1 TO DO LIST:

- The nicest we can do is reduce the dimensionality to 2 factors and plot by hue class. Or just pick 2 first factors from pca.
- More importantly we need some kind of scree plot. To justify working on the PCA. Implementation of the scree plot. We can jsut add the components explained variance with cumsum and it works.
- Correct way to use t-sne if we have space is to plot different perplexities.
- Tough to do what shak is doing with pca Example from sklearn

2 Preliminary data analysis

2.1 Basic facts

We have a dataset that consists of 5471 (n) samples with 4124 (p) columns. The features of our data are all continuous, in a logarithmic scale and they expression levels for genes. It is worth mentioning that the dataset is sparse, that is, a lot of cells have many gene expressions that are 0. While it is hard to have an objective measure of sparsity we can plot average gene expression levels across all our features, see Figure 1 below, most genes are non zero for a limited number of cells.

Figure 1: Possible figure of average gene expression level across all genes. We can do either histogram or histogram by label.

We have 2 distinct classes of cells, the TREG cells and the CD4+T cells. These are going to be our labels that we want to classify. We have some class imbalance, the ratio is 6/10 in favour of the CD4+T which is the dominant class. This is going to be important for the models we try to tune as some models have options to adjust for class imbalance. We will also try the option of tuning the threshold decision for classifying to one cell or another. By default the probabilistic models in Scikit-learn classify to the positive class if the conditional probability for the given model $\mathbb{P}(y|X) > 0.5$, see scikit references. We will tinker with this threshold give our class imbalance.

2.2 Visualization and dimensionality redcution

Both the fact that $n \approx p$ and the fact that the data is sparse point us towards using towards using regularization.

Make sure we explain why it doesn't work.

3 Training baseline models and hypertuning

- Most tuning is done, VSM give very good results.
-
- RESULTS FROM GRIDSEARCH ARE NOT GOING TO BE REPLICABLE, RESULTS FROM FINAL ESTIMATOR BASED ON IT SHOULD BE.

4 Our 3 models

- AdaBoost, just because we have seen it in class and it could be interesting.
- I wanted some classically strong classifier like random forest or GBDT?
- Play around with the PCA optimality. Or some other modern dimensionality reduction tool like t-SNE or UMAP.

5