A brief introduction to deep generative models

Clayton W. Seitz

January 24, 2022

Outline

Generative Models

References

The logic of generative modeling

Say we have a set of variables $\mathbf{x} = (x_1, x_2, ..., x_n)$ which might have some statistical dependence

The variable $\mathbf{x}$ might be an amino acid sequence, gene expression data, microscopy image, etc.

- ▶ Often we are handed a batch of empirical samples $\{\mathbf{x}_i\}_{i=1}^{N}$
- ▶ We want to know the generating distribution $p(\mathbf{x})$

In supervised generative learning, we try to explicity learn the joint distribution $p(\mathbf{x}) = \prod_{i=1}^{N-1} p(x_i | x_{i+1:N}) p(x_N)$, which is generally more difficult than discriminative learning.

Perks of generative modeling

- ▶ Fitting complete multivariate distributions $p(\mathbf{x})$ goes beyond correlation-based or clustering approaches
- ▶ Correlations cannot discover partial correlation in the context of other neighbors
- ▶ Fitting $p(\mathbf{x})$ permits sampling based inference

Why generative modeling is difficult

When describing a distribution over multiple variables, we may not know the proper normalization Z. That is,

$$p(\mathbf{x}) = \frac{1}{Z}\tilde{p}(\mathbf{x})$$

This very important situation arises in several contexts:

1. In Bayesian inference where $p(x_1|x_2) = p(x_2|x_1)p(x_1)/p(x_2)$ is intractable due to $Z = p(x_2) = \int p(x_2|x_1)p(x_1)dx_1$. This integral can be very difficult or impossible to compute.

2. In models from statistical physics, e.g. the Ising model, we only know $\tilde{p}(\mathbf{x}) = e^{-H(\mathbf{x})}$ where $H(\mathbf{x})$ is the Hamiltonian

Variational autoencoders (VAEs)
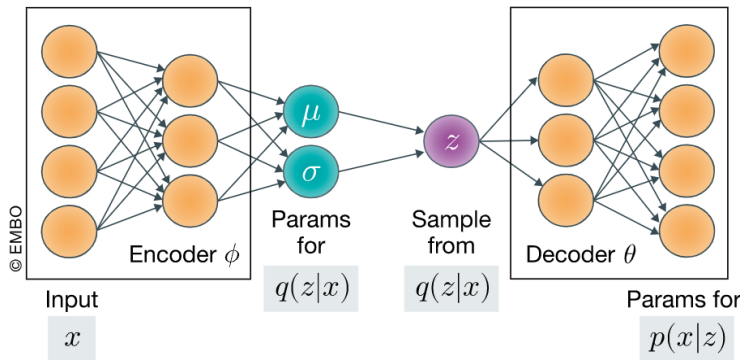
A variational solution to generative modeling

Figure 1: **Variational autoencoder architecture** Taken from Lopez 2020 in EMBO

Bayesian inference

The variable **x** has a latent representation or code **z**. We often say that **z** is the *causal source* of **x**. Ultimately, we would like to know the distribution $P_\Phi(\mathbf{x})$

$$P_\phi(\mathbf{x}) = \frac{P_\phi(\mathbf{x}|\mathbf{z})P_\phi(\mathbf{z})}{P_\psi(\mathbf{z}|\mathbf{x})}$$

in order to find the model parameters that maximize the likelihood of the observed data:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} - \log P_\Phi(\mathbf{x})$$

but we generally do not know $P_\psi(\mathbf{z}|\mathbf{x})$ due to the intractable integral $Z = \int P_\phi(\mathbf{x}|\mathbf{z})P_\phi(\mathbf{z})d\mathbf{z}$ (see slide 5)

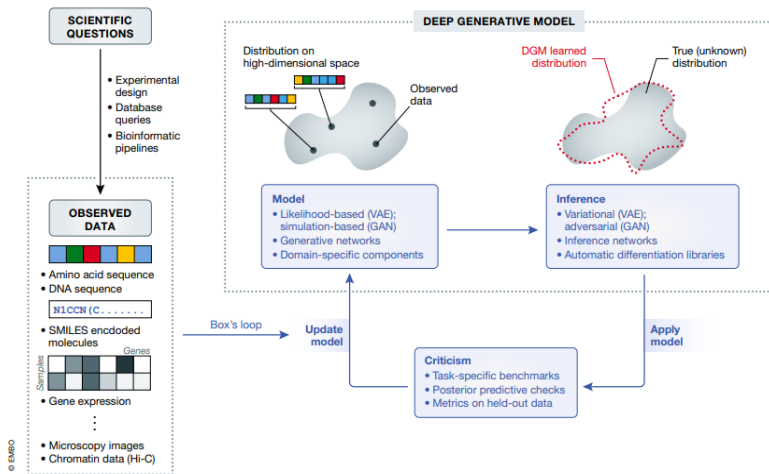# Computing the evidence

We can rewrite the evidence as

$$P_\phi(\mathbf{x}) = \int P_\phi(\mathbf{z}) P_\phi(\mathbf{x}|\mathbf{z}) d\mathbf{z}$$

$$= \int P_\phi(\mathbf{z}) P_\phi(\mathbf{x}|\mathbf{z}) \frac{P_\phi(\mathbf{z}|\mathbf{x})}{P_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{z} \sim P_\phi(\mathbf{z}|\mathbf{x})} \frac{P_\phi(\mathbf{z}) P_\phi(\mathbf{x}|\mathbf{z})}{P_\phi(\mathbf{z}|\mathbf{x})}$$

where $P_\phi(\mathbf{z}|\mathbf{x})$ is our model "encoder"

The evidence lower bound

$$\log P_\phi(\mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim P_\phi(\mathbf{z}|\mathbf{x})} \log \frac{P_\phi(\mathbf{z}) P_\phi(\mathbf{x}|\mathbf{z})}{P_\phi(\mathbf{z}|\mathbf{x})}$$

Applying deep generative models to biological data

A recent VAE for transcriptomics data: scV1

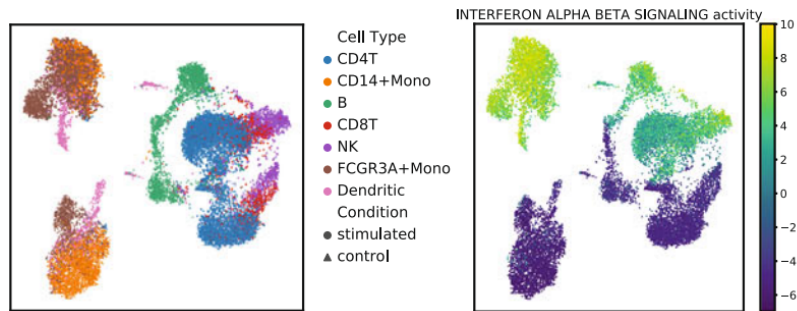# Recent improvements of interpretability

Figure 2: Phenotype segregation using a VAE on single-cell transcriptomics data. Taken from Seninge et al.

References I