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]

January 19, 2024

Inference problem

- Data $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- **Aim:** Estimate θ given data $\mathbf{x}$ (maximizing likelihood, sampling from posterior)

Inference problem

- Data $\mathbf{x} = \{x_i\}_{i=1}^{n} \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- **Aim:** Estimate θ given data $\mathbf{x}$ (maximizing likelihood, sampling from posterior)
- **Assumption:** Model is "correct", i.e., $\mathbb{Q}^n \in \mathcal{P}_\Theta$
- **Problem:** Model misspecification, i.e. $\mathbb{Q}^n \notin \mathcal{P}_\Theta \Rightarrow \nexists \theta \in \Theta$ s.t. $\mathbb{P}_\theta = \mathbb{Q}^n$

Inference problem

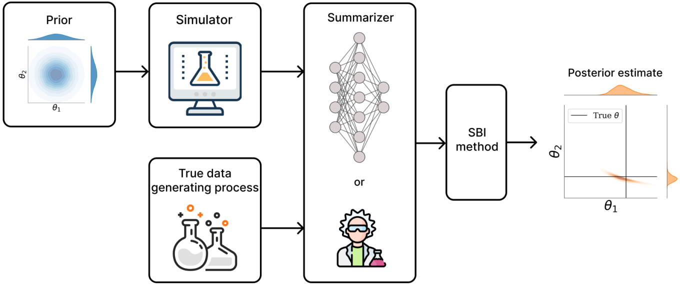
- Data $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- **Aim:** Estimate θ given data $\mathbf{x}$ (maximizing likelihood, sampling from posterior)
- **Assumption:** Model is "correct", i.e., $\mathbb{Q}^n \in \mathcal{P}_\Theta$
- **Problem:** Model misspecification, i.e. $\mathbb{Q}^n \notin \mathcal{P}_\Theta \Rightarrow \nexists \theta \in \Theta$ s.t. $\mathbb{P}_\theta = \mathbb{Q}^n$
  - Stochasticity in data collection process (outliers, missing data, broken independence assumption)
  - "All models are wrong..."
- Inference outcomes are unreliable under misspecification

Inference for simulators

- Data $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Simulator-based model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- $\mathbb{P}_\theta$ is intractable, but sampling $y \sim \mathbb{P}_\theta$ is straightforward
- **Aim:** Estimate θ given data $\mathbf{x}$ (~~maximizing likelihood, sampling from posterior~~)

Inference for simulators

- Data $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Simulator-based model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- $\mathbb{P}_\theta$ is intractable, but sampling $y \sim \mathbb{P}_\theta$ is straightforward
- **Aim:** Estimate θ given data $\mathbf{x}$ (~~maximizing likelihood, sampling from posterior~~)
- **Solution:** Simulation-based inference



Simulation-based inference (SBI)

Approximate Bayesian computation (ABC)

Repeat until m samples accepted:

- Sample from prior $\theta^\star \sim p(\theta)$
- Simulate data from model, $\mathbf{y} \sim \mathbb{P}_{\theta^\star}$
- If $d\left(\eta(\mathbf{y}), \eta(\mathbf{x})\right) < \epsilon$, accept $\theta^\star$

Simulation-based inference (SBI)

Approximate Bayesian computation (ABC)

Repeat until m samples accepted:

- Sample from prior $\theta^\star \sim p(\theta)$
- Simulate data from model, $\mathbf{y} \sim \mathbb{P}_{\theta^\star}$
- If $d\left(\eta(\mathbf{y}), \eta(\mathbf{x})\right) < \epsilon$, accept $\theta^\star$

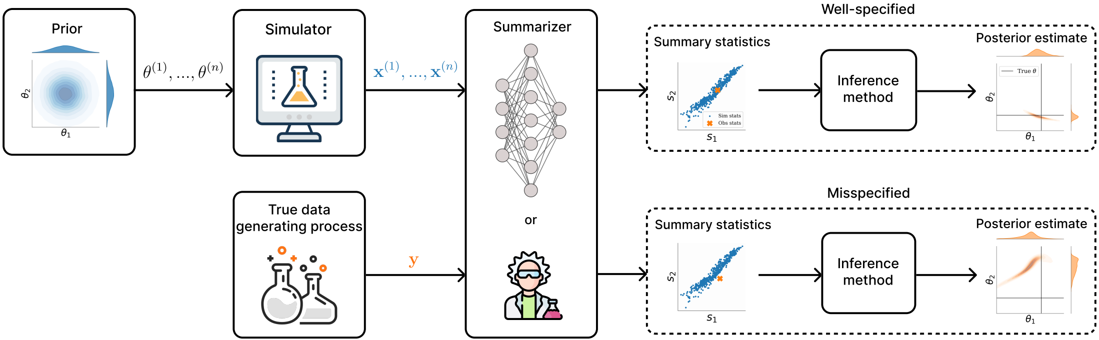
Neural posterior estimation (NPE)

- Sample from prior $\theta_1, \ldots, \theta_n \sim p(\theta)$
- Simulate data from model, $\mathbf{y}_i \sim \mathbb{P}_{\theta_i}, i = 1, \ldots, n$. Training data: $\{(\theta_i, \mathbf{y}_i)\}_{i=1}^n$
- Assume posterior is member of a distribution family q_ν
- Learn a map from the statistics $\eta(\mathbf{y})$ to the posterior (i.e. ν) using e.g. normalizing flows

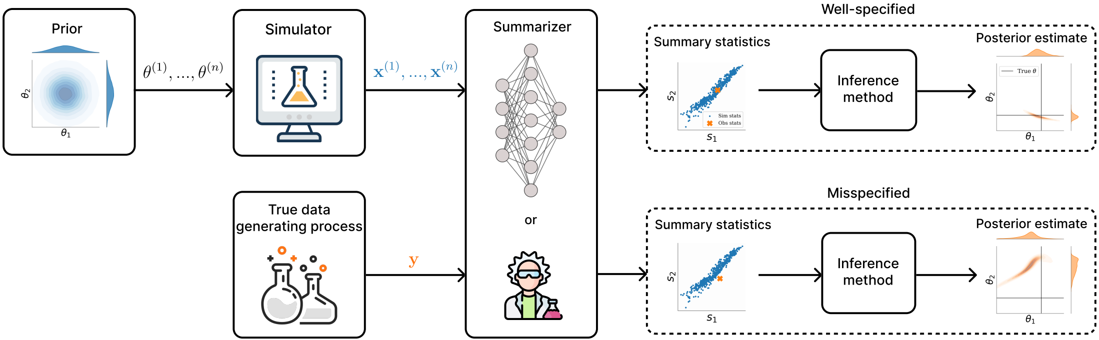
Inference for simulators

- Data $\mathbf{x} = \{x_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ denoted by empirical distribution $\mathbb{Q}^n$
- Simulator-based model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$
- $\mathbb{P}_\theta$ is intractable, but sampling $y \sim \mathbb{P}_\theta$ is straightforward
- **Aim:** Estimate θ given data $\mathbf{x}$ (~~maximizing likelihood, sampling from posterior~~)
- **Assumption:** Model is "correct", i.e., $\mathbb{Q}^n \in \mathcal{P}_\Theta$
- **Problem:** Model misspecification, i.e. $\mathbb{Q}^n \notin \mathcal{P}_\Theta \Rightarrow \nexists\theta \in \Theta$ s.t. $\mathbb{P}_\theta = \mathbb{Q}^n$
    - Stochasticity in data collection process (outliers, missing data, broken independence assumption, etc.)
    - "All models are wrong..."
    - Numerical approximations
- **Even more problem:** Inference is based on simulation from misspecified model!

Inference is based on summary statistics



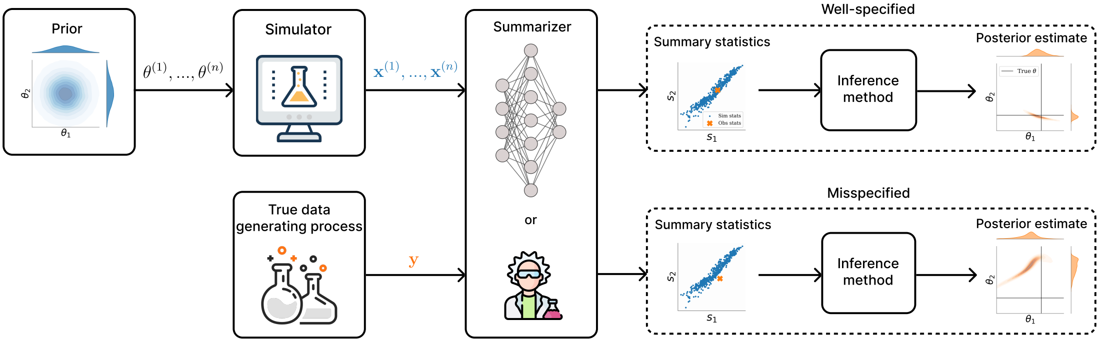
Inference is based on summary statistics



Insight 1: Even if model is misspecified ($\mathbb{Q}^n \notin \mathcal{P}_\Theta$), 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

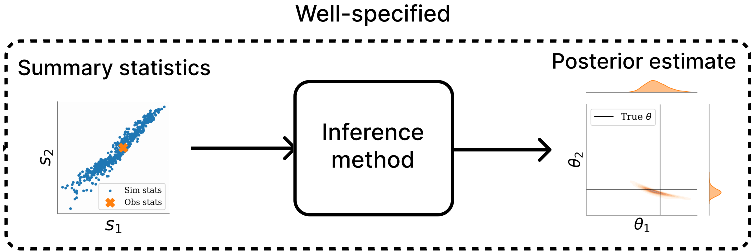
Inference is based on summary statistics

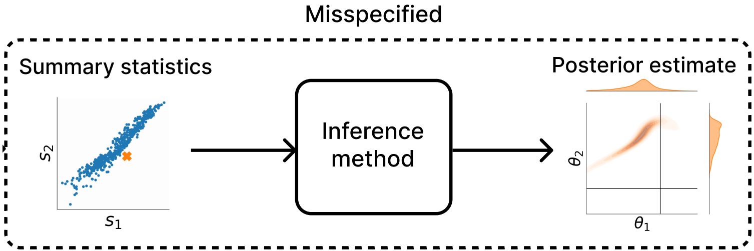


Insight 1: Even if model is misspecified ($\mathbb{Q}^n \notin \mathcal{P}_\Theta$), it may be well-specified w.r.t the statistics

- 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!

Inference is based on summary statistics





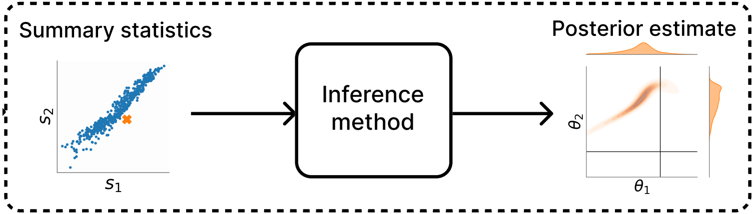
Insights

Insight 1: Even if model is misspecified ($\mathbb{Q}^n \notin \mathcal{P}_\Theta$), 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!

Insight 2: Under misspecification, observed statistic goes outside the set of simulated statistics

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



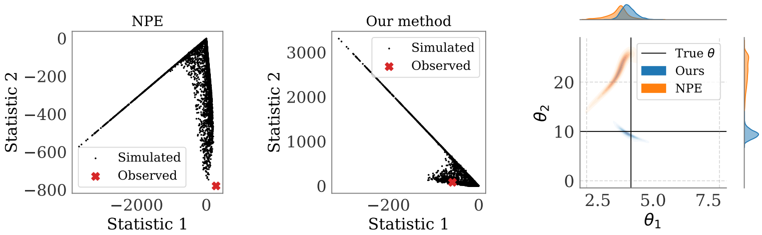
Learning robust statistics for SBI

proposed loss $=$ usual loss $+$ $\lambda\mathcal{D}$(simulated statistics, observed statistic)

Learning robust statistics for SBI

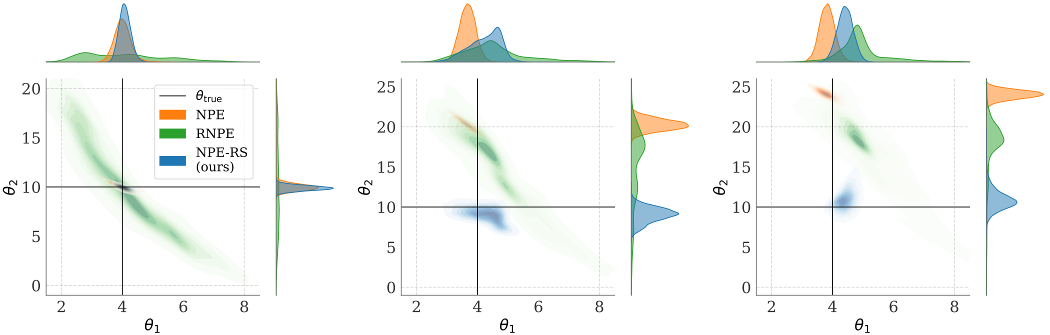
$$\text{proposed loss} = \text{usual loss} + \lambda\mathcal{D}(\text{simulated statistics, observed statistic})$$

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



Results

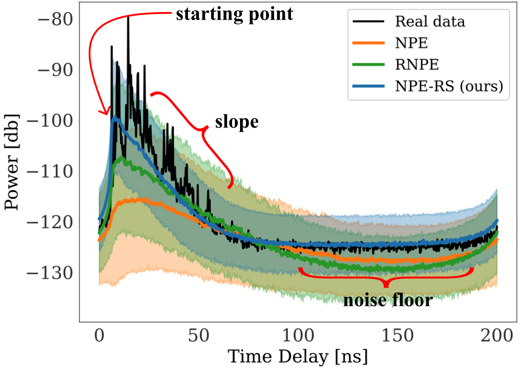
- **Ricker model:** 2 parameters
- **Inference method:** Neural posterior estimation (NPE)
- **ϵ-contamination model:** $\mathbb{Q} = (1-\epsilon)\mathbb{P}_{\theta_{\text{true}}} + \epsilon\mathbb{P}_{\theta_c}$



(a) Well-specified ($\epsilon = 0$) (b) Misspecified ($\epsilon = 10\%$) (c) Misspecified ($\epsilon = 20\%$)

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.