Data assimilation as simulation-based inference

Andry Gérôme

University of Liège

Master's thesis presentation (2022-2023)

Outline

Outline

Big picture

- We study dynamical systems characterized by their varying state x_t.
- We do not have direct access to those states. We can observe them through an observation process $y_t \sim \mathcal{O}(x_t)$.
- We study models of physical systems defined by a set of ODEs that characterizes the transition model $x_{t+1} \sim \mathcal{M}(x_t)$.

Formulation

In our setup

- $\mathcal{M}(.)$ is deterministic
- $\mathcal{O}(.)$ is linear w.r.t. x_t and Gaussian s.t. $y_t \sim \mathcal{N}(Ax_t, \Sigma_y)$

Unlike classical point estimation, we target the full posterior distribution $p(x \mid y_{t-T:t}, t)$ to incorporate uncertainty in the inference process.

Outline

Classical SBI vs Data assimilation

$$\boxed{p(\theta \mid x)}$$

$$\frac{p(x \mid \theta)p(\theta)}{p(x)}$$

- x is an observation
- θ is a parameter or a variable of interest

$$\boxed{p(x \mid y_{t-T:t}, t)}$$

$$\frac{p(y_{t-T:t} \mid x, t)p(x \mid t)}{p(y_{t-T:t} \mid t)}$$

- x is a state
- y is an observation
- t is the time index

Challenges

$$p(x \mid y_{t-T:t}, t) = \frac{p(y_{t-T:t} \mid x, t) p(x \mid t)}{p(y_{t-T:t} \mid t)}$$

- Time-varying posterior
- Scale of the problem
- Need proper evaluation of the estimator

Outline

Estimation methods

$$p(x \mid y_{t-T:t}, t) = \frac{p(y_{t-T:t} \mid x, t)p(x \mid t)}{p(y_{t-T:t} \mid t)}$$

Estimation methods

$$p(x \mid y_{t-T:t}, t) = \boxed{\frac{p(y_{t-T:t} \mid x, t)}{p(y_{t-T:t} \mid t)}}\, p(x \mid t)$$

Linked to the likelihood-to-evidence ratio

- **Direct posterior estimation**
- **Neural ratio estimation**

Estimation methods

$$\boxed{p(x \mid y_{t-T:t}, t)} = \boxed{\frac{p(y_{t-T:t} \mid x, t)}{p(y_{t-T:t} \mid t)}} \, p(x \mid t)$$

Density estimator as posterior surrogate

- **Direct posterior estimation**
- **Neural ratio estimation**
- **Neural posterior estimation**

Estimation methods

$$\boxed{x_t \sim} \; \boxed{p(x \mid y_{t-T:t}, t)} = \boxed{\frac{p(y_{t-T:t} \mid x, t)}{p(y_{t-T:t} \mid t)}} \; p(x \mid t)$$

Use SDE to reconstruct samples by estimating
$\nabla_{x(\tau)} \log p(x(\tau) \mid y_{t-T:t}, t)$

- **Direct posterior estimation**
- **Neural ratio estimation**
- **Neural posterior estimation**
- **Posterior score estimation**
- **Composed score estimation**

Outline

Simulators

Chaotic systems

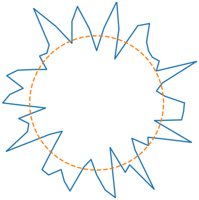


Figure 1: 1D - Lorenz96

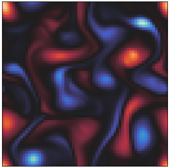


Figure 2: 2D - Turbulent flows

Used as a toy problem to benchmark different estimators. We gradually increase the problem scale from 8 to 256 nodes.

Used to push the scaling of our methods at much higher dimensions. This system is more physically realistic. We scale to 2048 nodes.

Comparison of considered techniques

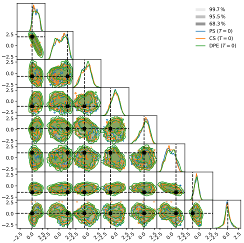
	Comparison	Density	Scale	Time	Interpretation
Loss					
Corner					
Qualitative					
Discriminator					

Comparison of considered techniques

	2D
Size N	**2048**
MAF ($T = 0$)	1095.204 ± 53.896
MAF ($T = 9$)	1184.298 ± 29.542
PS ($T = 0$)	0.044 ± 0.002
PS ($T = 9$)	0.042 ± 0.002
CS ($T = 0$)	0.050 ± 0.002

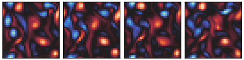
	Comparison	Density	Scale	Time	Interpretation
Loss	×	×	✓	✓	×
Corner					
Qualitative					
Discriminator					

Comparison of considered techniques



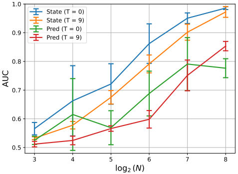
	Comparison	Density	Scale	Time	Interpretation
Loss	×	×	✓	✓	×
Corner	✓	✓	×	×	✓
Qualitative					
Discriminator					

Comparison of considered techniques



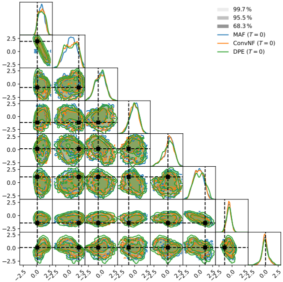
	Comparison	Density	Scale	Time	Interpretation
Loss	×	×	✓	✓	×
Corner	✓	✓	×	×	✓
Qualitative	✓	×	✓	×	✓
Discriminator					

Comparison of considered techniques

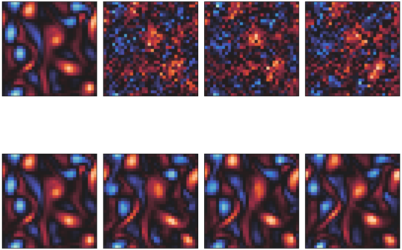


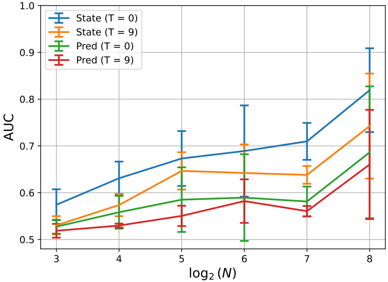
	Comparison	Density	Scale	Time	Interpretation
Loss	×	×	✓	✓	×
Corner	✓	✓	×	×	✓
Qualitative	✓	×	✓	×	✓
Discriminator	✓	×	✓	✓	×

Results

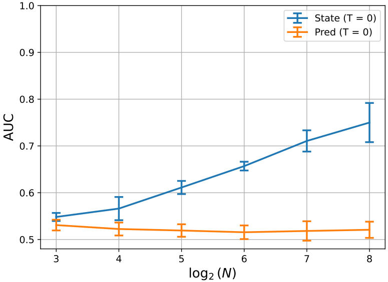


99.7%
95.5%
68.3%
MAF ($T = 0$)
ConvNF ($T = 0$)
DPE ($T = 0$)





State (T = 0)
State (T = 9)
Pred (T = 0)
Pred (T = 9)
AUC
$\log_2(N)$



State (T = 0)
Pred (T = 0)
AUC
$\log_2(N)$

Outline

Introduction

Adapting the SBI framework

Posterior estimation

Experiments

Conclusion and future work

Discussion

- We have to incorporate as much information as we have access to in our estimators.
- Convolutional architectures are preferred for good scaling. They are relevant regarding the data structure.
- Score-based models are promising despite they have certain defects.
- Convolutional flows must be further tested.
- It is of interest to study the impact of the time embedding.
- How can we adapt those methods to real-world scenarios with potential misspecified models ?

Any questions ?