Learning Robust Statistics for Simulation-based Inference under Model Misspecification

Daolang Huang*[1], Ayush Bharti*[1], Amauri Souza[1], Luigi Acerbi[2], Samuel Kaski[1,3]

Department of Computer Science, Aalto University[1]
Department of Computer Science, University of Helsinki[2]
Department of Computer Science, University of Manchester[3]

December 1, 2023

Simulation-based inference

- Data $\mathbf{y} = f y_i g_{i=1}^n$ R^d denoted by empirical distribution Q^n
- Simulator-based model $P_\Theta = f \mathrm{P}_\theta : \ 2\ \Theta g$
- P_θ is intractable, but sampling $\mathbf{x}$ P_θ is straightforward
- **Aim:** Estimate given data $\mathbf{y}$

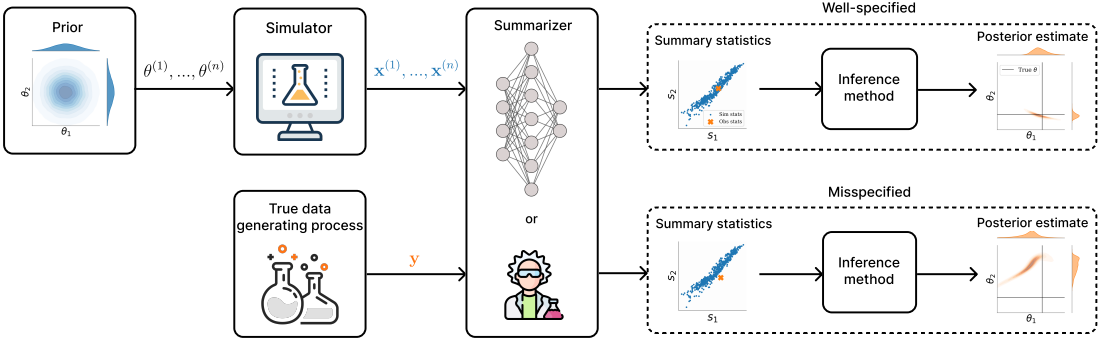
Simulation-based inference

- Data $\mathbf{y} = f y_i g_{i=1}^n$ R^d denoted by empirical distribution Q^n
- Simulator-based model $P_\Theta = f \mathrm{P}_\theta : \quad 2\ \Theta g$
- P_θ is intractable, but sampling $\mathbf{x}$ P_θ is straightforward
- **Aim:** Estimate given data $\mathbf{y}$
- **Solution:** methods based on distances like approximate Bayesian computation (ABC); methods based on deep neural networks like neural posterior estimation (NPE)

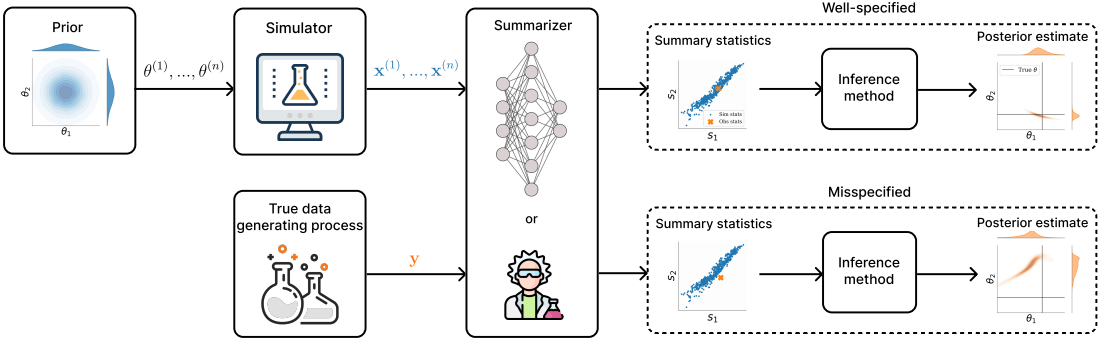
Simulation-based inference

- Data $\mathbf{y} = f y_i g_{i=1}^n$ R^d denoted by empirical distribution Q^n
- Simulator-based model $P_\Theta = f P_\theta : \ 2\ \Theta g$
- P_θ is intractable, but sampling $\mathbf{x}$ P_θ is straightforward
- **Aim:** Estimate given data $\mathbf{y}$
- **Assumption:** Model is "correct", i.e., $Q^n\ 2\ P_\Theta$
- **Problem:** Model misspecification, i.e. Q^n 2 P_Θ) @ $2\ \Theta$ s.t. $P_\theta = Q^n$
  - Stochasticity in data collection process (outliers, missing data, broken independence assumption, etc.)
  - "All models are wrong..."
- **Even more problem:** Inference is based on simulation from misspecified model!

Inference is based on summary statistics



Inference is based on summary statistics



Insight 1: Under misspecification, observed statistic goes outside the set of simulated statistics

⟩ SBI methods have to generalize outside their training data

Insights

Insight 1: Under misspecification, observed statistic goes outside the set of simulated statistics

$\Rightarrow$ SBI methods have to generalize outside their training data

Insight 2: Even if model is misspecified ($Q^n \not\simeq P_\Theta$), it may be well-specified w.r.t the statistics

- Example: Gaussian model, skewed data
- Misspecified if statistics are sample mean and sample skewness
- Well-specified if statistics are sample mean and sample variance
- If we pick statistics appropriately, we can be robust!

Learning robust statistics for SBI

proposed loss = usual loss + D(simulated statistics, observed statistic)

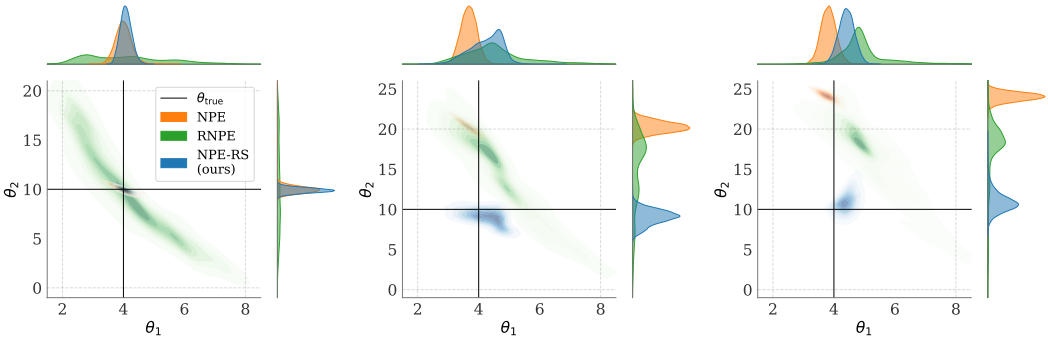
Learning robust statistics for SBI

proposed loss = usual loss + D(simulated statistics, observed statistic)

- For ABC or other SBI methods, usual loss is autoencoder's reconstruction loss
- For NPE, statistics and posterior can be learned jointly
- We want D to be outlier-robust. Hence, maximum mean discrepancy.
- Regularizer : encodes trade-off between accuracy and robustness

Results

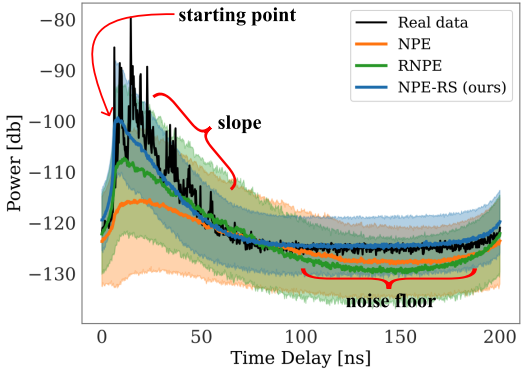
- **Ricker model:** 2 parameters
- **Inference method:** Neural posterior estimation (NPE)
- **-contamination model:** Q = (1)P$_{\theta_{\text{true}}}$ + P$_{\theta_c}$



(a) Well-speci ed (ϵ = 0) (b) Misspeci ed (ϵ = 10%) (c) Misspeci ed (ϵ = 20%)

Radio propagation example

- 4 parameters
- Data dimension: 801
- Model misspecified due to broken iid assumption



Conclusion

- We propose a simple solution for tackling misspecification of simulator-based models.
- Our method can be applied to any SBI method that utilizes summary statistics.
- Our method only has one hyperparameter balancing efficiency and robustness.
- We show robustness under misspecified scenarios with both synthetic and real-world data.