

1

2 **Supplementary Information for**

3 **Learning complex models with invertible neural networks: a likelihood-free Bayesian** 4 **approach**

5 **Stefan T. Radev, Ulf K. Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe**

6 **Stefan Radev.**

7 **E-mail: stefan.radev@psychologie.uni-heidelberg.de**

8 **This PDF file includes:**

- 9 Supplementary text
- 10 Fig. S1
- 11 Captions for Movies S1 to S3
- 12 Captions for Databases S1 to S2
- 13 References for SI reference citations

14 **Other supplementary materials for this manuscript include the following:**

- 15 Movies S1 to S3
- 16 Databases S1 to S2

17 Supporting Information Text

18 Results

19 **Performance metrics.** In the following, the computation of the performance metrics used throughout the main text is detailed.

Normalized Root Mean Squared Error. The normalized root mean squared error (NRMSE) between a sample of true parameters $\{\theta^{(i)}\}_{i=1}^n$ and a sample of estimated parameters $\{\hat{\theta}^{(i)}\}_{i=1}^n$ is given by:

$$NRMSE = \sqrt{\sum_{i=1}^n \frac{(\theta^{(i)} - \hat{\theta}^{(i)})^2}{\theta_{max} - \theta_{min}}} \quad [1]$$

20 Due to the normalization factor $\theta_{max} - \theta_{min}$, the NRMSE is scale-independent, and thus suitable for comparing the recovery
21 across parameters having different numerical ranges. The NRMSE is zero when the estimates are exactly equal to the true
22 values.

Coefficient of Determination. The coefficient of determination R^2 gives the proportion of variance in a sample of true parameters $\{\theta^{(i)}\}_{i=1}^n$ that is "explained" by a sample of estimated parameters $\{\hat{\theta}^{(i)}\}_{i=1}^n$. It is computed as:

$$R^2 = 1 - \sum_{i=1}^n \frac{(\theta^{(i)} - \hat{\theta}^{(i)})^2}{(\theta^{(i)} - \bar{\theta})^2} \quad [2]$$

23 where $\bar{\theta}$ denotes the mean of the true parameter samples. When R^2 equals 1, it means that the estimates are perfect
24 reconstructions of the true parameters.

Kullback-Leibler Divergence. The Kullback-Leibler divergence (D_{KL}) quantifies the increase in entropy incurred by approximating a target probability distribution P with a distribution Q . Its general form for absolutely continuous distributions is given by

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad [3]$$

where p and q denote the pdfs of P and Q . In the case where P and Q are both multivariate Gaussian distributions, the KL divergence can be computed in closed form (1):

$$D_{KL}(P||Q) = \frac{1}{2} \left[\log \frac{\det \Sigma_q}{\det \Sigma_p} + \text{Tr}(\Sigma_q^{-1} \Sigma_p) - d + (\mu_p - \mu_q)^T \Sigma_q^{-1} (\mu_p - \mu_q) \right] \quad [4]$$

where Σ_p and Σ_q denote the covariance matrices of p and q , μ_p and μ_q the respective mean vectors, and d the number of dimensions of the Gaussian. In the case of diagonal Gaussian distributions, Eq.4 reduces to:

$$D_{KL}(P||Q) = \sum_{i=1}^d \left(\log \frac{\sigma_{q,i}}{\sigma_{p,i}} + \frac{\sigma_{p,i}^2 + (\mu_{q,i} - \mu_{p,i})^2}{2\sigma_{q,i}^2} - \frac{1}{2} \right) \quad [5]$$

25 Even though the KL divergence is not a proper distance metric, as it is not symmetric in its arguments, it can be used to
26 quantify the error of approximation and serve as a metric for comparing different methods.

Simulation-Based Calibration. Simulation-based calibration is a recently proposed method for validating the accuracy of posterior samples generated by a Bayesian sampling method (2). It is based on the so called *self-consistency* of the Bayesian joint distribution. Given a sample from the prior distribution $\tilde{\theta} \sim p(\theta)$ and a sample from the data-generating process $\tilde{x} \sim p(x|\tilde{\theta})$, one can integrate $\tilde{\theta}$ and \tilde{x} out of the Bayesian joint distribution to recover back the prior of θ :

$$p(\theta) = \int p(\theta, \tilde{\theta}, \tilde{x}) d\tilde{x} d\tilde{\theta} \quad [6]$$

$$= \int p(\theta, \tilde{x}|\tilde{\theta}) p(\tilde{\theta}) d\tilde{x} d\tilde{\theta} \quad [7]$$

$$= \int p(\theta|\tilde{x}) p(\tilde{x}|\tilde{\theta}) p(\tilde{\theta}) d\tilde{x} d\tilde{\theta} \quad [8]$$

27 If the Bayesian sampling method produces samples from the exact posterior, the equality implied by Eq.8 should hold regardless
28 of the particular form of the posterior. Thus, any violation of this equality indicates some error incurred by the sampling
29 method. The authors of (2) propose **Algorithm 1** for visually detecting such violations:

30 **Algorithm 1** is justified, since Eq.8 implies that the rank statistic defined in line 5 should be uniformly distributed. Hence,
31 any deviations from uniformity indicate some error in the approximate posterior.

Algorithm 1 Simulation-based calibration (SBC) for a single parameter θ

```
1: for  $i = 1, \dots, n$  do  
2:   Sample  $\tilde{\theta}^{(i)} \sim p(\theta)$   
3:   Simulate a dataset  $\mathbf{x}^{(i)} = q(\tilde{\theta}^{(i)})$   
4:   Draw posterior samples  $\{\theta^{(l)}\}_{l=1}^L \sim p(\theta|\mathbf{x}^{(i)})$   
5:   Compute rank statistic  $r^{(i)} = \sum_{l=1}^L \mathbb{1}_{[\theta^{(l)} < \tilde{\theta}^{(i)}]}$   
6:   Store  $r^{(i)}$   
7: end for  
8: Create a histogram of  $\{r^{(i)}\}_{i=1}^n$  and analyze it for uniformity
```

32 **Model details.**

Simulation of the Ricker model. We place the following uniform priors over the Ricker model parameters:

$$\rho \sim \mathcal{U}(0, 15) \quad [9]$$

$$r \sim \mathcal{U}(1, 90) \quad [10]$$

$$\sigma \sim \mathcal{U}(0.05, 0.7) \quad [11]$$

33 These ranges are very broad, as datasets generated by extreme parameter values appear implausible in real-world scenarios.
34 Nevertheless, we stick to broad priors for training, even though parameter recovery might degrade at the extremes. Figure
35 XXX depicts some Ricker datasets generated by parameters drawn from the specified prior.

Simulation of the Lévy-Flight Model. The following uniform priors over the LFM parameters are used for simulation:

$$v_0 \sim \mathcal{U}(0, 6) \quad [12]$$

$$v_1 \sim \mathcal{U}(-6, 0) \quad [13]$$

$$zr \sim \mathcal{U}(0.3, 0.7) \quad [14]$$

$$a \sim \mathcal{U}(0.6, 3) \quad [15]$$

$$t_0 \sim \mathcal{U}(0.3, 1) \quad [16]$$

$$\alpha \sim \mathcal{U}(1, 2) \quad [17]$$

36 These priors are broad enough to cover the range of realistic reaction times observed in choice RT experiments.

Simulation of the stochastic SIR model. We place the following uniform priors over the two rate parameters of the stochastic SIR model:

$$\beta \sim \mathcal{U}(0.01, 1) \quad [18]$$

$$\gamma \sim \mathcal{U}(0.01, \beta) \quad [19]$$

37 Figure XXX depicts some SIR datasets generated by parameters drawn from the specified prior.

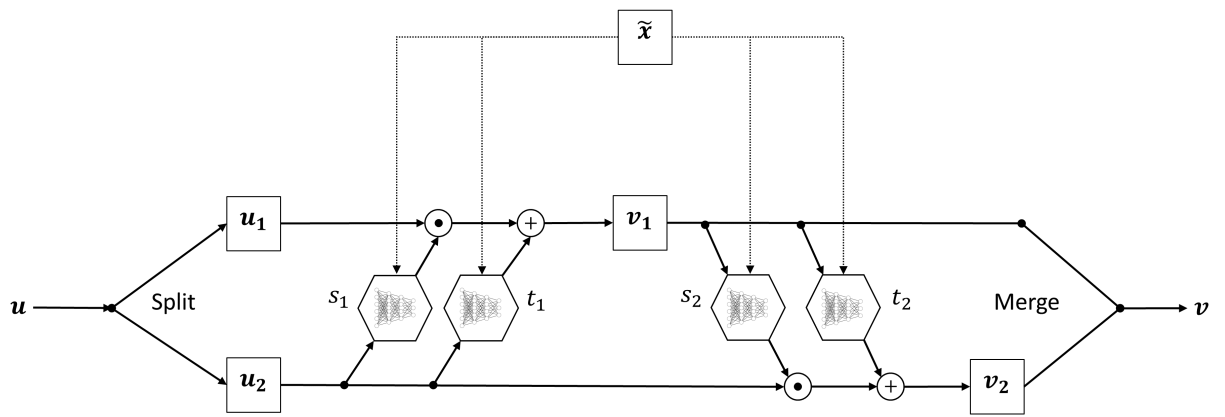


Fig. S1. Second figure

38 **Movie S1.** Type caption for the movie here.

39 **Movie S2.** Type caption for the other movie here. Adding longer text to show what happens, to decide on
40 alignment and/or indentations.

41 **Movie S3.** A third movie, just for kicks.

42 **Additional data table S1 (dataset_one.txt)**

43 Type or paste caption here.

44 **Additional data table S2 (dataset_two.txt)**

45 Type or paste caption here. Adding longer text to show what happens, to decide on alignment and/or indentations for
46 multi-line or paragraph captions.

47 **References**

- 48 1. Hershey JR, Olsen PA (2007) Approximating the kullback leibler divergence between gaussian mixture models in *2007*
49 *IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. (IEEE), Vol. 4, pp. IV–317.
- 50 2. Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2018) Validating bayesian inference algorithms with simulation-
51 based calibration. *arXiv preprint arXiv:1804.06788*.