Identifying Biologically-defined Groups as Clusters in Single-Cell Transcriptomics using Expectation-Maximization

Karthik Srinivasan, Karthik Desingu, Siddharth Viswanath, John Lazzari

May 8, 2024

1 Introduction

Clustering single-cells using their transcriptomic profiles has a wide range of applications. It is, in fact, one of the first steps in single-cell studies to detect distinct cell populations that can be annotated as known cell types or discovered as novel ones [1]. In this project, we use the Expectation-Maximization (EM) algorithm [2] to identify clusters in single-cell transcriptomic data. We applied it on a single-cell RNA sequencing dataset (scRNAseq) to test its ability to reproduce biologically-defined groups of cells as clusters.

The EM algorithm takes the expected number of clusters, n, and the prior probability distribution as hyperparameters to define a mixture model composed of n individual prior distributions. Then, it iteratively alternates between an *expectation (E)* step and a *maximization (M)* step to find parameters of the n distributions that maximize likelihood of the mixture model, given the data. Since the method works off of a specified number of expected clusters, the inference problem can become particularly challenging when there are multiple sources of variation defining the biological groups. In such scenarios, we expect that the algorithm should identify the clusters defined by the source that explains most of the variance, whilst allowing other sources of variation to mix within each cluster. This would allow us to repeat the inference process within each cluster to identify sub-clusters stemming from other sources of variation.

Notably, we found that the choice of prior was crucial for the algorithm to converge well. To validate the implementation, we first applied the algorithm to identify clusters in synthetically generated datasets sampled from defined distributions. The choice of different priors with these datasets, and their corresponding performance metrics and convergence plots then gave us a framework to assess the choice of priors for the real dataset. *Overall*, we found that the EM algorithm, with a suitable choice of prior as assessed based on convergence plots, was able to reproduce biologically-defined clusters in the data. In addition, EM-based clustering and con-

vergence graphs can be used to get an impression about the underlying distribution of the features in the data.

2 Data

We tested the EM approach in identifying logical clusters in two datasets: first, a synthetic dataset generated by sampling Poisson and Gaussian distributions; second, the open-access, real-world *Tabula Muris* consortium [3].

To first validate the EM method and its implementation, we generated two synthetic datasets with 1000 data points and 100 features. To generate the first synthetic dataset $\mathcal{D}^{syn}_{pois}$, we sampled 500 values for each feature from two randomly chosen (from $uniform(0, 20)$) to form two distinct clusters. We then generated the second synthetic dataset $\mathcal{D}^{syn}_{gaus}$ in a similar manner, this time sampling from a gaussian distribution with means and standard deviations in the two clusters chosen from $uniform(0, 20)$ and $uniform(0, 5)$ respectively.

To test EM clustering on a real-world dataset, we used the *Tabula Muris* dataset. The dataset contains transcriptomics of $100,000$ cells from across 20 different tissues and organs. This worked as a suitable, realistic single-cell compendium to analyze the effectiveness of the proposed EM-based clustering approach since it represents a good collection of single cells with potential variability arising from annotated sources including sex, tissue, and anatomical region among others.

3 Methodology

We adopted an unsupervised machine learning method called Expectation-Maximization to identify the biologically-defined groups as clusters. The data preparation to run and test the approach, and the mechanics of the EM-based clustering method are described here.

3.1 Data preparation

The Tabula Muris dataset comprises transcriptomic data from cells of different sexes and ontology classes of mice. To analyze the clustering ability and robustness of the EM method, we prepared two different types of data subsets. In the first set of data subsets, we drew cells from a single ontology class but from different sexes of mice ($\mathcal{D}_{sex}$). In the second, we included cells from two ontology classes at a time while also including both sexes of mice ($\mathcal{D}_{sex\times ont}$). Data subsets from $\mathcal{D}_{sex}$ and $\mathcal{D}_{sex\times ont}$ were used as the real-world dataset to evaluate the clustering algorithm.

Further, to quantitatively validate the EM algorithm and its implementation, we used the generated synthetic datasets, $\mathcal{D}^{syn}_{pois}$ and $\mathcal{D}^{syn}_{gaus}$, with known underlying distributions and gauged the clustering performance with different choices of priors.

These observations would also help assess the quality of clustering performance on the real-world dataset relative to an ideal, synthetic dataset.

3.2 EM-based clustering

To identify clusters in each data subset, both real-world and synthetic, we used the EM algorithm with two different prior distribution settings. Concretely, for a dataset with d data points with n clusters, the joint mixture model is defined as shown in equation 1.

$$p(x; \alpha, \theta) = \sum_{k=1}^{n} z_{ki} \alpha_k p(x; \theta_k) \tag{1}$$

where α_k denotes the mixing coefficients between the cluster distributions, $p(..; ..)$ denotes the probability density function of the chosen prior distribution, and θ_k denotes the current parameter setting of the prior distribution for cluster k.

The EM clustering algorithm is then sought to maximize the log-likelihood of the joint probability distribution over all d data points described in equation 2.

$$L(\alpha, \theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{d} \sum_{k=1}^{n} z_{ki} log(\alpha_k p(x_i; \theta_k)) \tag{2}$$

where $\mathbf{z}$ is a matrix of $n \times d$ indicator latent variables, one per data point per cluster, such that its elements z_{ki} are 1 when the data point i belongs to cluster k and 0 otherwise. However, the unobserved latent variable $\mathbf{z}$ makes the maximization of $L(\alpha, \theta; \mathbf{x}, \mathbf{z})$ intractable. Hence, we instead choose to maximize its lower bound by substituting $\mathbf{z}$ for its conditional expected value $\hat{\mathbf{z}}$ [4, 5]. This lower bound becomes the objective of maximization and is shown in equation 3.

$$\mathbb{E}_{\mathbf{z}|\mathbf{x};\theta^{(t)}}[L(\alpha, \theta; \mathbf{x}, \mathbf{z})] = \sum_{i=1}^{d} \sum_{k=1}^{n} \hat{z}_{ki} log(\alpha_k p(x_i; \theta_k)) \tag{3}$$

where the expectation $\hat{z}_{ki}$ is given by $\frac{\alpha_k p(x_i;\theta_k)}{\sum_{m=1}^{n} \alpha_m p(x_i;\theta_m)}$. In each iteration of the EM algorithm, the E step computes the expectation $\hat{z}_{ki}$ and the M step then computes parameters α_k and θ_k that maximize equation 3.

In the context of our problem, we tried Poisson and Gaussian priors on each data subset. Data subsets, as described above, were chosen such that they had either one ($\mathcal{D}_{sex}$) or two ($\mathcal{D}_{sex \times ont}$) known sources of variation defining the biological groups in the data; the EM algorithm was run to identify these groups as $n = 2$ distinct clusters, and to get an impression about the underlying distribution of the data based on convergence and performance measures.

4 Implementation

We implemented the EM clustering algorithm on the synthetic and real-world datasets under the following settings.

Preprocessing. Each gene expression value in the dataset was first *normalized* by dividing by the sum of across samples, and then *standardized* by removing the mean and scaling to unit variance. The normalization and standardization steps on features were performed across samples within each data subset.

Hyperparameters and Convergence. The EM clustering method was set to maximize likelihoods of $n = 2$ clusters in the data. We tried two choices of priors for each real-world data subset, *Gaussian* and *Poisson*. Clustering iterations were run for 10 iterations, with an early stopping criterion requiring an increase in log-likelihood of at least $1e - 02$ to continue.

Software libraries. All experiments were carried out in Python on Google Colab. We used SciPy for all statistical tools, Pandas and NumPy for data handling, Matplotlib for visualization, and Scikit Learn for metrics, principal component analysis, and data preprocessing. The EM clustering method was implemented as a Python class. The complete implementation and data used can be accessed from the code base (see code and data availability statement).

5 Results

To quantitaively assess the quality of clusters obtained against expected biological grouping, we used the Adjusted Rand Index (ARI). ARI is a measure of similarity between two data clusterings, that evaluates to 1 when completely similar and -1 when completely dissimilar. We found it to be a suitable metric, since the EM method starts by randomly assigning points to clusters first and the ARI metric accordingly adjusts for the fact that some agreement between the expected and obtained clusterings can occur by chance. Then, to visually assess the clusters, we performed principal component analysis (PCA) on the normalized and standardized data points and plotted them on the plane defined by the first two components, assigning color attributes on the plot based on expected and obtained clusters.

We first validated the implementation of the EM clustering algorithm by running it on $\mathcal{D}^{syn}_{pois}$ and $\mathcal{D}^{syn}_{gaus}$ using Poisson and Gaussian priors respectively. The clustering results were per expectation, achieving ARI scores of 1.0 each (see Table 1).

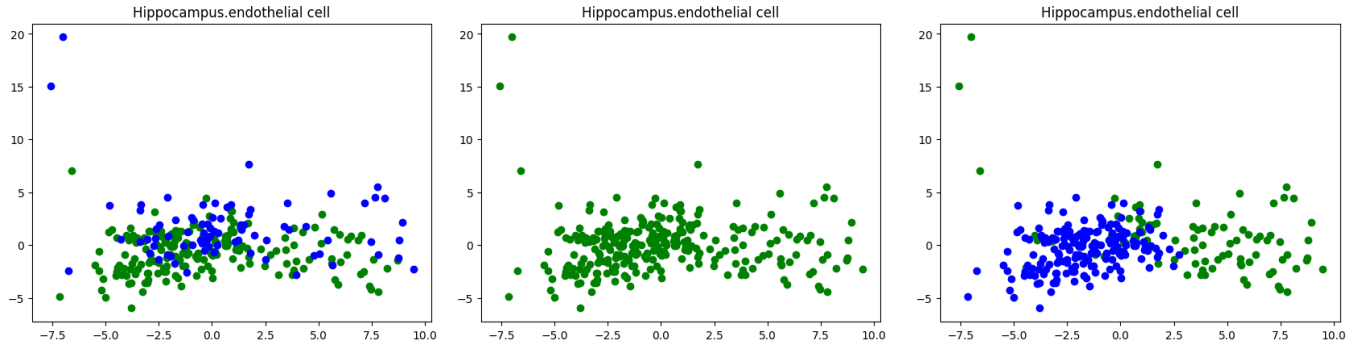
With the real-world dataset, we first used the Poisson prior to find clusters expecting that this would be a good choice of prior since gene transcription can generally be described as a Poisson process. However, we found that EM clustering did not

Dataset	Organ:Ontology	Sources of Biological Variability	ARI Score	
			Poisson	Gaussian
$\mathcal{D}_{pois}^{syn}$	-	Sampled params	-	1.0
$\mathcal{D}_{gaus}^{syn}$	-	Sampled params	1.0	-
$\mathcal{D}_{sex}$	Cerebellum:Endothelial	Sex	0*	0.0903
$\mathcal{D}_{sex}$	Cerebellum:Neuron	Sex	0*	0.0174
$\mathcal{D}_{sex}$	Cerebellum:Astrocyte	Sex	0*	-0.0255
$\mathcal{D}_{sex}$	Cortex:Astrocyte	Sex	0*	0.0384
$\mathcal{D}_{sex}$	Cortex:Endothelial	Sex	0*	0.5742
$\mathcal{D}_{sex}$	Cortex:Oligo	Sex	0*	-0.0169
$\mathcal{D}_{sex}$	Hippocampus:Endothelial	Sex	0*	0.0238
$\mathcal{D}_{sex}$	Hippocampus:Oligo	Sex	0*	0.4617
$\mathcal{D}_{sex}$	Striatum:Endothelial	Sex	0*	0.0831
$\mathcal{D}_{sex}$	Striatum:Neuron	Sex	0*	-0.0682
$\mathcal{D}_{sex}$	Striatum:Oligo	Sex	0*	0.6333
$\mathcal{D}_{sex\times ont}$	Cerebellum:Endothelial-Neuron	Sex, Ontology	0*	0.9602
$\mathcal{D}_{sex\times ont}$	Cerebellum:Endothelial-Oligo	Sex, Ontology	0*	0.4736
$\mathcal{D}_{sex\times ont}$	Cerebellum:Neuron-Oligo	Sex, Ontology	0*	0.4069
$\mathcal{D}_{sex\times ont}$	Cortex:Astrocyte-Endothelial	Sex, Ontology	0*	1.0
$\mathcal{D}_{sex\times ont}$	Cortex:Astrocyte-Oligo	Sex, Ontology	0*	0.7732
$\mathcal{D}_{sex\times ont}$	Cortex:Endothelial-Oligo	Sex, Ontology	0*	0.7733
$\mathcal{D}_{sex\times ont}$	Hippocampus:Endothelial-Oligo	Sex, Ontology	0*	0.7557
$\mathcal{D}_{sex\times ont}$	Striatum:Endothelial-Neuron	Sex, Ontology	0*	0.9212
$\mathcal{D}_{sex\times ont}$	Striatum:Endothelial-Oligo	Sex, Ontology	0*	0.7041
$\mathcal{D}_{sex\times ont}$	Striatum:Neuron-Oligo	Sex, Ontology	0*	0.9799

Table 1: Clustering performance on different datasets and their subsets with two choices of prioirs – Poisson and Gaussian. ** indicates that the algorithm did not converge.*

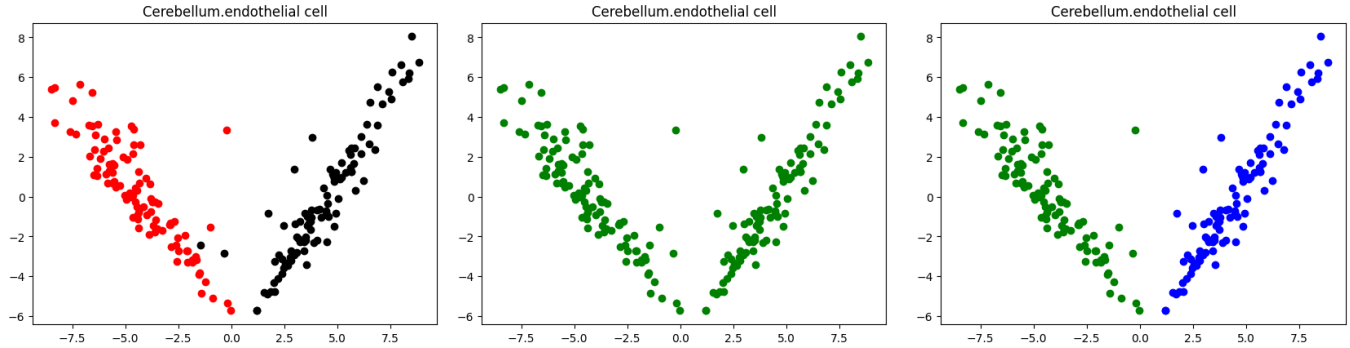
converge and performed poorly with this choice of prior on both real-world datasets $\mathcal{D}_{sex}$ and $\mathcal{D}_{sex\times ont}$ (see Table 1). We speculate that this could be arising either from a dimensionality issue, where the number of samples in the data subsets might be too few to represent variability information from all the genes, or because the expression values are, in fact, not following a Poisson distribution.

We then sought to use a Gaussian distribution as the clustering prior. Once again, after validating it first on the synthetic dataset $\mathcal{D}_{gaus}^{syn}$, we applied it to the real-world datasets. The ARI scores obtained on each of the real-world data subsets are noted in Table 1. We found that the normal prior was able to identify clusters defined by cell ontology groups very well. However, it failed to converge when applied to the dataset containing only sex as a variation. Furthermore, when both sources of variation were combined, the EM method was able to delineate the cell ontology classes while still failing to identify the sex-defined groups. This is in agreement



(a) Ground truth (b) Poisson prior (c) Gaussian prior

Figure 1: PCA visualizations of the Hippocampus:Endothelial data subset in $\mathcal{D}_{sex}$ and the predictions obtained with two different choices of priors.



(a) Ground truth (b) Poisson prior (c) Gaussian prior

Figure 2: PCA visualizations of the Cerebellum:Endothelial-Neuron data subset in $\mathcal{D}_{sex\times ont}$ and the predictions obtained with two different choices of priors.

with our expectation stated above that the EM method will capture the largest variability source when identifying clusters — the cell ontology classes, in this case. The PCA visualizations are shown in Figures 1 and 2. The performance scores can be seen in Table 1.

We conjecture that the failure to capture sex-variability as clusters might be because the transcriptomic data does not show significant variation in gene expressions across the sexes. This is supported by the PCA-space plots of ground truth we obtain (see Figure 1), where even the highest variance capturing principal components are not able to delineate the sex-defined clusters well. Moreover, the number of sex-specific genes in healthy mice are a small proportion of the total number genes whose expression is measured. Hence, this variability may be suppressed by the variability introduced by other factors in the dataset. In conclusion, we find that EM-based clustering is an effective approach to identify variate groups in a dataset when we have some knowledge of the underlying distribution and the number of clusters is known. It is, however, a computationally heavy algorithm with a computational complexity of $\mathcal{O}(dn^i)$, for d data points, n clusters, and i iterations. It can, in general, be used as a downstream algorithm after we have a good idea about the data distribution and the number of clusters to expect, perhaps from performing

agglomerative hierarchical clustering first.

Code and Data Availability

The code for all the experiments performed in the project can be accessed from: `https://github.com/karthik-d/em-clustering-sc-transcriptomics`. The real-world data used for this project was adopted from the Tabula Muris Consortium [3] and is available for public access at: `https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733`. The specific data subsets used in this project can be accessed from the code repository at: `https://github.com/karthik-d/em-clustering-sc-transcriptomics/tree/main/data`.

References

[1] Isabella N Grabski, Kelly Street, and Rafael A Irizarry. Significance analysis for clustering with single-cell rna-sequencing data. *Nature Methods*, 20(8):1196–1202, 2023.

[2] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[3] Tabula Muris Consortium, Overall coordination Schaum Nicholas 1 Karkanias Jim 2 Neff Norma F. 2 May Andrew P. 2 Quake Stephen R. quake@ stanford. edu 2 3 f Wyss-Coray Tony twc@ stanford. edu 4 5 6 g Darmanis Spyros spyros. darmanis@ czbiohub. org 2 h, Logistical coordination Batson Joshua 2 Botvinnik Olga 2 Chen Michelle B. 3 Chen Steven 2 Green Foad 2 Jones Robert C. 3 Maynard Ashley 2 Penland Lolita 2 Pisco Angela Oliveira 2 Sit Rene V. 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, and Computational data analysis Batson Joshua 2 Botvinnik Olga 2 Castro Paola 2 Croote Derek 3 Darmanis Spyros 2 DeRisi Joseph L. 2 27 Karkanias Jim 2 Pisco Angela Oliveira 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.

[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[5] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961, 2012.