Validation Diagnostics for SBI algorithms based on Normalizing Flows

Julia Linhart
Université Paris-Saclay, Inria, CEA
Palaiseau 91120, France
julia.linhart@inria.fr

Alexandre Gramfort
Université Paris-Saclay, Inria, CEA
Palaiseau 91120, France
alexandre.gramfort@inria.fr

Pedro L. C. Rodrigues
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
Grenoble 38000, France
pedro.rodrigues@inria.fr

Abstract

Building on the recent trend of new deep generative models known as Normalizing Flows (NF), simulation-based inference (SBI) algorithms can now efficiently accommodate arbitrary complex and high-dimensional data distributions. The development of appropriate validation methods however has fallen behind. Indeed, most of the existing metrics either require access to the true posterior distribution, or fail to provide *theoretical guarantees* on the consistency of the inferred approximation beyond the one-dimensional setting. This work proposes easy to interpret validation diagnostics for multi-dimensional conditional (posterior) density estimators based on NF. It also offers theoretical guarantees based on results of local consistency. The proposed workflow can be used to check, analyse and guarantee consistent behavior of the estimator. The method is illustrated with a challenging example that involves tightly coupled parameters in the context of computational neuroscience. This work should help the design of better specified models or drive the development of novel SBI-algorithms, hence allowing to build up trust on their ability to address important questions in experimental science.

1 Introduction

Recent advances in computing have led to a new generation of expressive simulators used to study complex systems in many scientific fields [4]. They implicitly encode the intractable likelihood $p(x \mid \boldsymbol{\theta})$ of the underlying mechanistic model which relates observed data $x \in \mathbb{R}^d$ to scientifically meaningful internal parameters $\boldsymbol{\theta} \in \mathbb{R}^m$. To perform statistical inference in this setting, one can recur to simulation based inference (SBI) [4] to approximate the posterior distribution $p(\boldsymbol{\theta} \mid x)$ using samples from the joint pdf $p(x, \boldsymbol{\theta})$. In this work, we consider SBI methods based on normalizing flows [14], which are invertible neural networks that can be trained via maximum likelihood. Once the flow is trained, and for any new observation x, one can directly evaluate the estimated density over the parameter space (i.e. $\boldsymbol{\theta}$-space) [6], draw samples to construct confidence regions [11], etc.

Flow-based SBI has been used in many recent applied works [10, 1, 6], but it lacks an important feature before becoming a technology for experimental science: *validation*. Ideally, one would like to have a method that provides finite-sample guarantees of nominal coverage (or calibration) of the estimated posterior regions, but also ensures that the approximation fits the true underlying posterior of the model when new data is observed. While solutions have been proposed to provide such finite-sample guarantees [5, 11], assessing the convergence and consistency of the underlying

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

inference remains a challenging task [10]. Simulation based calibration (SBC) [18] is arguably the most popular metric for validating posterior approximations in the applied SBI community [3], but it only provides necessary conditions for consistency and fails to give any insight on the *local* behavior of the estimator. Moreover, it is a univariate procedure, thus ignoring any information about the coupling between parameters. Zhao et al. [19] recently proposed local coverage tests (LCT). They leverage machine learning to evaluate their test quantities on different locations of the feature space (i.e. x-space). In 1D, the probability integral transform (PIT) provides *necessary and sufficient* conditions for consistency, which is not the case for the proposed multivariate extensions (e.g. HPD).

In this work, we present a multivariate version of LCT for density models based on normalizing flows. Our method comes with the same theoretical guarantees on local consistency and interpretable diagnostics as provided by [19] in 1D. We also present a workflow that can be used as a practical user guide. Lastly, we provide numerical illustrations on a well known model from computational neuroscience [8], and demonstrate the importance of trustworthy diagnostics for multivariate posterior distributions when correlations between parameter variables can play an important role.

2 Methods

Our conditional density estimator q_ϕ is a normalizing flow defined for samples $\boldsymbol{\theta} \in \mathbb{R}^m$ and is conditioned on observations $x \in \mathbb{R}^d$. It uses a Gaussian base distribution $p(\boldsymbol{z}) = \mathcal{N}(0, \boldsymbol{I}_m)$ and a bijective transformation defined for every x, $T_\phi(.;x) := (T_{\phi,1}(.;x), \dots, T_{\phi,m}(.;x))$, with Jacobian $J_{T_\phi}(.;x)$, such that $q_\phi(\boldsymbol{\theta} \mid x) = p(T_\phi(\boldsymbol{z};x)) = p(\boldsymbol{z}) \left|\det J_{T_\phi}(\boldsymbol{z};x)\right|^{-1}$.

Our goal is to evaluate the *local consistency* [9] of q_ϕ with respect to the true posterior density, i.e. whether for a given x the following null hypothesis holds:

$$\mathcal{H}_0(x) : q_\phi(\boldsymbol{\theta} \mid x) = p(\boldsymbol{\theta} \mid x), \quad \forall \boldsymbol{\theta} \in \mathbb{R}^m \; . \tag{1}$$

We define the multivariate *probability integral transform* (PIT) of $\boldsymbol{\theta}$ at x and associated to q_ϕ as the vector of m one-dimensional projections:

$$\mathrm{PIT}_m(\boldsymbol{\theta}, x, q_\phi) = [P_1(\boldsymbol{\theta}, x), \dots, P_m(\boldsymbol{\theta}, x)], \quad P_i(., x) = F_{\mathcal{N}(0,1)} \circ T^{-1}_{\phi,i}(.;x), \; \forall i \in [1, m] \tag{2}$$

where $F_{\mathcal{N}(0,1)}$ is the cumulative distribution function (c.d.f.) of the univariate normal distribution.

Theorem 1: Local Consistency and multivariate PIT. *For any $x \in \mathbb{R}^d$, the null hypothesis $\mathcal{H}_0(x)$ holds if, and only if, the covariates of $PIT_m(\boldsymbol{\theta}, x, q_\phi)$ conditioned on x are mutually independent and uniformly distributed over* $(0, 1)$. We refer to Appendix A for a detailed proof of this result.

$$\begin{aligned} p(\boldsymbol{\theta} \mid x) = q_\phi(\boldsymbol{\theta} \mid x) &\iff p(T_\phi^{-1}(\boldsymbol{\theta}, x) \mid x) = p(\boldsymbol{z}) \\ &\iff p(T^{-1}_{\phi,1}(\boldsymbol{\theta}, x), \dots T^{-1}_{\phi,m}(\boldsymbol{\theta}, x) \mid x) = \mathcal{N}(0, \boldsymbol{I}_m) \\ &\iff \begin{cases} p(P_i(\boldsymbol{\theta}, x) \mid x) = \mathcal{U}(0,1) \quad \forall i \in [1, m] \quad \text{and} \\ \{P_i(\boldsymbol{\theta}, x) \mid x\}_{i=1,\dots,m} \quad \text{are mutually independent} \end{cases} \end{aligned}$$

We can now verify the null hypothesis of local consistency $\mathcal{H}_0(x)$ (1) via m statistical tests for the uniformity of the 1D local PIT covariates and an additional test for their mutual independence.

Multivariate Local Coverage Tests (LCT). Noting that for every $i = 1, \dots, m$ we have

$$P_i(\boldsymbol{\theta}, x) \mid x \sim \mathcal{U}(0,1) \iff \forall \alpha \in [0,1] \quad r_{i,\alpha}(x) = \mathbb{P}(P_i(\boldsymbol{\theta}, x) \leq \alpha \mid x) = \alpha \tag{3}$$

and following the same approach as in [19], we propose m test statistics

$$\mathcal{T}_i(x) := \frac{1}{|G|} \sum_{\alpha \in G} \left(\widehat{r}_{i,\alpha}(x) - \alpha\right)^2 \quad \forall i \in [1, m] \, , \tag{4}$$

where G is a grid of α-values and the estimators $\widehat{r}_{i,\alpha}$ are obtained by regressing $\mathbb{I}_{\{P_i(\boldsymbol{\theta},x) \leq \alpha\}}$ on x, which is optimal for $\mathbb{E}[\mathbb{I}_{\{P_i(\boldsymbol{\theta},x) \leq \alpha\}} \mid x] = r_{i,\alpha}(x)$ (when using appropriate loss-functions [12]). We refer to Algorithm 1 (resp. 2) in [19] for computing the p-values (resp. confidence bands) associated to each test. Since there are m independent tests, we recur to a Bonferroni correction of the p-values (resp. confidence levels) [2]. These tests come with interpretable graphical diagnostics such as PP-plots or histograms that depict distributional deviations in 1D, as shown in Figure 1. If any of the uniformity tests is rejected, we reject $\mathcal{H}_0(x)$. If not, we proceed to the mutual independence test.

We now assume that the covariates of PIT_m conditioned on x are uniformly distributed over $(0, 1)$, which also means that every coordinate of the flow-transformation $T_\phi^{-1}(\boldsymbol{\theta}, x)$ follows a normal distribution *given* x (cf. Theorem 1). Their mutual independence is thus characterized by a covariance matrix equal to the identity $\mathbf{I}_m$. We are currently working on how to perform this check in practice.

The workflow in practice. Let $\mathcal{D} = \{x_n, \boldsymbol{\theta}_n\}_{n=1}^N$ be a calibration dataset with $(x_n, \boldsymbol{\theta}_n) \sim p(x, \boldsymbol{\theta})$ which were *not* used to train q_ϕ. We use $\mathcal{D}$ to calculate the PIT values $\mathrm{PIT}_m(\boldsymbol{\theta}_n, x_n, q_\phi)$ (2) and estimate our local test quantities as functions of x. We investigate the *consistency* of q_ϕ in two parts:

(1) *Global consistency check*: first, we look at the global PIT-distribution, i.e. on average over the entire x-space. More specifically, we directly compute the empirical approximation of

$$r_{i,\alpha} = \mathbb{P}(P_i(\boldsymbol{\theta}, x) \leq \alpha) \approx \tfrac{1}{N} \textstyle\sum_{n=1}^N \mathbb{I}_{\{P_i(\boldsymbol{\theta}_n, x_n) \leq \alpha\}}, \quad \forall i \in [1, m]$$

with samples from $\mathcal{D}$ to test the global uniformity of each PIT-covariate and check which one(s) might be responsible for making q_ϕ deviate from the true posterior distribution. Note that passing such global test is only a neccessary condition for consistency of q_ϕ, as it is insensitive to covariate transformations in x-space [19] (and ignores the condition on mutual independence).

(2) *Local consistency check*: we construct $m \times |G|$ transformed datasets from $\mathcal{D}$

$$\mathcal{D}_{i,\alpha} = \{(x_n, W_n^{i,\alpha})\}_{n=1}^N, \quad \alpha \in G, \quad \forall i \in [1, m]$$

where $W_n^{i,\alpha} = \mathbb{I}_{\{P_i(\boldsymbol{\theta}_n, x_n) \leq \alpha\}}$. We can then compute the test statistics defined in (4) for any new observation x and check whether $\mathcal{H}_0(x)$ should be rejected or not. In the latter case we proceed to the mutual independence test. Only then can we conclude on the validity of $\mathcal{H}_0(x)$ according to Theorem 1. If the check in (1) passes, this allows us to *guarantee* (or reject) consistency anywhere in x-space. Even if the check in (1) does not pass, analyzing local consistency allows to 'open the box' and better understand why and where in x-space the estimator fails. In such situations – as the goal is *not to guarantee* local consistency – it can be enough to test for the covariate-wise uniformity of PIT_m, putting the test for mutual independence aside. This approach was adopted for our numerical illustrations in Section 3.

3 Numerical illustrations

We apply our method to check the validity of a posterior estimate q_ϕ of the Jansen and Rit neural mass model (JRNMM) [8]. This well known model from computational neuroscience takes parameters $\boldsymbol{\theta} = (C, \mu, \sigma, g) \in \mathbb{R}^4$ as input and generates time series $x \in \mathbb{R}^{1024}$ with properties similar to brain signals obtained in neurophysiology. Parameter C influences the oscillatory behavior of the signals and (μ, σ) characterize their amplitude. The gain factor g rescales the signals and models the effects of the amplifier used for measuring them in practice. Note that the coupling-effect of g and (μ, σ) on the amplitude leads to intrinsic indeterminacies in the inversion of the model [16]. Our approximation q_ϕ is a conditioned masked autoregressive flow (MAF) [13] with 10 layers and implemented in the `sbi` package [20]. Based on the setup in [16], we train q_ϕ on 50 000 simulations from the JRNMM with a uniform prior defined on physiologically relevant values (cf. pair-plots in Figure 2).

Results and discussion. All test quantities are computed with a calibration dataset $\mathcal{D}$ containing 10 000 simulations and a grid of $|G| = 100$ α-values in [0, 1]. Following the workflow described in Section 2, we first check the global consistency of q_ϕ. In the left part of Figure 1, our graphical diagnostics illustrate how the c.d.f. for every PIT-covariate deviates from the identity function (black dashed line), outside of the 95%-confidence region (in gray), thus rejecting the null hypothesis of global consistency. We compare our method to SBC (right part of Figure 1) as implemented in the `sbi` package. We observe that SBC is unable to detect any inconsistencies in q_ϕ.

We proceed to investigate the local consistency of q_ϕ on different locations of x-space. Note that since the global test for consistency has not passed, we may focus on the uniformity of the local PIT-covariates (cf. (2) of the workflow in Section 2). We consider a 1D subspace in $\boldsymbol{\theta}$-space, where (C_0, μ_0, σ_0) are fixed and the gain g_0 varies in $[-20, 20]$ to generate observations x_0. The upper-left part of Figure 2 shows how the test statistics evolve with g_0. We observe a strong deviation from

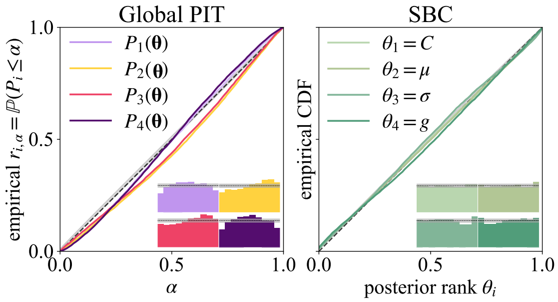

Figure 1: Global Consistency of the JRNMM posterior estimator: PP-plots and corresponding histograms obtained for every global PIT covariate $P_i(\boldsymbol{\theta})$ (right) and for SBC applied to every element θ_i of $\boldsymbol{\theta}$ (left). The gray zone indicates the 95%-confidence region of acceptance outside of which the uniformity test is rejected. SBC fails to reject the null hypothesis of global consistency.

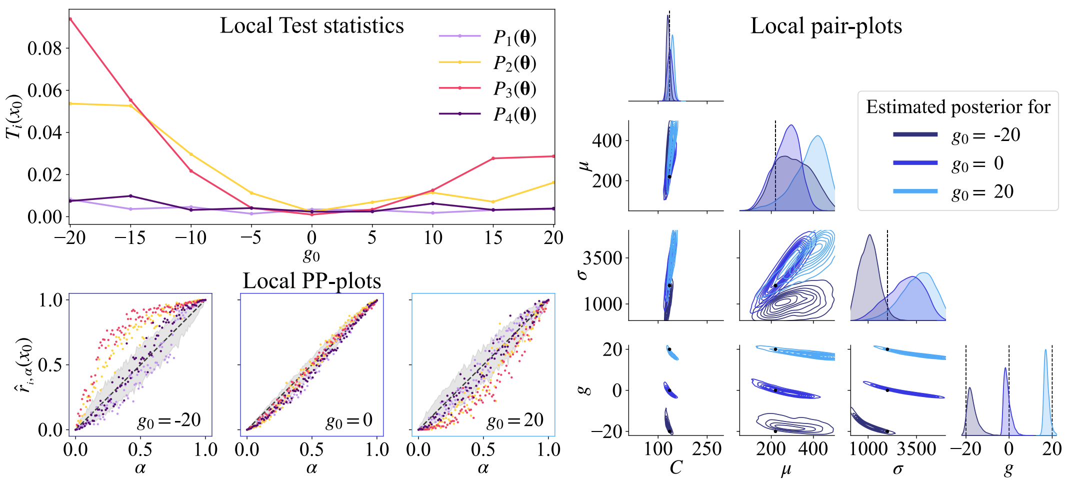

Figure 2: Local Consistency Analysis. (Left) Diagnostics obtained for every PIT-covariate, when evaluated for different x_0, simulated via the JRNMM with fixed parameters (C_0, μ_0, σ_0) and variable g_0 in $[-20, 20]$. The test-statistics are plotted as a function of g_0 and indicate inconsistent behavior (values are not constantly close to zero). Below, the PP-plots report the nature (bias/dispersion) of these inconsistencies for $g_0 \in \{-20, 0, 20\}$. The gray zone indicates the 95%-confidence region of acceptance outside of which the uniformity test is rejected. (Right) Pair-plots of $q_\phi(\boldsymbol{\theta} \mid x_0)$ with ground-truth parameters $\boldsymbol{\theta}_0 = (C_0, \mu_0, \sigma_0, g_0)$ (black dots and dashed lines) for $g_0 \in \{-20, 0, 20\}$.

uniformity for covariates P_2 and P_3: they vary smoothly in a 'U-shape', with *higher values* as g_0 deviates from zero. We also generate local PP-plots (lower-left part of Figure 2) and observe positive (resp. negative) bias for small (resp. high) values of g_0. Our procedure shows that there are certain locations in x-space (here $g_0 = 0$) where q_ϕ performs well (i.e. test statistics are close to zero and PP-plots show little deviations from the 95%-confidence region in gray), even though global consistency does not hold. These observations can be compared to the pair-plots (right part of Figure 2) which shows the estimated posterior density over the entire $\boldsymbol{\theta}$-space. We observe that gain values at the boundary of the prior support ($g_0 = -20$ and $g_0 = 20$) induce discrepancies (cf. marginals plotted in the diagonals of the pair-plots in Figure 2) that are detected by our PIT-based diagnostics.

The above results are obtained by directly applying the method from [19] using their default regressor, `MLPClassifier` from scikit-learn [15]. Indeed, this model is well suited for the binary target variables $W^{i,\alpha}$ and, based on neural networks, is able to scale to high-dimensional x-spaces. Furthermore, it has been shown [12] that for a large enough dataset $\mathcal{D}$ (and number of training iterations), the cross-entropy loss converges to the optimal solution of our regression problem. Note that although such method allows to capture and visualize local discrepancies, it does not cope well with our high-dimensional, complex data-distributions (high variances represented by large confidence regions computed over 100 trials). Also, it requires the training of a large number ($m \times |G|$) of regressors (one for each $\boldsymbol{\theta}$-dimension and α-value). Finally, this method is only applicable for the uniformity tests. We are currently investigating algorithms that are computationally more efficient and more accurate in estimating the *full* multivariate local PIT-distribution.

Conclusions. Numerical illustrations demonstrate that our diagnostics capture well the inconsistencies in q_ϕ with respect to the true JRNMM posterior. To be precise, the proposed validation method is statistically more powerful and computationally more efficient than SBC and, importantly, it allows for a local analysis that reveals where in x-space the estimations should be improved. Our method exploits useful properties of modern normalizing flows. Indeed, contrarily to other SBI-strategies [7], they allow for efficient density evaluation and use bijective transformations involving the mixing between elements of $\boldsymbol{\theta}$. These transformations can be interpreted as a set of 1D-projections of $\boldsymbol{\theta}$, including information about its joint p.d.f. (i.e. interactions between its elements θ_i). The covariates of our multivariate PIT therefore define fast and easy to compute 1D test-quantities on which we can perform univariate LCTs that do not completely ignore correlations in the $\boldsymbol{\theta}$-space. Combined with a test to check their mutual independence, this would provide theoretical guarantees for consistency, which is not the case for HPD, the existing multivariate version of LCT proposed in [19].

Ongoing work will investigate the link between PIT-covariates and their deviations from uniformity with the actual parameters of the model. This could reveal the true nature of their coupling which can be used for the development of better specified models (e.g. hierarchical posterior estimation [16, 17]). Ultimately, the goal is to use our diagnostics as a tool for model selection.

Broader Impact

This work tackles an important open question in simulation based inference. We introduce theoretically valid and interpretable validation diagnostics that scale to both high-dimensional data and parameter spaces. Our contribution should help to further improve SBI methods and drive the design of better specified models, hence allowing to build up trust on their ability to address important questions in experimental science. Moreover, our work is in line with many other tentatives in the machine learning community of ensuring the quality and calibration of complex models based on neural networks.

Acknowledgments

Julia Linhart is recipient of the Pierre-Aguilar Scholarship and thankful for the funding of the Capital Fund Management (CFM). Alexandre Gramfort thanks the support of the ANR BrAIN (ANR-20-CHIA0016) grant.

References

[1] Sean R. Bittner, Agostina Palmigiano, Alex T. Piet, Chunyu A. Duan, Carlos D. Brody, Kenneth D. Miller, and John Cunningham. Interrogating theoretical models of neural computation with emergent property inference. *eLife*, 10, 7 2021. ISSN 2050084X. doi: 10.7554/eLife.56265.

[2] J. Martin Bland and Douglas G. Altman. Multiple significance tests: the bonferroni method. *BMJ (Clinical research ed.)*, 310:170, 1 1995. ISSN 0959-8138. doi: 10.1136/BMJ.310.6973.170.

[3] Jan Boelts, Jan-Matthis Lueckmann, Richard Gao, and Jakob H. Macke. Flexible and efficient simulation-based inference for models of decision-making. *bioRxiv*, page 2021.12.22.473472, 6 2022. ISSN 2050084X. doi: 10.1101/2021.12.22.473472.

[4] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences (PNAS)*, 117:30055–30062, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1912789117.

[5] Niccolò Dalmasso, David Zhao, Rafael Izbicki, and Ann B. Lee. Likelihood-free frequentist inference: Confidence sets with correct conditional coverage. 7 2021. doi: 10.48550/arxiv.2107.03920.

[6] Pedro J. Gonçalves, Jan Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F. Podlaski, Sara A. Haddad, Tim P. Vogels, David S. Greenberg, and Jakob H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:1–46, 9 2020. ISSN 2050084X. doi: 10.7554/ELIFE.56261.

[7] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4239–4248. PMLR, 13–18 Jul 2020. doi: 10.48550/arxiv.1903.04057.

[8] Ben H. Jansen and Vincent G. Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics 1995 73:4*, 73: 357–366, 9 1995. ISSN 1432-0770. doi: 10.1007/BF00199471.

[9] Ilmun Kim, Ann B. Lee, and Jing Lei. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13:5253–5305, 12 2018. ISSN 19357524. doi: 10.48550/arxiv.1812.08927.

[10] Jan-Matthis Lueckmann, Jan Boelts, David S Greenberg, Pedro J Gonçalves, and Jakob H Macke. Benchmarking simulation-based inference. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (PMLR)*, 130:343–351, 4 2021. doi: 10.48550/arxiv.2101.04653.

[11] Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, and Ann B. Lee. Simulation-based inference with waldo: Perfectly calibrated confidence regions using any prediction or posterior estimation algorithm. 5 2022. doi: 10.48550/arxiv.2205.15680.

[12] John W. Miller, Rod Goodman, and Padhraic Smyth. On loss functions which minimii to conditional expected values and posterior probabilities. *IEEE TRANSACTIONS ON INFORMAITON THEORY*, 39, 1993.

[13] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2339–2348, 12 2017. ISSN 10495258. doi: 10.48550/arxiv.1705.07057.

[14] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22:1–64, 2021. ISSN 15337928. doi: 10.48550/arxiv.1912.02762.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] Pedro L C Rodrigues, Thomas Moreau, and Gilles Louppe. Hnpe: Leveraging global parameters for neural posterior estimation. *Neural Information Proceedings (NeurIPS)*, 12 2021. ISSN 10495258. doi: 10.48550/arxiv.2102.06477.

[17] Louis Rouillard and Demian Wassermann. ADAVI: Automatic Dual Amortized Variational Inference Applied To Pyramidal Bayesian Models. In *ICLR 2022*, Virtual, France, April 2022. doi: 10.48550/arXiv.2106.12248.

[18] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. 4 2018. doi: 10.48550/arxiv.1804.06788.

[19] David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B. Lee. Diagnostics for conditional density models and bayesian inference algorithms. *PMLR*, 2 2021. ISSN 2640-3498.

[20] Álvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J Gonçalves, David S Greenberg, Jakob H Macke, Computational Neuroengineering, and T U Munich. Sbi – a toolkit for simulation-based inference. *Journal of Open Source Software*, 5:2505, 7 2020. ISSN 2475-9066. doi: 10.21105/joss.02505.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Theoretical results and workflow in Section 2, numerical illustrations and discussion in Section 3.
   (b) Did you describe the limitations of your work? [Yes] See Section 3: last paragraph in "results and discussion" and "conclusions".
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] At the beginning of Section 2, right before introducing the multivariate PIT.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The exact script for results reproduction is not provided. However, we refer (in section 3) to the code from different works that were directly used to simulate data, train the normalizing flow and estimate local test-quantities (modula minor changes like python package updates in `sbi`).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Data splits (size of training and calibration datasets) and number of layers for the normalizing flow are explicitly mentioned. All other hyper parameters are implicit by referring to the used implementations.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] As our work does not compare ML-models, no such results are needed.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] This includes the implementation of the simulator, the normalizing flow, posterior estimation and the univariate LCT method.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No need, we work with *simulated* data and the implementation for the simulator is publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No need, we work with *simulated* data.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Proof of Theorem 1. Let $x \in \mathbb{R}^d$. We consider the random variables defined on $\mathbb{R}^m$: $\Theta \sim p(\boldsymbol{\theta} \mid x)$ and $\Theta_\phi \sim q_\phi(\boldsymbol{\theta} \mid x)$. Our local null hypothesis (cf. Equation 1), can be rewritten as

$$\mathcal{H}_0(x) \ : \ \text{Prob}\Big(\Theta \in \Omega_\theta\Big) = \text{Prob}\Big(\Theta_\phi \in \Omega_\theta\Big) \quad \forall \Omega_\theta \subset \mathbb{R}^m \tag{5}$$

From the definition of our normalizing flow q_ϕ with bijective transformation T_ϕ and normal base distribution, there exists $Z \sim p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}_m)$, such that $\Theta_\phi = T_\phi(Z, x)$.

Replacing Θ_ϕ by $T_\phi(Z, x)$ in (5) and defining $\Omega_z = T_\phi^{-1}(\Omega_\theta; x)$, we have

$$\text{Prob}\Big(\Theta \in \Omega_\theta\Big) = \text{Prob}\Big(T_\phi(Z; x) \in \Omega_\theta\Big) \quad \iff \quad \text{Prob}\Big(T_\phi^{-1}(\Theta; x) \in \Omega_z\Big) = \text{Prob}\Big(Z \in \Omega_z\Big)$$

Considering $T_{\phi,i}^{-1}$, the i-th coordinate of T_ϕ^{-1}, and $Z_1, \dots, Z_m$ the mutually independent and normally distributed covariates of $Z \sim \mathcal{N}(0, \mathbf{I}_m)$, we can rewrite the right part of the above equivalence as:

$$\text{Prob}\Big(T_{\phi,1}^{-1}(\Theta; x) \in \Omega_{z,1} \cap \dots \cap T_{\phi,m}^{-1}(\Theta; x) \in \Omega_{z,m}\Big) = \prod_{i=1}^{m} \text{Prob}\Big(Z_i \in \Omega_{z,i}\Big) \tag{6}$$

where the $\Omega_{z,i}$ are projections of the set Ω_z on each of its dimensions.

Note that since the above equivalence is true for *any* choice of Ω_θ, hence of Ω_z, we can write what happens when we fix a given $\Omega_{z,1}$ and let $\Omega_{z,2} = \dots = \Omega_{z,m} = \mathbb{R}$:

$$\text{Prob}\Big(T_{\phi,1}^{-1}(\Theta; x) \in \Omega_{z,1}\Big) = \text{Prob}\Big(Z_1 \in \Omega_{z,1}\Big)$$

Doing the same for all other coordinates and using these equalities in (6), we end up with

$$\forall i \in [1, m], \quad \forall \Omega_{z,i} \subset \mathbb{R}, \quad \text{Prob}\Big(T_{\phi,i}^{-1}(\Theta; x) \in \Omega_{z,i}\Big) = \text{Prob}\Big(Z_i \in \Omega_{z,i}\Big) \tag{7}$$

$$\text{and} \quad \text{Prob}\Big(T_{\phi,1}^{-1}(\Theta; x) \in \Omega_{z,1} \cap \dots \cap T_{\phi,m}^{-1}(\Theta; x) \in \Omega_{z,p}\Big) = \prod_{i=1}^{m} \text{Prob}\Big(T_{\phi,i}^{-1}(\Theta; x) \in \Omega_{z,i}\Big) \tag{8}$$

where (8) is the definition of mutual independence itself.

We now apply the c.d.f. $F_{\mathcal{N}(0,1)}$ to each random variable in (7) and (8). Mutual independence stays true and the probability integral transform theorem (in 1D) states that $F_{\mathcal{N}(0,1)}(Z_i) \sim \mathcal{U}(0, 1)$. We therefore get that the null hypothesis $\mathcal{H}_0(x)$ holds if, and only if,

$$\begin{cases} P_i(\Theta, x) = F_{\mathcal{N}(0,1)}(T_{\phi,i}^{-1}(\Theta; x)) \sim \mathcal{U}(0, 1), \quad \forall i = [1, m] \quad \textit{and} \\ \{P_i(\Theta, x)\}_{i=1,\dots,m} \quad \textit{are mutually independent} \end{cases} \tag{9}$$

The result in Theorem 1 directly follows from rewriting (5) and (9) with initial notations from Section 2: remember that $\Theta \sim p(\boldsymbol{\theta} \mid x)$, so $P_i(\Theta, x) \sim p(P_i(\boldsymbol{\theta}, x) \mid x)$ and we get:

$$p(\boldsymbol{\theta} \mid x) = q_\phi(\boldsymbol{\theta} \mid x) \iff \begin{cases} p(P_i(\boldsymbol{\theta}, x) \mid x) = \mathcal{U}(0, 1) \quad \forall i \in [1, m] \quad \textit{and} \\ \{P_i(\boldsymbol{\theta}, x) \mid x\}_{i=1,\dots,m} \quad \textit{are mutually independent} \end{cases}$$

where $P_i(\boldsymbol{\theta}, x)$ is the i^{th} covariate of $\text{PIT}_m(\boldsymbol{\theta}, x, q_\phi)$, the multivariate PIT of $\boldsymbol{\theta}$ at x, associated to q_ϕ.

Conclusion: The null hypothesis $\mathcal{H}_0(x)$ holds if, and only if, the covariates of PIT_m conditioned on x are mutually independent and uniformly distributed over $(0, 1)$.