SCVAE: Variational Auto Encoders for single cell gene expression data

Paul Caucheteux, Marc Chevrière, Marion Hubler
Master MVA, ENS Paris Saclay

We worked on the articles : **Deep generative modeling for single-cell transcriptomics** written by Romain Lopez and al. [1] and on **Variational auto-encoders for single-cell gene expression data** written by Christopher H Grønbechet and al. [2].

Article [1] focuses on the performances, on various tasks, of a new VAE model : **SCVI**. Article [2] introduced a Gaussian Mixture prior on the latent space, to try and improve the model on the specific task of clustering in the latent space (**GMVAE**).

Our goal was to evaluate whether the GMVAE improves clustering performance over standard VAEs, determine if GMVAE enhances data representation, and compare the impact of different reconstruction likelihoods on performance.

Single cell RNA sequencing dataset

Single cell RNA sequencing (scRNA-seq) is a technology that allows the measurement of gene expressions at **individual cell level**.
scRNA-seq allows a deep understanding of cellular diversity.

Example of a scRNA-seq :

		Gene expression count					
		g_1	g_2	...	g_g	...	g_G
Cell	c_1	x_{11} : raw count data	x_{12}	...	x_{1g}	...	x_{1G}
	c_2	x_{21}	x_{22}	...	x_{2g}	...	x_{2G}
	...	...	...	...	...	...	...
	c_n	x_{n1}	x_{n2}	...	x_{ng}	...	x_{nG}
	...	...	...	...	...	...	...
	c_N	x_{N1}	x_{N2}	...	x_{Ng}	...	x_{NG}

Tasks we examined:

- **Imputation:** Addresses missing data due to **dropouts**, occurring when genes are not detected by the RNA sequencing due to: **transcriptional noise** (transcription pauses when cells divide) and the **sensitivity issues** (low RNA amounts are variably detection and cause gaps in the datasets).
- **Visualization and Clustering:** Involves visualizing and clustering the data (e.g., in a latent space of the VAE) to uncover patterns and relationships, often with the goal of identifying distinct cell types within the dataset.
- **Differential Expression Analysis:** Identifies genes with significantly different expression between cell groups, helping to understand which genes are overexpressed or underexpressed in a given context.

Models: Simple VAE and GMVAE

We use a simple VAE as a baseline to test the performance of the GMVAE.
We recall that we train the simple VAE by maximizing the ELBO :

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right),$$

The GMVAE is trained by maximizing the ELBO :

$$\begin{aligned}\mathcal{L}(\theta, \phi\,; x) = &\mathbb{E}_{y\sim q_\phi(y|\mathbf{x})}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x},y)}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, y) \,\|\, p_\theta(\mathbf{z} \mid y)\big)\right] \\ &- \mathrm{KL}\big(q_\phi(y \mid \mathbf{x}) \,\|\, p_\theta(y)\big).\end{aligned}$$

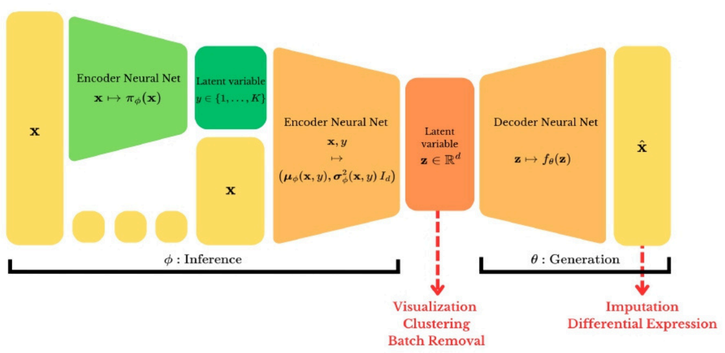
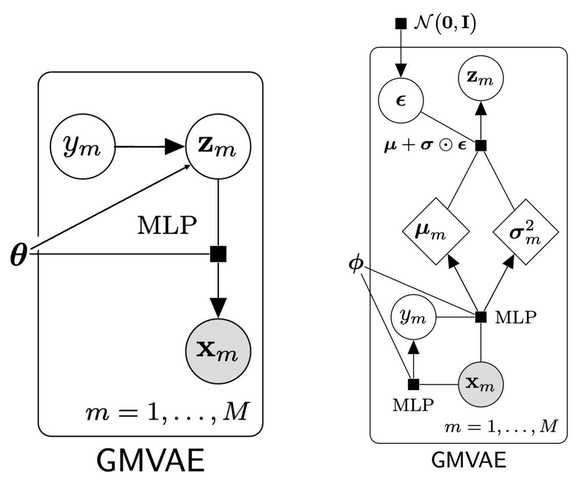


Figure 1: GMVAE Architecture

We detail below the generative and inference process for the GMVAE and the law of the different variables.

$$\begin{aligned}\mathbf{z} &\sim \mathcal{N}(\boldsymbol{\mu}_\theta(y), \sigma^2_\theta(y)\,\mathbb{I}) \\ q_\phi(\mathbf{z}|\mathbf{x}, y) &\sim \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{x}, y), \boldsymbol{\sigma}^2_\phi(\mathbf{x}, y)\right) \\ q_\phi(y|\mathbf{x}) &\sim \text{Categorical}\left(\boldsymbol{\pi}_\phi(\mathbf{x})\right) \\ p_\theta(\mathbf{x}|\mathbf{z}) &\sim \text{Likelihood}\left(f_\theta(\mathbf{z})\right)\end{aligned}$$



(a) Generative Process (b) Inference Process

Choice of the Likelihood

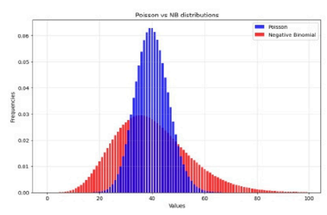


Figure 2: P vs NB

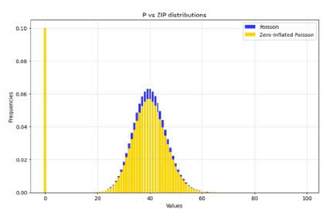


Figure 3: P vs ZIP

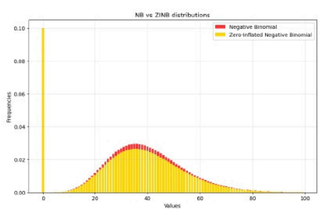


Figure 4: NB vs ZINB

- **Poisson (P):** Assumes the mean equals the variance, a simplistic assumption often unrealistic for single-cell data.
- **Negative Binomial (NB):** Captures over-dispersion, allows variance to exceed the mean.
- **Zero-Inflated Distribution (ZINB and ZIP):** Adds binary process for excess zeros to the original distributions.

Visualization in the latent space

To visualize our latent representation, we take the expectation of the variational posterior distribution. We then project the results on $\mathbb{R}^2$ using UMAP. Figure 5 displays results for GMVAE, VAE and for raw data on which we perform PCA.

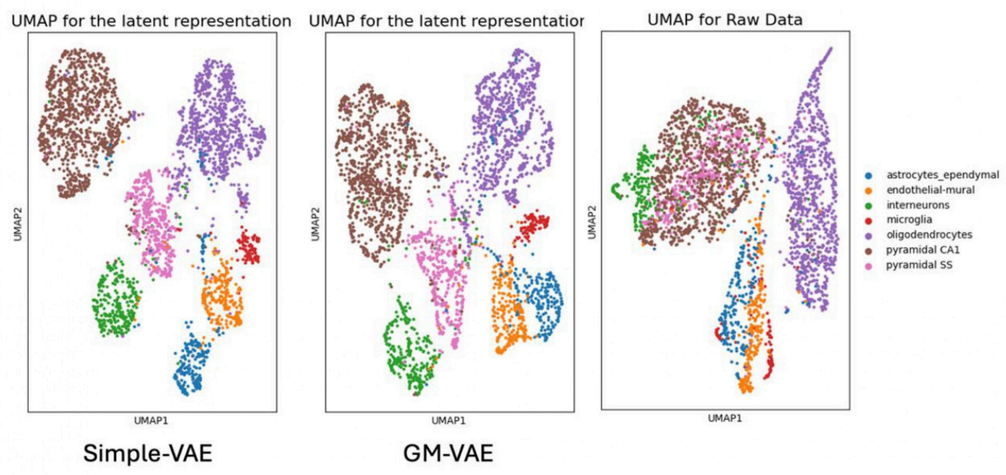


Figure 5: Visualization in the latent space for VAE, GMVAE and for raw data

Clustering and Imputation

To perform the **clustering**: for the GMVAE we simply take the `argmax` of the categorical distribution ; for the VAE we perform a K-Means on the latent space. Then we evaluate the clustering for some metrics (Adjusted Rand Index (ARI), Normalized Mutual Information (NMI))
To evaluate the **imputation**: we corrupt the data and try to reconstruct the original data by passing it into the model. We then compute the distance between the original and imputed only on the corrupted entries, and condense these distances into metric by either taking their median or mean.

Likelihoods	Simple VAE		GMVAE	
	ARI	NMI	ARI	NMI
Poisson	**0.6327**	**0.6972**	0.4026	0.5207
Zero-inflated Poisson	**0.6511**	**0.7054**	0.3684	0.4817
Negative Binomial	**0.5875**	**0.6338**	0.3787	0.4968
Zero-inflated Negative Binomial	**0.6803**	**0.7074**	0.3751	0.4959

(a) Imputation Results.

Likelihoods	Simple VAE		GMVAE	
	Median	Mean	Median	Mean
Poisson	**0.8889**	**1.819**	0.9044	1.900
Zero-inflated Poisson	**0.9345**	**2.298**	0.9527	2.393
Negative Binomial	**2.768**	**4.212**	2.779	4.244
Zero-inflated Negative Binomial	**4.481**	**5.665**	5.283	6.649

(b) Clustering Results

The simple VAE outperforms the GMVAE on clustering tasks across all likelihoods, possibly due to a misalignment between the GMVAE's Gaussian mixture prior and the biological structure of the data. For imputation, non-zero-inflated likelihoods perform better, as zero-inflated models may misinterpret artificially introduced dropouts as true biological zeros, impacting performance.

Références

[1] Christopher Heje Grønbech et al. "scVAE: variational auto-encoders for single-cell gene expression data". In: *Bioinformatics* (2020).

[2] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature methods* (2018).