# 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]

ISBA 2024

# Inference problem

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

# Inference problem

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

# Inference for simulators

- Data $\mathbf{y} = \{y_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ 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 $x \sim \mathbb{P}_\theta$ is straightforward
- **Aim:** Estimate $\theta$ given data $\mathbf{y}$ (~~maximizing likelihood, sampling from posterior~~)
- Simulators in sciences and engineering:
  - Population genetics [Pritchard et al., 1999]
  - Ecology and evolution [Beaumont, 2010]
  - Astrophysics [Akeret et al., 2015]
  - Epidemiology [Kypraios et al., 2017]
  - Radio communications [Bharti et al., 2022]
  - Economics [Dyer et al., 2022]
- **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, $\mathbf{x} \sim \mathbb{P}_{\theta^\star}$
- 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, $\mathbf{x} \sim \mathbb{P}_{\theta^\star}$
- If $d\left(\eta(\mathbf{y}), \eta(\mathbf{x})\right) < \epsilon$, accept $\theta^\star$

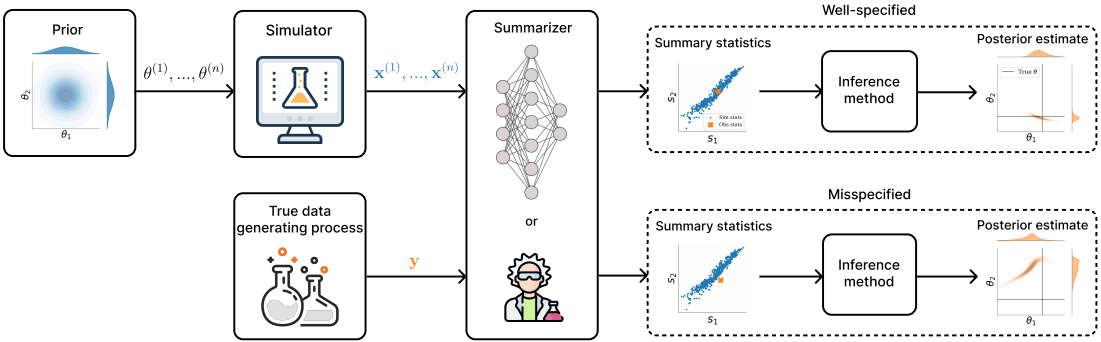
## Neural posterior estimation (NPE)

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

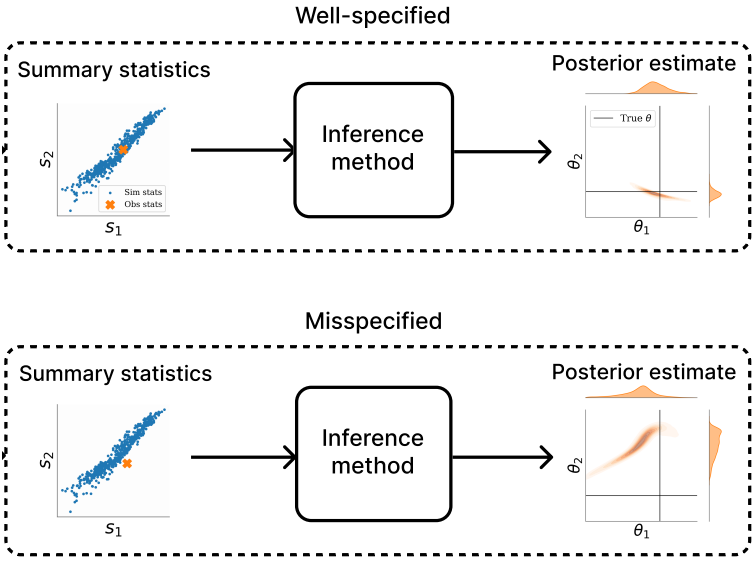
# Inference for simulators

- Data $\mathbf{y} = \{y_i\}_{i=1}^n \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ 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 $\theta$ given data $\mathbf{y}$ (~~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



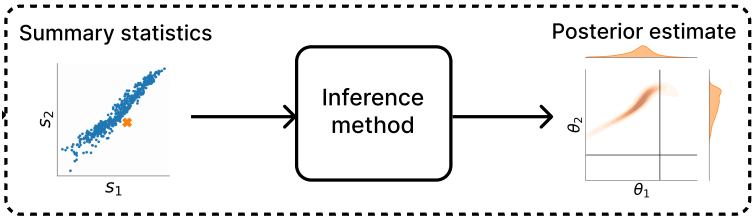
# Inference is based on summary statistics



# Insight 1

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

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



# Insight 2

**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

# Insight 2

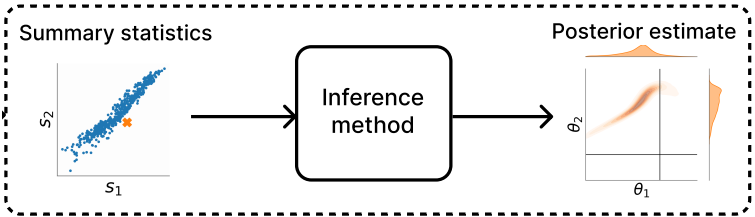
**Insight 1:** Even if model is misspecified ($\mathbb{Q}^n \notin \mathcal{P}_\Theta$), it may be well-specified w.r.t the statistics

- 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!

# 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 ($\mathbb{Q}^n \notin \mathcal{P}_\Theta$), it may be well-specified w.r.t the statistics

$\Rightarrow$ If we learn statistics s.t. the observed statistic is not an OOD sample, we can be robust!

# Learning robust statistics for SBI

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

# Learning robust statistics for SBI

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

- We want $\mathcal{D}$ to be outlier-robust. Hence, maximum mean discrepancy (MMD)
- Regularizer $\lambda$: encodes trade-off between accuracy and robustness

When learning inference and summary network jointly (as in NPE):

$$\hat{\mathcal{L}}(\phi, \psi) = -\frac{1}{m} \sum_{i=1}^{m} \log q_{h_\phi(\eta_\psi(\mathbf{x}_{1:n,i}))}(\theta_i) + \lambda \text{MMD}_k^2 \left[ \{\eta_\psi(\mathbf{x}_{1:n,i})\}_{i=1}^{l}, \eta_\psi(\mathbf{y}_{1:n}) \right]$$

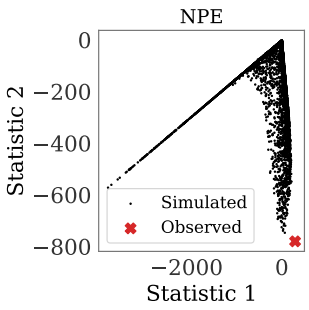
When learning statistics using an autoencoder (for ABC or other SBI methods):

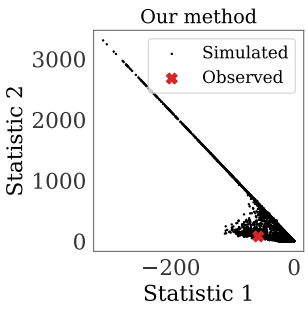
$$\hat{\mathcal{L}}(\psi, \psi_d) = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_{1:n,i} - \tilde{\eta}_{\psi_d}(\eta_\psi(\mathbf{x}_{1:n,i})))^2 + \lambda \text{MMD}_k^2 \left[ \{\eta_\psi(\mathbf{x}_{1:n,i})\}_{i=1}^{l}, \eta_\psi(\mathbf{y}_{1:n}) \right]$$

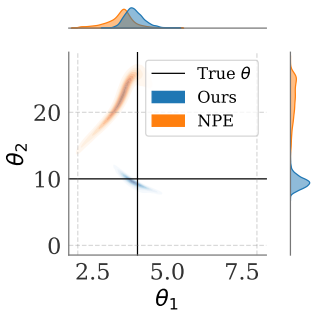
# Results

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

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

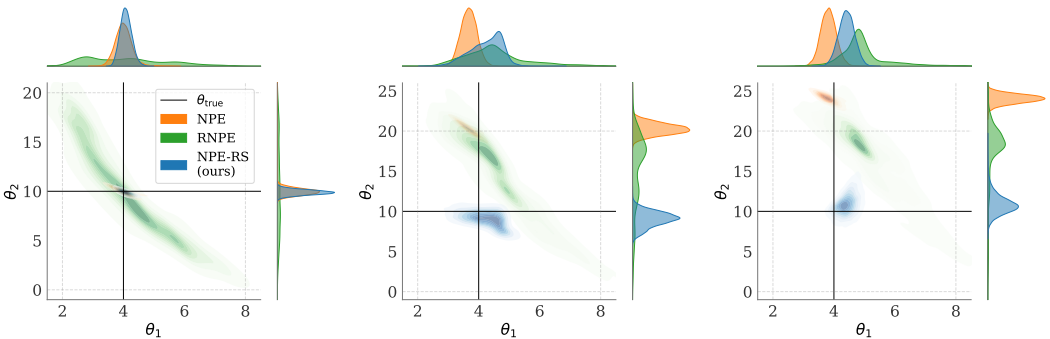


# Results

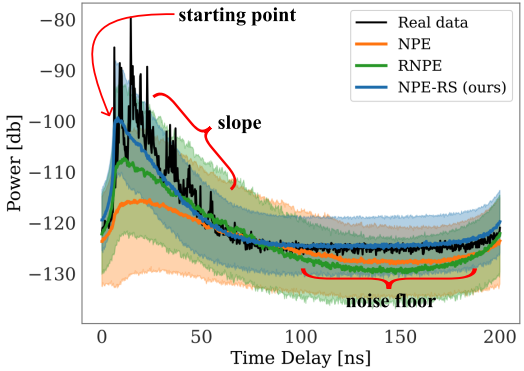
- **Ricker model:** 2 parameters
- **Inference method:** Neural posterior estimation (NPE)
- $\epsilon$**-contamination model:** $\mathbb{Q} = (1-\epsilon)\mathbb{P}_{\theta_{\text{true}}} + \epsilon\mathbb{P}_{\theta_c}$




(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.
- **Limitation:** NPEs are not amortized anymore.

Akeret, J., Refregier, A., Amara, A., Seehars, S., and Hasner, C. (2015).
Approximate Bayesian computation for forward modeling in cosmology.
Journal of Cosmology and Astroparticle Physics, 2015(08):043–043.

Beaumont, M. A. (2010).
Approximate Bayesian computation in evolution and ecology.
Annual Review of Ecology, Evolution, and Systematics, 41(1):379–406.

Bharti, A., Briol, F.-X., and Pedersen, T. (2022).
A general method for calibrating stochastic radio channel models with kernels.
IEEE Transactions on Antennas and Propagation, 70(6):3986–4001.

Dyer, J., Cannon, P. W., and Schmon, S. M. (2022).
Amortised likelihood-free inference for expensive time-series simulators with signatured ratio estimation.
In International Conference on Artificial Intelligence and Statistics, volume 151, pages 11131–11144.

Kypraios, T., Neal, P., and Prangle, D. (2017).

A tutorial introduction to bayesian inference for stochastic epidemic models using approximate bayesian computation.
Mathematical Biosciences, 287:42–53.


Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999).
Population growth of human y chromosomes: a study of y chromosome microsatellites.
Molecular Biology and Evolution, 16(12):1791–1798.