	Average	Average KS	Gneiting	Gneiting KS
Importance Resampling		0.078		0.041		0.128		0.121		0.059
Block Resimulation		0.029		0.085		0.034		0.052		0.069

(a) mSBC results for regression with outlier detection

Gaussian mixture model with an unknown number of mixture components

```
function imcmc_kernel(tr)
    tr, = mh(tr, gibbs_update_w, ())
    tr, = mh(tr, gibbs_update_ξ, ())
    tr, = mh(tr, gibbs_update_means, ())
    tr, = mh(tr, gibbs_update_vars, ())
    tr, = mh(tr, gibbs_update_allocations,
        ↪  ())
    tr, = split_merge(tr)
    tr
end
```

Inference	Hilbert	Hilbert KS	Peano	Peano KS	MST	MST KS	Band-Depth	Band-Depth KS	Average	Average KS	Gneiting	Gneiting KS
Importance Resampling		0.503		0.502		0.422		0.288		0.473		0.165
Involutive MCMC		0.071		0.123		0.219		0.046		0.13		0.069

(b) mSBC results for a GMM with an unknown number of mixture components

References

1. Talts, Sean, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. "Validating Bayesian inference algorithms with simulation-based calibration." arXiv preprint arXiv:1804.06788 (2018).
2. Gneiting, Tilmann, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson. "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds." Test 17, no. 2 (2008): 211-235.
3. Thorarinsdottir, Thordis L., Michael Scheuerer, and Christopher Heinz. "Assessing the calibration of high-dimensional ensemble forecasts using rank histograms." Journal of computational and graphical statistics 25, no. 1 (2016): 105-122.