SAD2 - project 1

Stanisław Janik

December 2022

1 Task 1

1.1 a

Train dataset consists of 72208 observations each of 5000 variables, meanwhile test dataset consists of 18052 observations each of 5000 variables.

1.2 b & d

Histograms of raw and preprocessed dataset may be found in the Figure 1. Values on the X-axis were clipped to 10 and when showing data including zeroes logscale Y axis were used.

1.3 c

Max value in the raw dataset is lower than max value in preprocessed dataset, so log1p transformation was not used on it, since $\ln(x+1)$ is greater than x for all x greater than 0.

1.4 e

Data distribution does not look like a normal distribution. It seems to follow gamma distribution when zeroes are ommited, and like exponential distribution with zeroes taken into acconut.

The abundance of zeros means that most of the genes are not very active in most of the cells. This may be because different cells play different functions, i.e. red blood cell has different expressed genes than lymphatic cell due to their functions. There are many types of cells which means that we can find many expressed genes, however each of these genes won't be expressed in many cells.

1.5 f

This dataframe contains information about each sample, such as information about donor of the sample (age, BMI, blood type, race and more), cell type, batch, and more.

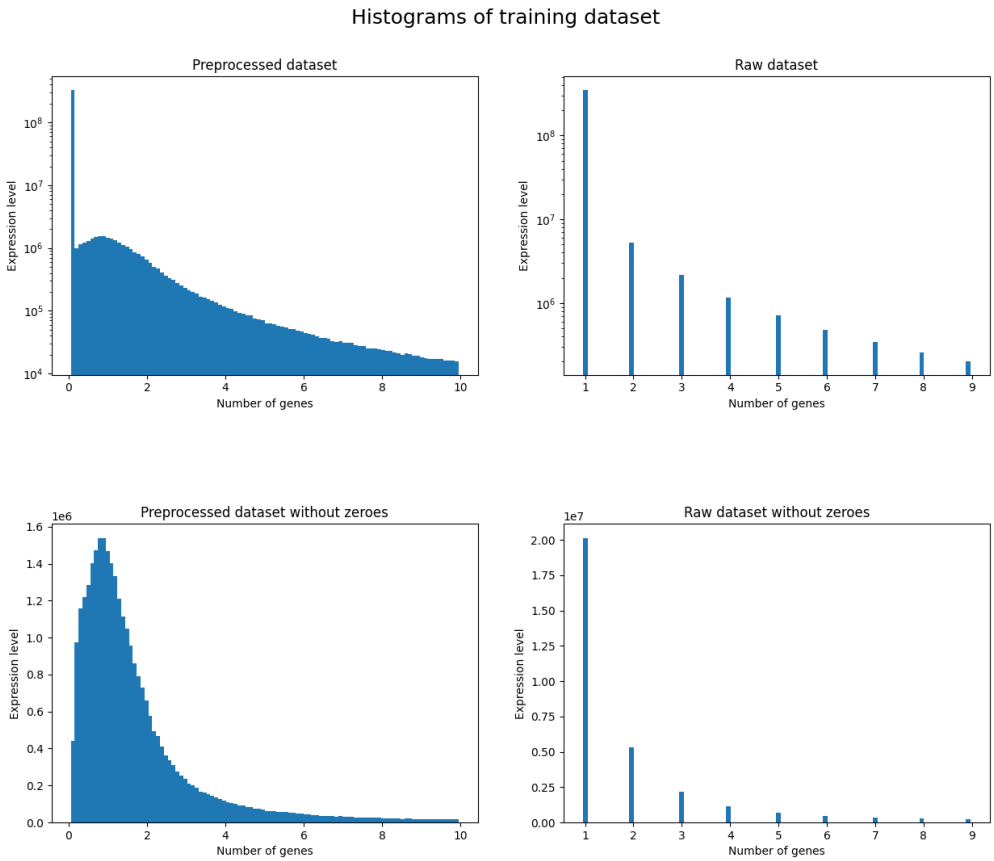

Figure 1: Histogram of data in the training dataset

Latent size	2	64	200
Loss on test set	4091509.553	2786982.16	2739841.37

Table 1: Comparison of loss and latent space

Number of unique patients is 9. Number of unique labs is 4. Number of unique cell types is 45.

2 Task 2

2.1 a

Model learns, however $-ELBO$ stays quite high. See Figure 2.

2.2 b

As we can see there is not much difference in loss between bigger sizes of latent space. However the increase is significant when we take very small latent space sizes e.g. 2. See Table 1.

For latent size 2 2 primary components explain 95% variance, for latent sizes 64 and 200 the number of components needed is 3.

Looking at plots coloured by cell type in figures 3, 4, 5 we may see that latent space size of 64 is somewhat worse clustered than the ones of sizes 2 and 200.

2.3 c

See Figures 3, 4, 5

2.4 d

For my final model I've chosen the model with hidden layers 2048, 1024, 256 and latent space size 64.

I've chosen preprocessed data, because it has continuous distribution, and the raw data doesn't.

Since comparing loss obtained by using different latent space sizes showed that it is quite similar no matter which size is chosen I decided to choose latent space of 64. Probably smaller latent sizes could be used, since we know from the PCA analysis that as little as 3 primary components are responsible for 95% in variance, however I opted for bigger size since I wanted as much variance remaining as possible.

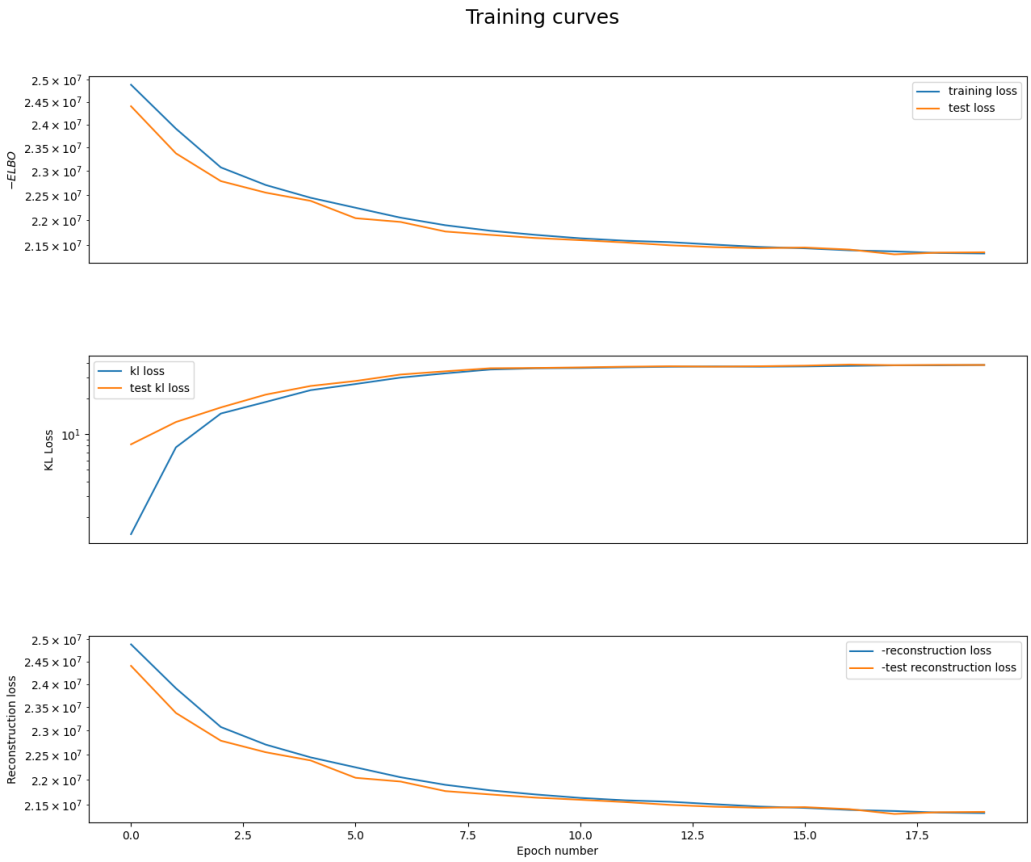

Figure 2: Training curves of VAE with break down to regularisation and reconstruction losses.

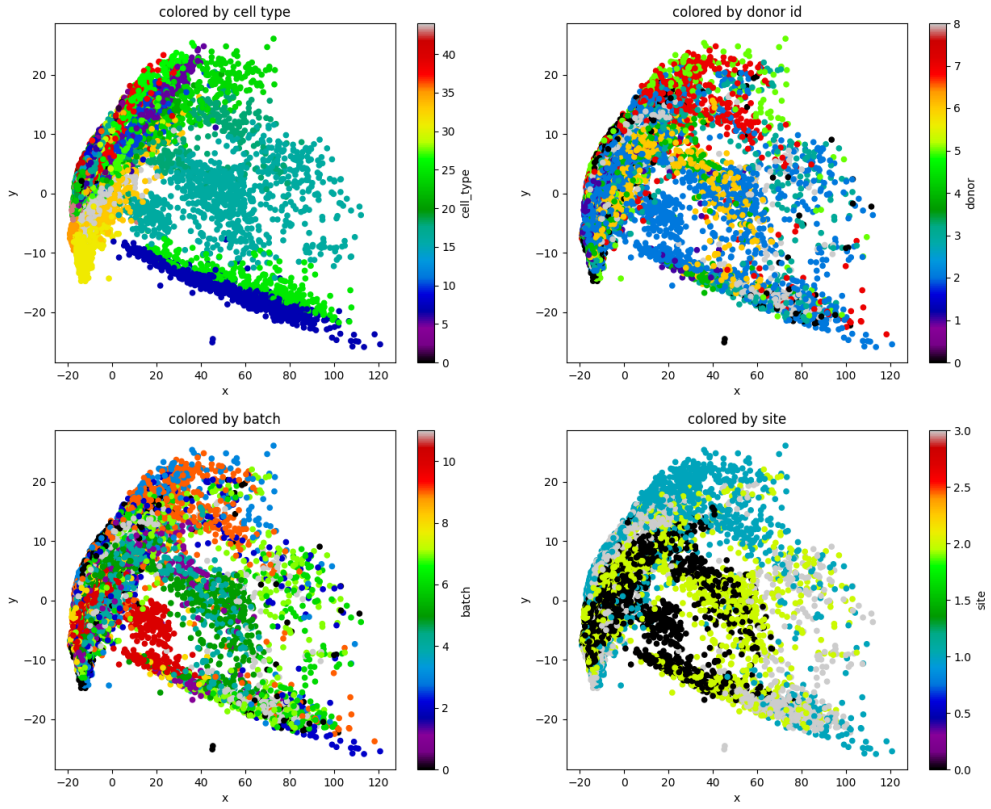

Figure 3: PCA projection of Gauss latent space of size 64 colored by information contained in *test.obs*

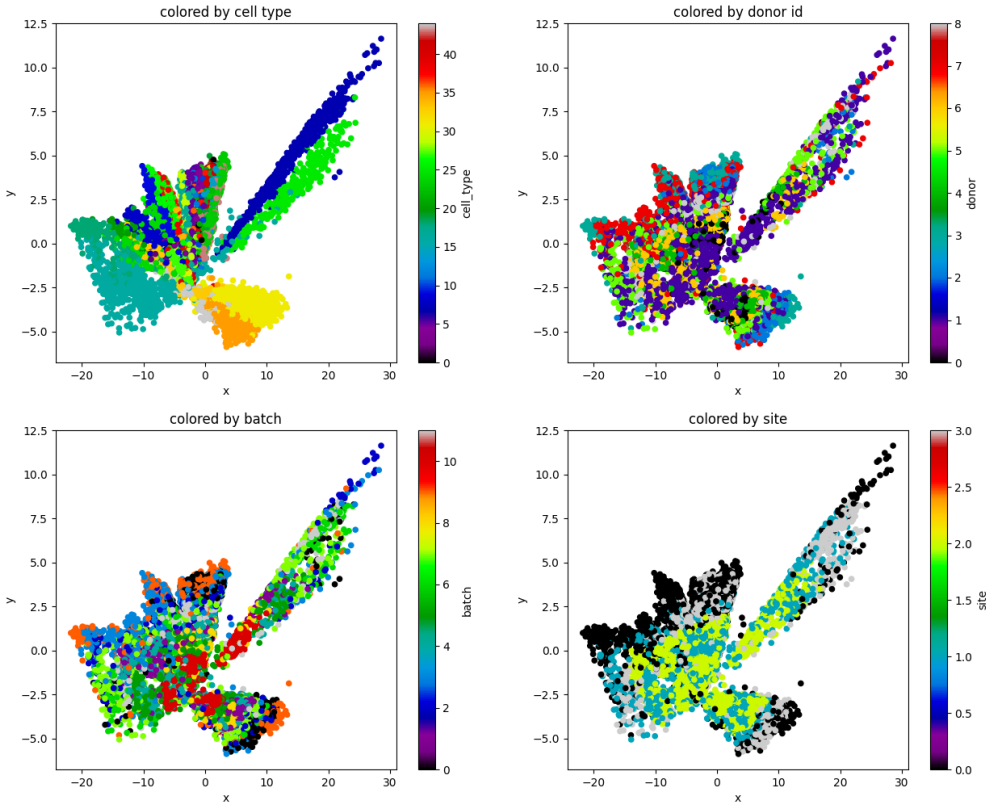

Figure 4: PCA projection of Gauss latent space of size 2 colored by information contained in *test.obs*

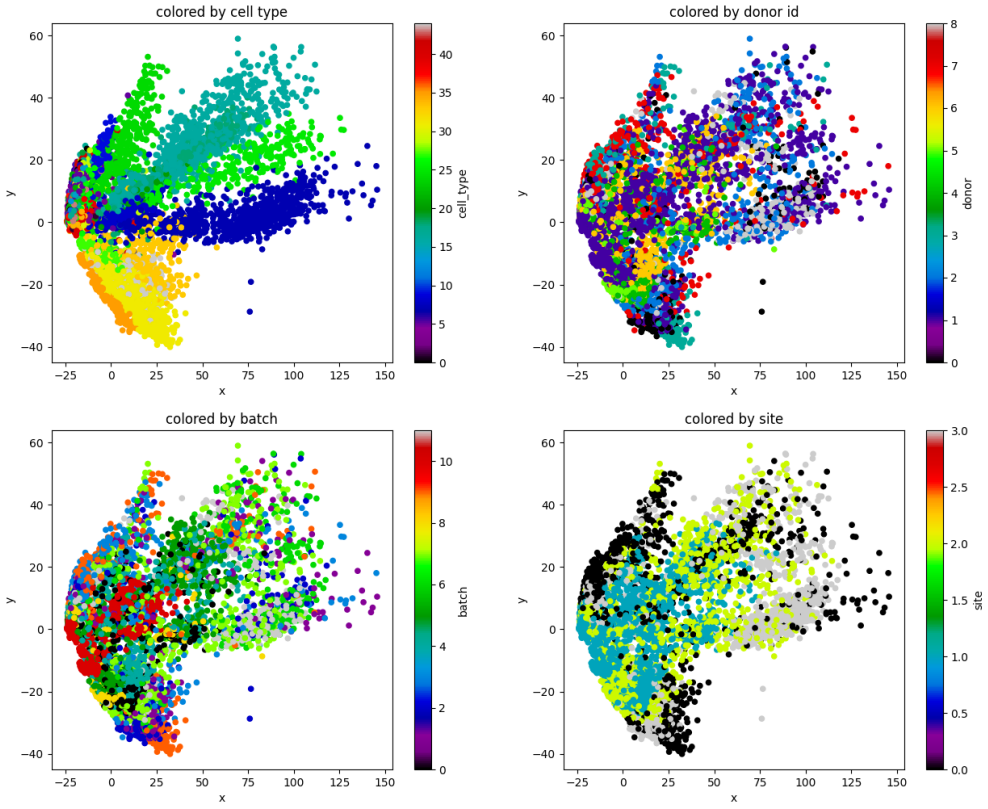

Figure 5: PCA projection of Gauss latent space of size 200 colored by information contained in *test.obs*

3 Task 3

3.1 a

I've chosen Exponential distribution to model the data. This is because It is a distribution that looks similarly to histogram of the preprocessed data. Probably better solution would be to choose some mixed distribution, for example mix of Gamma distributions, where we would try to estimate probability of each data point being one that comes from some Gamma distribution that models expressed genes or from one that models non expressed genes(this one would be more similar to Exponential distribution). However, I couldn't get VAE with such mixture distribution to work, so I tried implementing one with Exponential distribution.

3.2 b

I chose the same architecture for my model with Exponential encoder as I did for one with Gaussian encoder, since I wanted for PCA to be comparable even in the slightest. Because of that latent space size is 64.

As we can see in Figure 6 $-ELBO$ of this encoder is negative. Since sum log probs of continuous distributions may be positive so can be $-ELBO$. Releveant stackoverflow thread, and another one.

That said it seems Exponential decoder fits better since loss is substantially lower, however due to change in distribution some other metric should be used to assess which model is better.

3.3 c

Comparing plots colored by cell type in figrures 3 and 7 I don't see much difference in quality, PCA obtained from Exponential decoder may be little bit divided into clusters, however to assess it properly we would need some kind of metric, since it's hard to differentiate 45 different colours in an image.

4 Task 4

4.1 a

These plots may be found in figures 3, 4, 5 7.

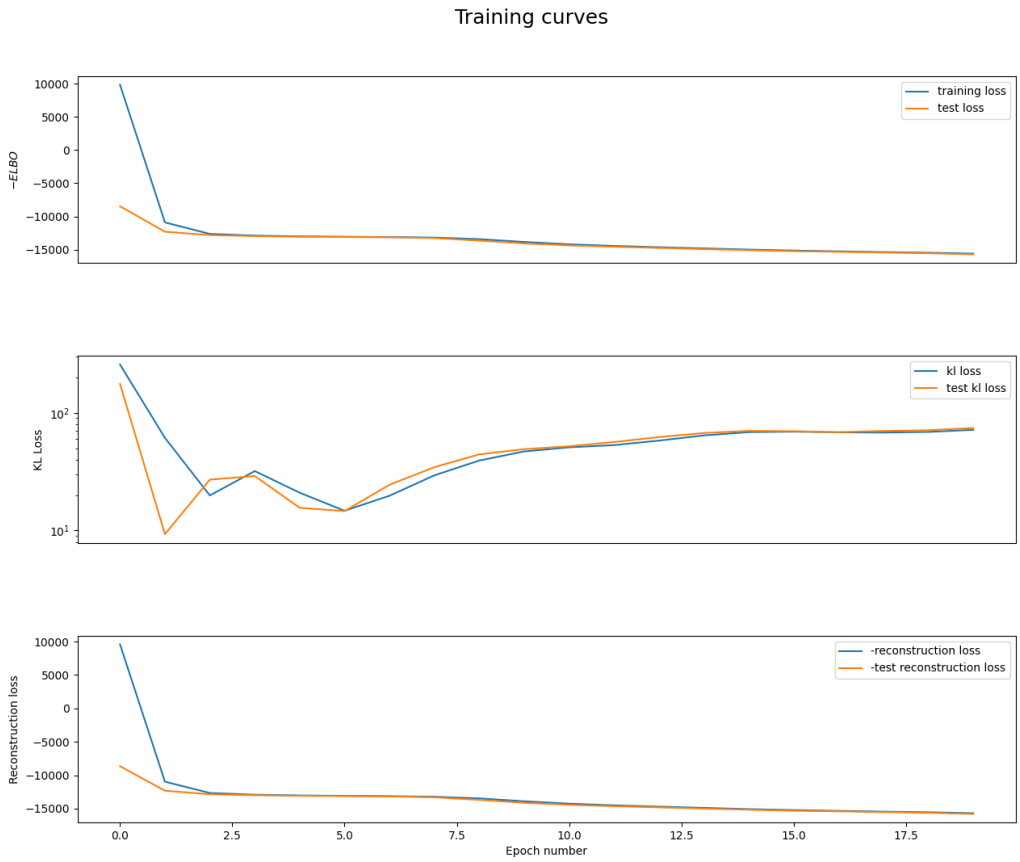

Figure 6: Loss plot for Exponential encoder

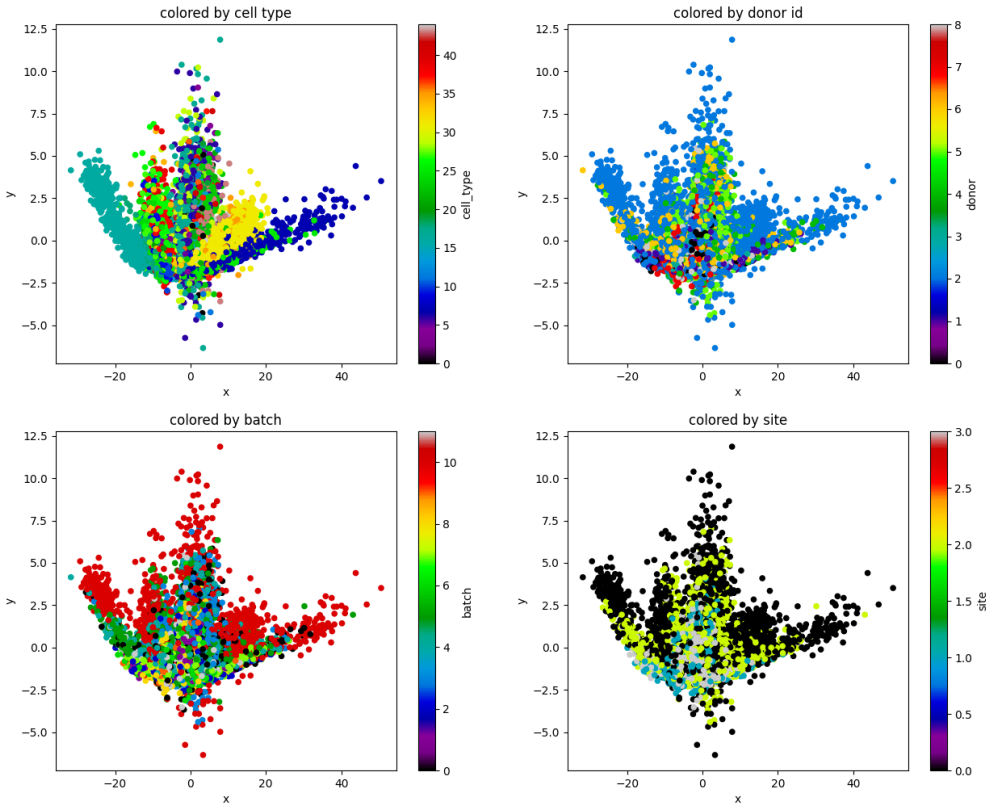

Figure 7: PCA of Exponential latent space