BPD Variational Autoencoder Quick Report

Jack Stanley (with the help of Justin Sing)
June 20, 2021

Data

Bronchopulmonary dysplasia (BPD) is a chronic lung disease present in preterm infants. These infants with BPD require long-term oxygen, and ultimately the disease results in significant mortality. There are many antenatal and postnatal factors that may harm infant lungs and result in BPD; clinically, prematurity and low birthweight are the most prominent indicators of possible development of BPD. Other clinical factors may also play a role in the development of BPD, and some of these are explored further below.

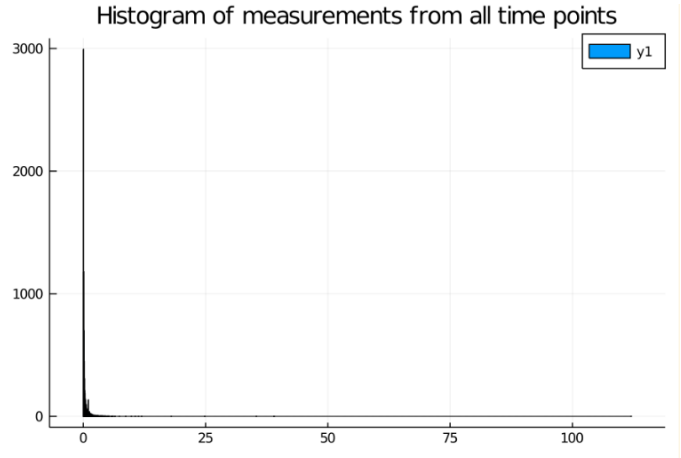
In our exploration, we are looking at metabolite abundances (normalized to creatinine and DSS) collected from the urine of infants at four different timepoints post-birth (Day 1 = A, Day 4 = B, Day 7 = C, Day 14 = D). In total, there are 405 samples with 73 different metabolites measured. Here is a further breakdown by time point:

- Time point A: 106 samples
- Time point B: 106 samples
- Time point C: 101 samples
- Time point D: 92 samples

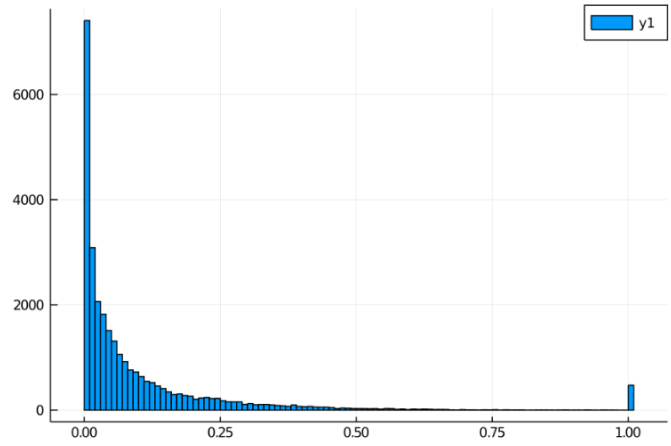
Note that there are a total of 121 unique samples, with 58 samples having BPD, and 63 control samples. Note also that there are 60 male samples and 61 female samples.

Unfortunately, only 62 samples out of the total 121 had readings for all four time points, and only 41 additional samples had readings at three time points. This lower number of samples may have exacerbated some of the inherent issues with our model, and this will be explored later.

Plotting a histogram of this data shows that it is not ideal to work with. Most metabolite measurements are quite low, while a few are very very high. As well, the most common measurement by far was 0, which gives the data this heavy right tail.



Now, the distribution of this data is not necessarily incompatible with the learning method (to be discussed in detail later) used, but the untransformed data runs incredibly slowly due to the quite variable orders of magnitude of some metabolite readings. To rectify this, I simply divided the measurements of every sample by the largest measurement for a particular metabolite. In this way, the range of values is from 0 to 1, but the metabolite ratios between samples are preserved. This works because the learning method used does not compare metabolites to each other, but compares an individual metabolite of one sample to that same metabolite from another sample. I did compare the results from the transformed and untransformed data, and they were exactly the same. The only difference is that the transformed data was able to run ~100x faster. Here's a distribution of the transformed data:



Note that a log transform also produces similar results, but is a bit slower than the transformation described above. Most reasonable transformations did not significantly affect the output.

In addition, there were a few clinical parameters that were thrown into this analysis including antenatal steroid use, gestation type, chorioamnionitis status, method of delivery, sepsis status, birth weight, sex, and gestational age.

Purpose

Ideally, this analysis would reveal some key metabolites (or an array of metabolites) that serve as early markers for BPD. Further, it should be clearly seen that the separation between healthy metabolite profiles and disease metabolite profiles increases as time increases. Each time point should serve to "update" the model as to the disease progression of a particular sample, and in this way the temporal aspect of the data could be fully leveraged.

Currently, physicians are able to predict with some precision which infants are at risk for BPD based entirely on clinical information, such as birth weight, gestational age, and sex. Ultimately, this analysis aims to identify metabolite markers that, taken in combination with clinical data, improve the accuracy of BPD predictions and thus allow for more urgent care to be given to infants.

Methods & Rationale

The first thing that comes to mind when thinking about analyzing this array of metabolite data is certainly PCA. We have a relatively large number of metabolites compared to the number of samples, so some sort of dimensionality reduction is needed. PCA produced decent separation of time points, but very little separation in terms of disease (see Justin Sing's work for more results and info regarding PCA). Thus, I aimed to implement a method that behaved similarly to PCA but captured more potential non-linear relationships between the samples and metabolites.

For this data, I settled on a machine learning method called a "variational autoencoder" or "VAE". Typically, VAEs have been used in image recognition and classification; the prototypical use case for a VAE is identifying and categorizing hand-written digits. Essentially the model takes in images and maps the information contained in each image to a low dimensional (in most cases 2D) latent space, then attempts to reconstruct the image from only the latent space information. The reconstruction is compared to the original input, and the parameters of the model (both input and output) are tweaked in order to get the closest match for all images. The "variational" is included in the name of this method because the latent space is actually sampled from instead of being fixed (any distribution can be used for the latent space, but a Gaussian distribution is common). Ideally this gives a smoother latent space and allows trends to be seen more clearly.

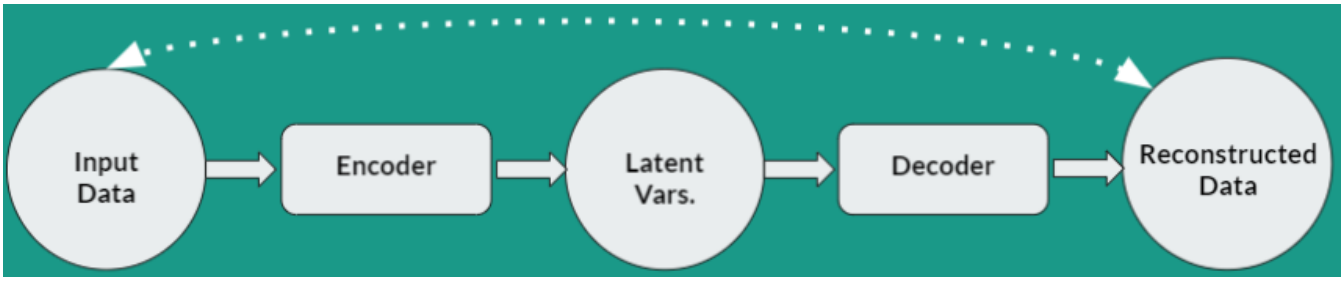
In terms of our metabolite data, each sample can be thought of as an image, and each metabolite reading for a given sample is akin to a specific pixel for each image. Then, mapping each sample to a 2D latent space representation will theoretically allow us to easily visualize any differences between the samples. The advantage of this VAE approach is that it takes into account every single metabolite (and the clinical data as well), and automatically picks up on the largest signals in the data to differentiate the samples by comparing the reconstruction to the original data. We can also

investigate the trained model parameters themselves to see which metabolites result in the largest signals. The only added difficulty in dealing with these metabolites is that they are continuous measurements, while pixels are typically Bernoulli (in the case of grayscale images) or at least discrete.

In practical terms, the model is structured with the following components:

- Encoder: maps the input data to a 2D latent space using a *tanh* nonlinearity; parameters will be trained
- Decoder: reverts the latent variables back to the metabolite data format, again using a *tanh* nonlinearity; parameters will be trained
- Log prior distribution: simply a standard Normal distribution that serves as our distribution for generating the latent space
- Log likelihood distribution: a truncated Normal distribution that computes the likelihood of our metabolite data given the latent space; mean and standard deviation vary depending on the sample and training parameters
- Joint log density: computes a joint density of the log prior and the log likelihood; this is the basis for comparing the reconstruction to the original data

This might seem a little confusing at first, so I've drawn up a diagram showing the flow of the data through this model. Note that *both* the parameters of the Encoder and the Decoder can be tweaked using gradient descent.

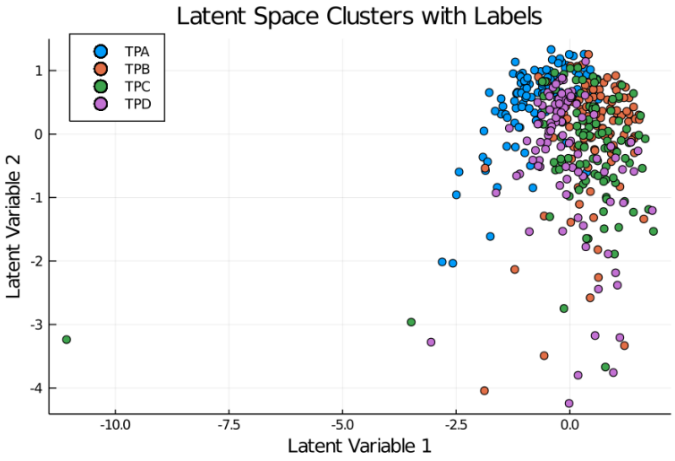


The code for all this was written in Julia largely from scratch. Some Python packages have built-in VAEs, but I found they did not work well for this kind of data. If any of the above seems confusing, the code is fairly well commented, so have a look at that for clarification.

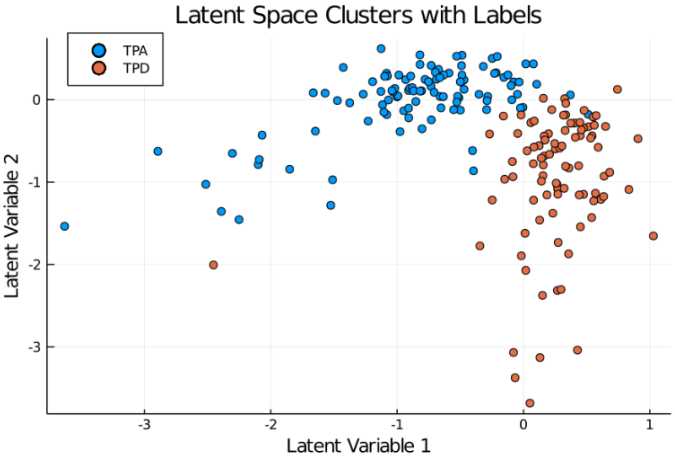
I should note that I also looked at tSNE and UMAP for this analysis, but the results were less promising than those from the VAE, so I won't be discussing them here.

Results

Early results for this analysis were fairly promising with just metabolites (no clinical data) included; I was able to see a reasonable amount of separation between the time points. Starting off, I ran all four time points together at once. Note that each dot on the graph below represents a single sample, and because the latent space is randomly sampled, the location of each sample in one run cannot be accurately compared to the next.

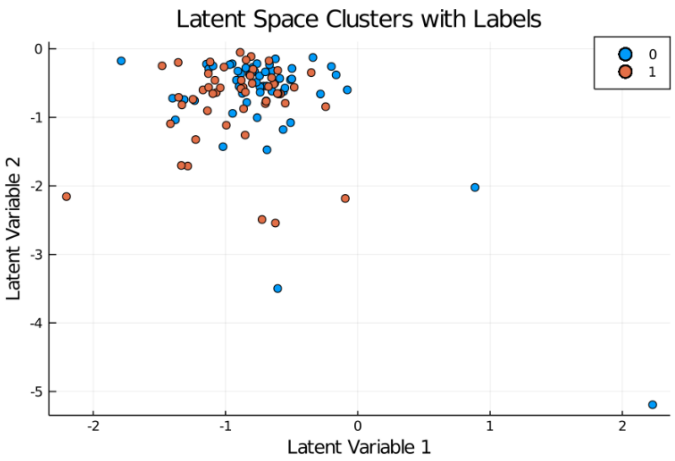


We can see some pretty decent separation of all time points here, particular with Time Point A and Time Point D. There is more overlap than I would like overall though, and this appears to be only marginally better than PCA. We can see this separation even more clearly by running only Time Point A and Time Point D together:

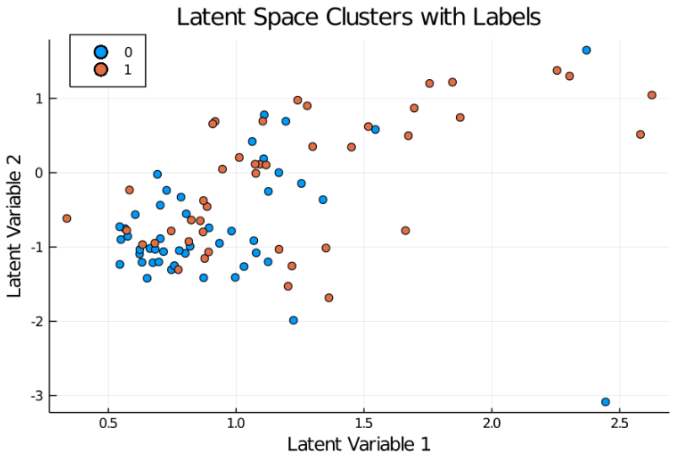


The separation is clearly much better and more pronounced here, indicating that there are clear signals that differentiate these two time points from each other. This is quite a bit better than PCA.

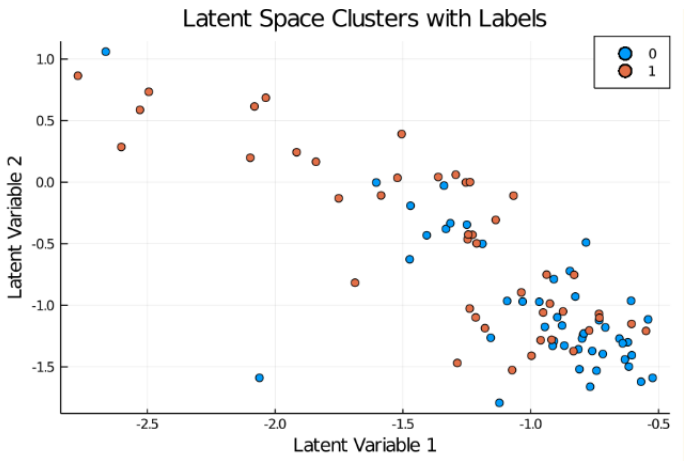
Now that we have shown that there is a latent distinction between earlier and later metabolite profiles, we can see if this sort of separation is present in healthy vs disease samples. Unfortunately, running just Time Point A did not produce any differentiation. Both A and D cluster together indiscriminately.



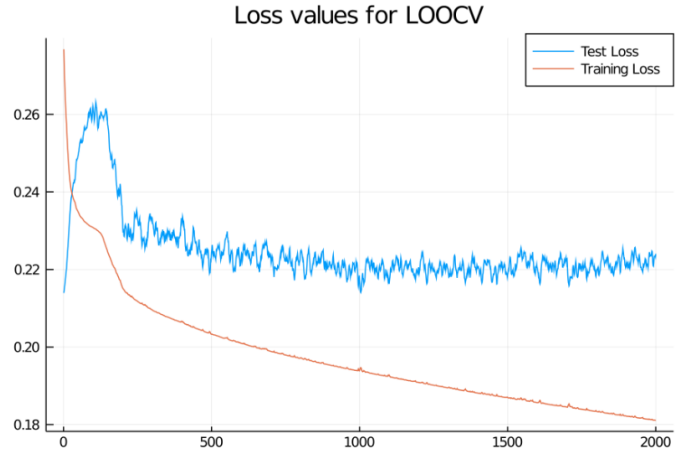
Next, I tried running only Time Point D to see if I could reveal more of a separation between the healthy and disease samples. I think here the separation is slightly better, but certainly not conclusive enough to say for certain if there are any inherent differences between the metabolite profiles.



Adding clinical data did not help as much as I hoped that it would have, but here is a plot of Time Point D with metabolite and clinical data included. There is perhaps a little more separation, but not as much as I would have hoped.



Digging into the model output a little bit more, we can kind of see why this VAE method might not be super well suited to our data. After running leave-one-out cross validation, our model accuracy (in predicting whether an individual sample was healthy or disease) was only 63.52% at Time Point D. Unfortunately that's worse than real world accuracy using only clinical information. So our model here is not incredibly useful for predicting disease state. It is however nearly 100% accurate in predicting which time point a sample came from (either A or D), so it is not a total failure. Here are the loss values for the training set and a test set (Time Point D).



We can see that the test loss is quite variable, indicating model instability, and that the test loss actually starts to go up as time goes on, indicating that we may be overfitting the data at a certain point. I have several thoughts on why this may be the case, given below.

Conclusion & Final Remarks

Although it is unfortunate that this VAE model was not very accurate at estimating whether an infant has BPD or not, it does seem to at least function in a basic sense. The fact that different time points are able to be distinguished from each other shows that the model does have some merit, and that there is some underlying difference between the metabolite profiles at different times after an infant's

birth. It also does seem that the separation between the healthy and disease samples increased slightly as time went on; I would really have liked to have had additional time points taken much later on to see if that trend continued.

That really is the story of this particular exercise; the sample size may have been too small to successfully make use of this VAE method. The unstable and oscillating test loss shown above drives this point home; ideally the test loss would not be so uneven, and would not start to increase after a relatively short period of time. It is also possible that the array of metabolites that were assayed simply have very little to do with BPD. Perhaps another metric combined with the

And finally, I would have liked a few more time points (either from farther out or more frequent) to leverage the temporal aspect of the data. Four time points really isn't enough for any meaningful temporal analysis. I even tried to build a recurrent neural network to work in all the different time points, but was unfortunately unable to get any meaningful results. There may in fact be something useful in these metabolite profiles, and the VAE may be a reasonable method to extract signals from this data, but I think we would need more samples and more time points to really take advantage of this sort of data. Or, using only the data we are given, a different model would likely be optimal. The VAE is designed to handle thousands or hundreds of thousands of images, so 121 metabolite samples is probably not enough information to properly train this sort of model.