ICML | 2022

How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative

Ahmed M. Alaa, Boris van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar

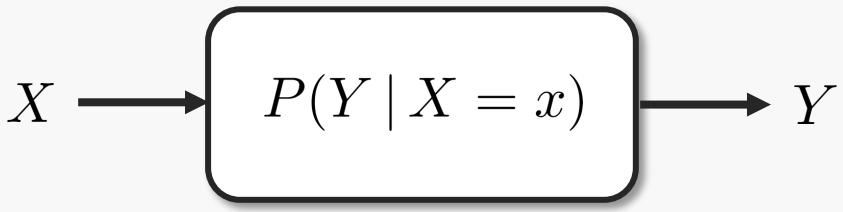
UNIVERSITY OF CAMBRIDGE

UCLA

Evaluating generative models

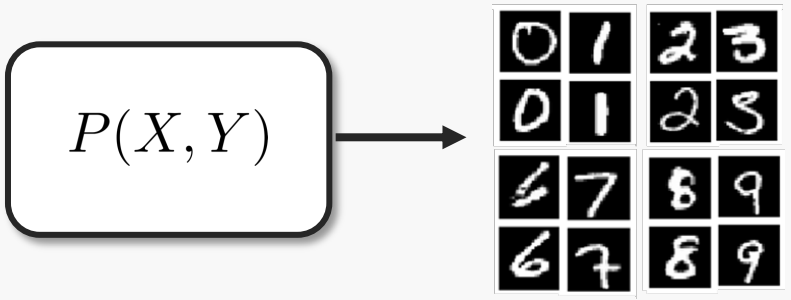
- Synthetic data can be sampled from **generative models** of $P(X, Y)$
- **How do we know if the synthetic data is of a high quality? What does "quality" mean?**

Discriminative models



- Validation against ground-truth labels

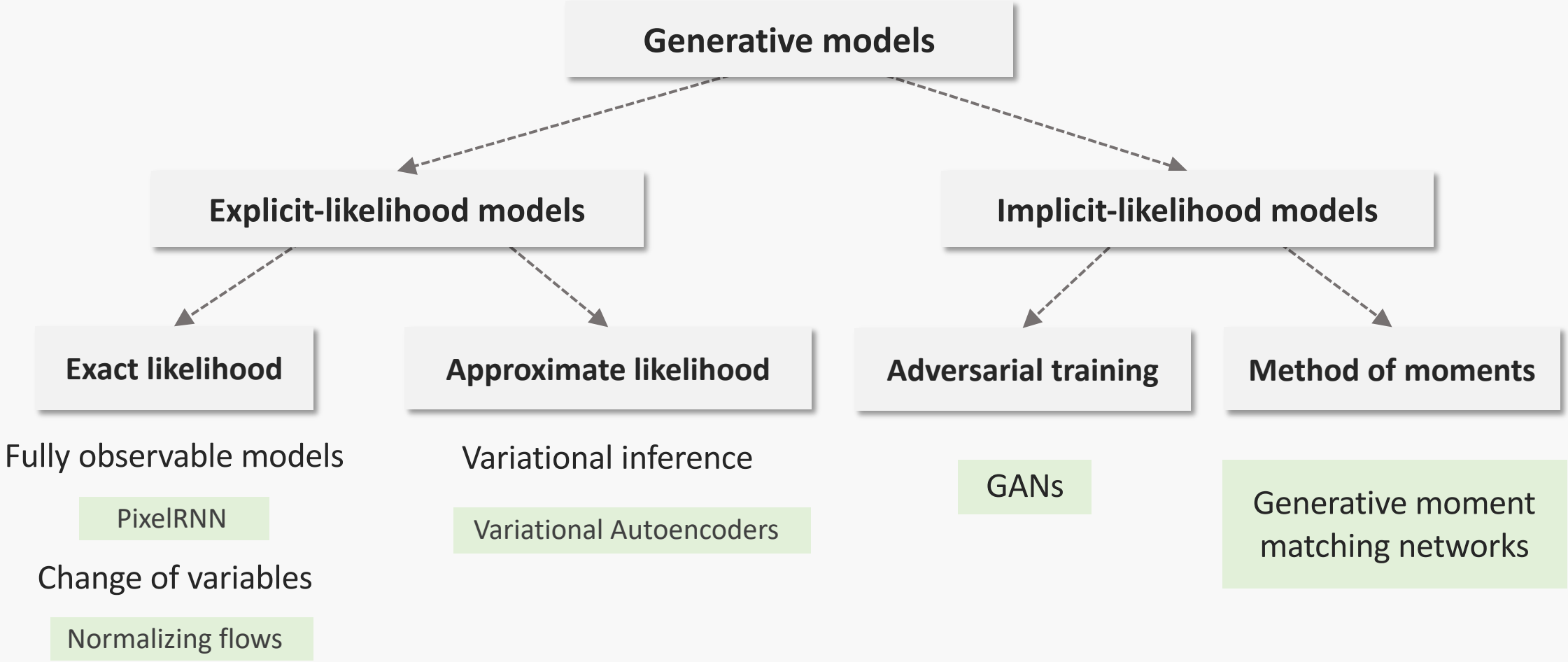
Generative models



- No ground-truth

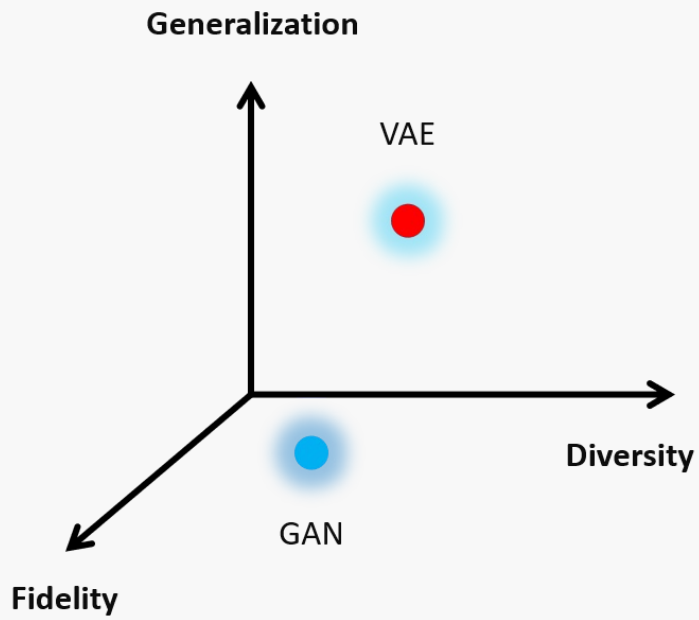
Not all models have explicit likelihoods!

- **Our goal:** A model- and domain-agnostic evaluation metric for generative models



A three-dimensional sample-level metric

- A model's performance can be viewed as a point in a 3D space…
  - **Fidelity:** How "good" the synthetic samples are?
  - **Diversity:** How much of the real data is covered?
  - **Generalization:** How often does the model copy training data?
- Each sample is evaluated w.r.t each of the above criteria
- Model performance = average performance over samples



Evaluating *Fidelity* through *α-Precision*

- Builds on the precision-recall analysis framework proposed in **[Sajjadi et al, 2018]**

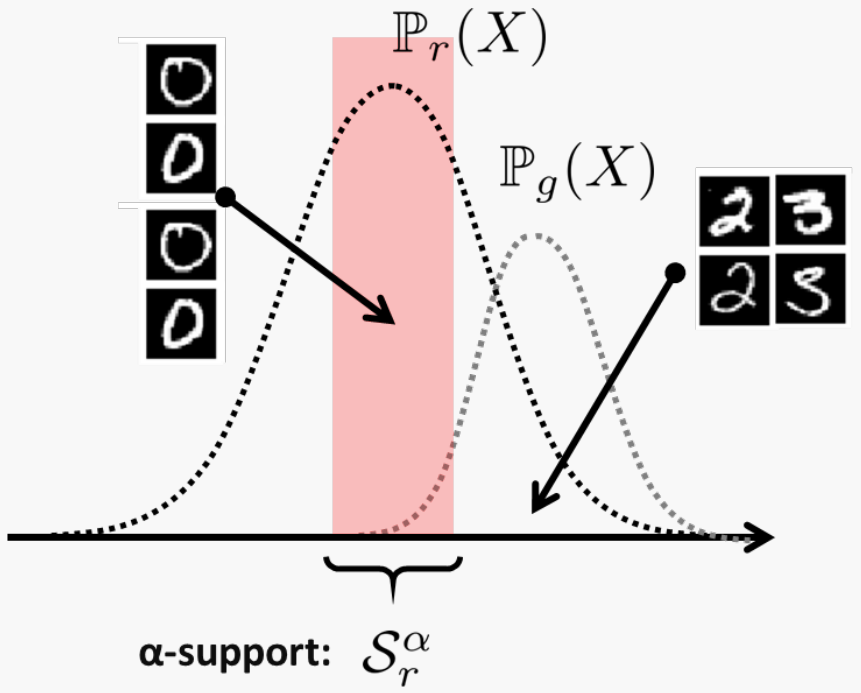
$\mathcal{S}_r^\alpha$ = Minimum-volume α-support of real data distribution

α-Precision

$$P_\alpha = \mathbb{P}(X_g \in \mathcal{S}_r^\alpha)$$

The fraction of synthetic Samples that resemble the α most "typical" samples in real data

- **α-Precision** measures sample *fidelity*.



Evaluating *Diversity* through β-*Recall*

- Builds on the precision-recall analysis framework proposed by

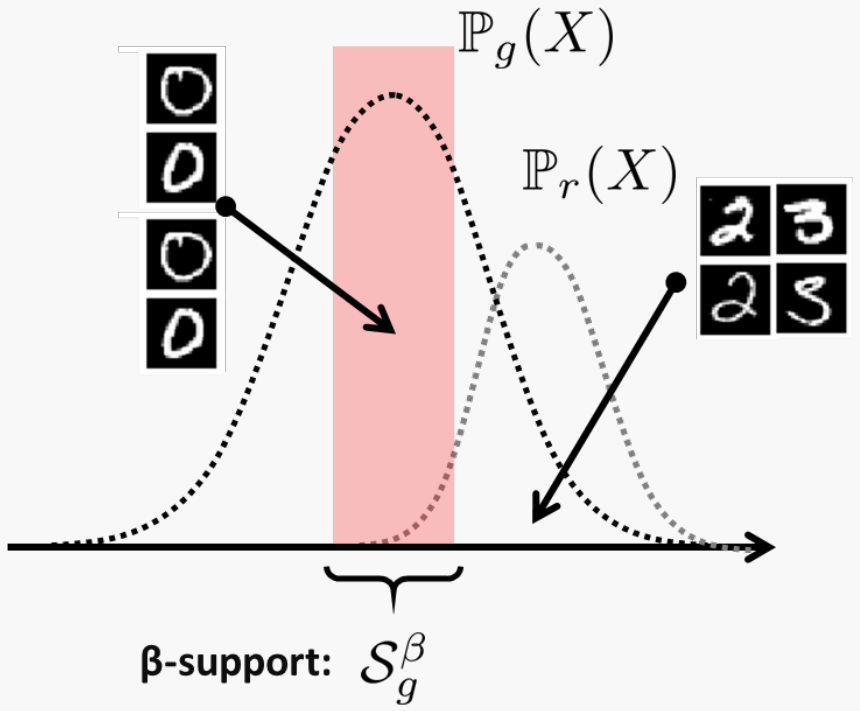
$\mathcal{S}_g^\beta$ = Minimum-volume β-support of synthetic data distribution

β-Precision

$$R_\beta = \mathbb{P}(X_r \in \mathcal{S}_g^\beta)$$

The fraction of real samples covered by the β most typical synthetic samples

■ **β-Recall** measures sample *diversity*.



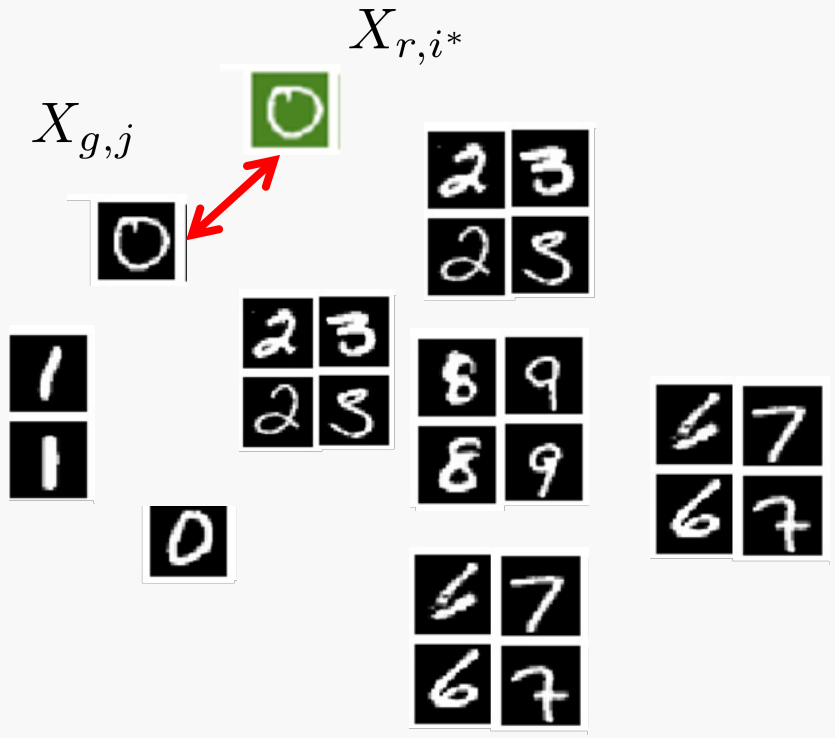
Evaluating *Generalization* through the *Authenticity* metric

- We can generate diverse and high-fidelity data by re-sampling real data (memorization)
- How to test if a model is truly synthesizing new samples?

Authenticity metric

$$\mathbb{P}(d(X_{g,j}, \mathcal{D}_{real}) < d(X_{r,i^*}, \mathcal{D}_{real}/\{X_{r,i^*}\}))$$

How often does the model generate samples that are closer to real data than the closest real sample?



Post-hoc model auditing

- Remove samples that are memorized or imprecise