scvis

Pascale Walters

July 12, 2018

1 Introduction

Single cell gene expression (scRNA-seq) data is commonly used to study the properties of cancer and non-cancer cells, whereby sequencing RNA from a cell and mapping it to genes in the DNA gives an idea of how the cell is functioning. Since there are so many genes of interest in cancer research and a great number of cells are sequenced at once, cell by gene matrices can become very large. Dimensionality reduction must therefore be performed before any sort of clustering or feature extraction can occur.

Two methods commonly used for dimensionality reduction are principal component analysis (PCA) and the t-distributed stochastic neighbour embedding algorithm (t-SNE). However, neither of these are optimized for the type of data obtained from scRNA-seq. While t-SNE is more suited for dimension reduction than PCA, it is still impossible to add more data to an existing t-SNE embedding. In addition, t-SNE is sensitive to hyperparameters, does not scale well, does not output uncertainties of the embedding and can miss subclusters that make up a larger cluster [1].

These drawbacks lead to the development of the scvis algorithm [1], which is designed to perform dimension reduction in scRNA-seq data to allow for further analysis. scvis uses a probabilistic generative model and allows for the addition of new data points to an existing embedding. This document describes the R package that has been developed for the use of this algorithm.

2 scvis Package

scvis has been implemented as an R package to interact with scRNA-seq data stored as SingleCellExperiment objects. The algorithm can be used to create a new mapping from a high-dimensional gene expression matrix to a low-dimensional representation or to add scRNA-seq data to an existing embedding.

As an example to be used for demonstration in this report, scRNA-seq data from ascites will be used. This dataset contains values for 500 cells and 5000 genes.

Before using scvis, noise in the data should be reduced by performing an initial dimension reduction on the data. For the example dataset, the number of dimensions has been reduced from 5000 to 100 using PCA.

2.1 Training the Model

The initial mapping model is generated by running the scvis_train function. For example, to create a mapping on the SingleCellExperiment instance example_sce:

```
scvis_train_sce <- scvis_train(example_sce, "output/", use_reducedDim = TRUE,
    reducedDim_name = "PCA")
```

This carries out the scvis algorithm on the PCA results, which have been stored in a reducedDim slot called PCA. It creates a new SingleCellExperiment instance scvis_train_sce with the same value as example_sce, except the results of dimension reduction are in a reducedDim slot called scvis.

The results of scvis have also been stored in a .csv file in output/, as well as the log likelihoods, elbo results and t-SNE costs.

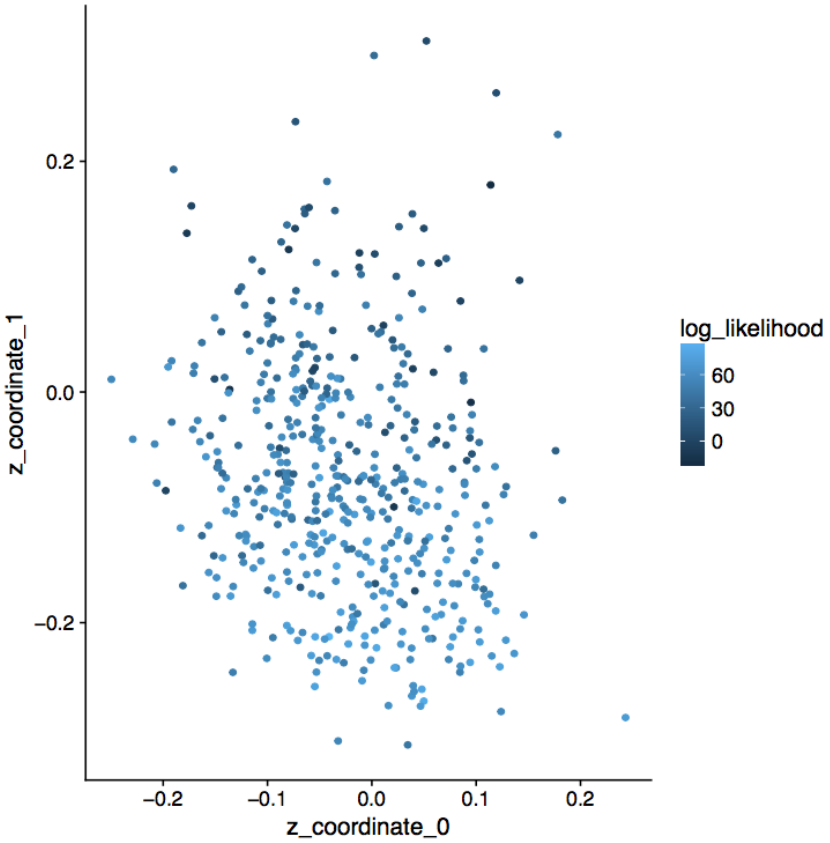

Figure 1: scvis Dimensionality Reduction Results

2.2 Plotting Objective Function Results

The results of the objective function can be plotted directly from the .csv files that have been generated during the training of the model.

2.3 Adding New Data to an Existing Embedding

Suppose we had an existing trained model saved in output/model/ and we wanted to add more data to it. We can use scvis_map with another SingleCellExperiment object containg scRNA-seq data.

```
scvis_map_sce <- scvis_map(example_sce, "output/", use_reducedDim = TRUE,
    reducedDim_name = "PCA")
```

The data that is added must have the same number of dimensions as the data used to train the model. For example, because we used data that had been reduced to 100 dimensions, we must use data in the scvis_map function that also has 100 dimensions.

This function adds the data from the PCA reducedDim slot to the model stored in output/model/. It also writes the results to a .csv file in output/ as well as the log likelihoods. The reduced dimensions are also added to the scvis reducedDim slot of scvis_map_sce.

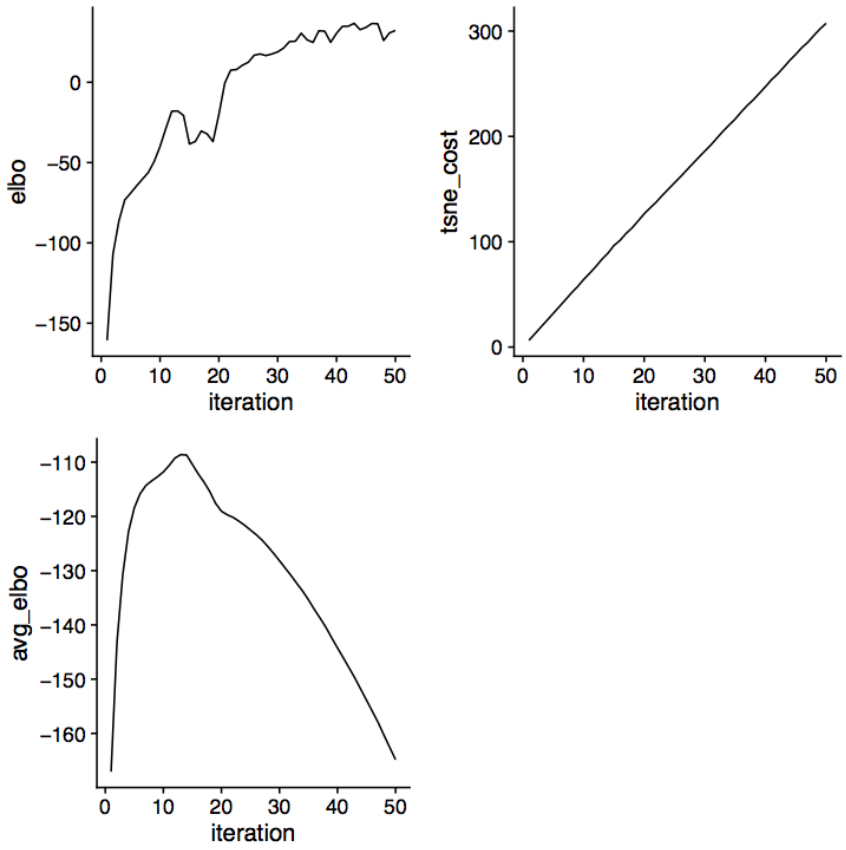

Figure 2: Elbo Results and t-SNE Costs after 50 Iterations

References

[1] Jiarui Ding, Anne Condon, and Sohrab P. Shah. *Interpretable dimensionality reduction of single cell transcriptome data with deep generative models.* Nature Communications, (2018)9:2002, 2018.