Generative Models for scRNA

Jackson Loper

March 12, 2018

Status Succesfully trained the new "iterative model," looked at it

Todo Compare clusters against Allen, compare held-out log likeloohd

1 Overview

The Allen Institute has developed a method for clustering, one that has apparently allowed them to distinguish 116 unique cell types using single cell RNA data. The method follows the following iterative procedure:

Algorithm 1: itclust

Data: RNA expression for a collection of cells

```
Run PCA on the data (e.g. take first 16 components);
Run T-SNE on the resulting low-d representation;
Circle things that look like clusters;
if there is more than one cluster then
    for each cluster c do
        run itclust on the cells inside c;
    end
end
```

This method is nice because the important low-dimensional representations *within* a cluster may be quite different from the ones that help you distinguish *between* two subclusters within that cluster. This is what allows them to achieve the level of granularity that they have.

Unfortunately, it is a bit difficult to grasp the uncertainty involved in the process. There is certainly noise in the data, and different practitioners of this basic procedure may have different understandings of what this noise is. The result is that when Bosilijka Tasic (one of the lead experimentalists at Allen) gives two data scientists the same data, they each come back with different answers. And they have no way to compare or evaluate these two different answers. One essential question is "what level of granularity does the signal-to-noise ratio support?" The algorithm above, as stated, runs iteratively until subclusters cannot be distinguished via T-SNE. But this is obviously a very subjective matter. What to do?

We propose to help clarify some of these issues by building a fully generative model of gene expression, one that rigorously measures some kinds of uncertainty in every step of the process. To be most useful to the Allen Institute, however, we also want this generative model to fit into the scientific paradigm under which they are operating. In particular, the treatment of clusters should be hierarchical in nature, and it is desirable that the model

have latent representations that roughly correspond to the low-dimensional representations found in the Allen algorithm.

Towards this end, we propose the following generative model for gene expression of a single cell. Let H denote a tree of cell types.

1. $T \sim \text{Categorical}(\pi)$, one of the leaves of H. Let $t_0, t_1(T), t_2(T), \cdots, t_\ell(T)$, indicate the nodes in the unique path in H which leads from the root to the leaf T. Here t_0 is the root of the tree, ℓ is the depth of the leaf T, and $t_\ell(T) = T$. To consolidate notation, we take $T_i \triangleq t_i(T)$.
2. For each $i < \ell$ let $Z_{T_i} \sim \mathcal{N}(\mu_{T_{i+1}}, \Sigma_{T_{i+1}})$, and take $Z_{T_\ell} \sim \mathcal{N}(0, I)$
3. For each gene g, let $\lambda_g \in \mathbb{R}^p$ denote a set of parameters. These will be a deterministic function of the Zs:
$$\lambda_g = f_g\left(\sum_{i=0}^{\ell} h_{T_i} Z_{T_i}\right)$$
4. Conditioned on Z, the gth gene of the cell be distributed according to some $p(x_g; \lambda_g)$.

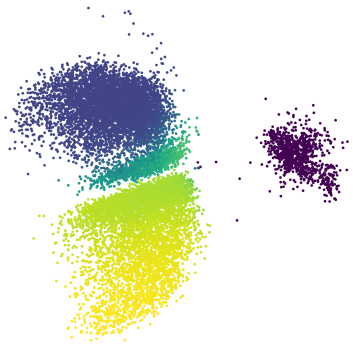
Let us unpack some of this:

- The latent representation Z_i corresponds to features of that cell which are specific to all the cells which are descendents of T_i.
- The the cell's subcluster T_{i+1} effects the distribution of its latent features at the ith level, Z_i. That is, the latent distribution on the ith latent representation is governed by what subcluster it is part of within T_i.
- The distribution of the final latent representation, Z_T, is simply a standard normal.
- The functions h_{T_i} indicates how these cell-type-specific features alter the distribution on the gene expressions for cells in that branch of the tree.

2 Empirical results

At the current moment, we have only actually implemented this model in the case the H is a tree with depth 1, i.e. the hierarchy is trivial. We use amortized variational methods for training and inference. It looks fairly promising. For a simple qualitative assessment, consider the following.

Scatterplot of the aggregate posterior distribution of Z_{T_0}, colored by posterior mean of the cluster assigment:



Notice that there are only so many clusters that are really clearly distinct in this latent space. However, let's focus on that cluster on the right-hand side. What happens if we look at the Z_{T_1} variables for those cells? We get substantial subclustering apparent in the Z_{T_2} space: