EE 377 Project Final Report

Govinda Kamath and Jesse Zhang

13 March 2016

1 Background

With recent advancements in RNA sequencing (RNA-seq) technologies, biologists have enjoyed a surplus of gene expression data at the resolution of individual cells. One of the most interesting problems in biology is that of determining the differentiation patterns of cells such as the process of a stem cell evolving (or differentiating) into a heart cell. A recent experiment [1] has produced a single-cell RNA-seq dataset consisting of 271 human primary myoblasts collected at four different times. The authors attempt to order these cells according to a so-called "pseudotime" or a latent variable that quantifies how differentiated a given cell is. Attempts of studying cell differentiation have been performed by various groups using various methods. For example, Mueller et al. [2] attempt to order *batches* of sampled cells according to the distributions of certain transcripts (or features) within clusters. For this project, we hope to combine the notion of measuring the distance between cells based on distributions of transcripts with the concept of "pseudotime." We hope that by doing this, we can obtain a more biologically meaningful (and statistically sound) ordering of the 271 primary myoblasts. The success of this project could lead to meaningful insights in modeling cell differentiation; these techniques can perhaps be applied to other single-cell datasets studying the growth of cancer, for example.

2 Problem statement

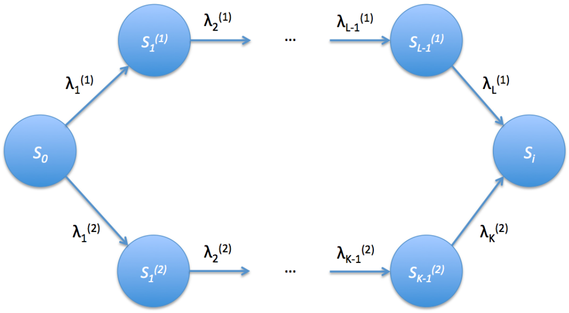
At sampling time t ($t = 1, \dots, T$), we are given n_t length-d vectors representing the n_t cells collected at time t. Each cell is described with d features, which can be gene expression levels, transcript expression levels, or binary markers indicating whether a particular mutation or epigenetic marker was observed. We assume that each data point is some mixture of two or more populations (or cell types). In other words, each point (cell) lived somewhere on a continuum *between* cell types before being sampled (collected). The goal is to learn something about the structure of this continuum.

As stated in our project proposal, one potential approach taken in [3] is to cluster the $N = \sum_t n_t$ cells into k clusters and then draw a minimum-spanning tree (MST) through the k centroids. In this case, the MST supposedly captures the underlying structure. We feel that a more intuitive model is a continuous-time Markov chain (CTMC) where states represent cell types and transition rates represent times required to transition from one cell state to another. After assigning a probability density to each state, we would want to do inference on such a model. One reason for choosing the CTMC is that it's easier to analyze statistically than arbitrary clustering algorithms.

3 Identifiability

We first show that if the CTMC does not have a tree-like structure, then estimating the exponential rate parameters is suffers an identifiability issue. In other words, if the CTMC structure contains multiple paths to the same node, then no number of observations can help us identify the rate parameters. Figure 1 demonstrates why this the case. Consider a node S_i that has at least two unique paths to it from the starting node S_0. Then there exists at least two sets of λ values that produces the same set of observation values. We can set the rate parameter for the last edge on path 1 to ∞, resulting in one solution, or we can set the rate parameter for the last edge on path 2 to ∞, resulting in a second solution. In either case, setting the last edge on a path to ∞ results in all rate parameters for the other path becoming free parameters. They can be whatever they need to be to produce what was observed. We therefore assume that the CTMC has a tree-like structure.

Figure 1: An example of a CTMC with two paths from the starting node S_0 to some other node S_i. Path 1 has L edges with $\lambda_L^{(1)}$ being the last edge, and path 2 has K edges with $\lambda_K^{(2)}$ being the last edge.



4 Analysis using a tree-structured CTMC

We quickly realized that there were several layers of complexity to this problem even with the CTMC assumption. For real-life datasets, one may not necessarily know N_S, the number of states, $\{\lambda_i\}_{i=1}^{M}$, the exponential rate parameters associated with the M edges, $p_S(x)$, the noise model associated with each state (i.e. how observations of cells in state S are distributed), or even which of the $\binom{N_S}{2}$ pairs of states should have edges connecting them.

We assume that we know the underlying structure of the CTMC. Inspired by the results from [1], we assume that samples come from one of 4 cell types described by the CTMC in Figure 2. In the case of [1], state A would be a proliferating cell, state C would be a differentiating myoblast, state D would be an interstitial mesenchymal cell, and state B would be some intermediate hybrid of the first 3 states. Another key assumption we make here is that we can look at the same cell multiple times (or once at each observation time).

Given this CTMC, we can simulate data by setting the number of samples, the values of $\lambda_1, \lambda_2, \lambda_3$, and the observation times, as shown in Figure 3. Using $\lambda_1 = 0.5, \lambda_2 = 1, \lambda_3 = 0.25$, we generate 100 samples. All 100 samples start as cell type A.

Given the structure of the CTMC and these observations, we can easily estimate all rate parameters. Let $P_t(S_1, S_2)$ represent the probability of going from state S_1 to state S_2 after t time has passed, and let t_S represent the amount of time a cell spends in state S. Then

$$P_t(A, A) = P(t_A \geq t) = e^{-\lambda_1 t} \implies \hat{\lambda}_1 = -\frac{1}{t} \log \frac{\text{\# type } A \text{ cells at } t}{\text{\# type } A \text{ cells at time } 0}$$

After estimating λ_1, we can estimate λ_2 and λ_3 using the fact that the rate of exiting state B equals $\lambda_2 + \lambda_3$

Figure 2: Assumed underlying structure for the CTMC. Here, cells start in state A. After spending $t_A \sim \mathsf{Exp}(\lambda_1)$ in state A, a cell transitions to an intermediate state B before ending as either a type C cell or a type D cell. After arriving at state B, the end state of this particular cell is determined by the outcome of an exponential race between λ_2 and λ_3.

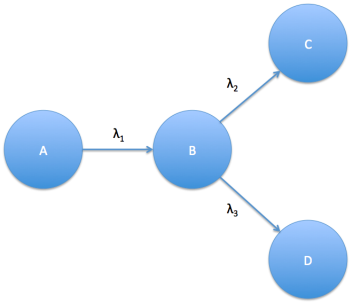
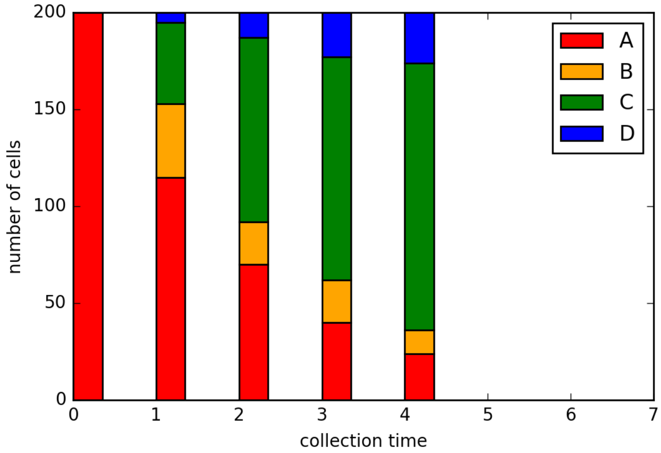


Figure 3: Simulation using the CTMC from Figure 2. 200 cells are sampled and tracked at 5 observation times. We see that as the observation time increases, cells evolve from type A to ultimately type C or D.



(exponential race):

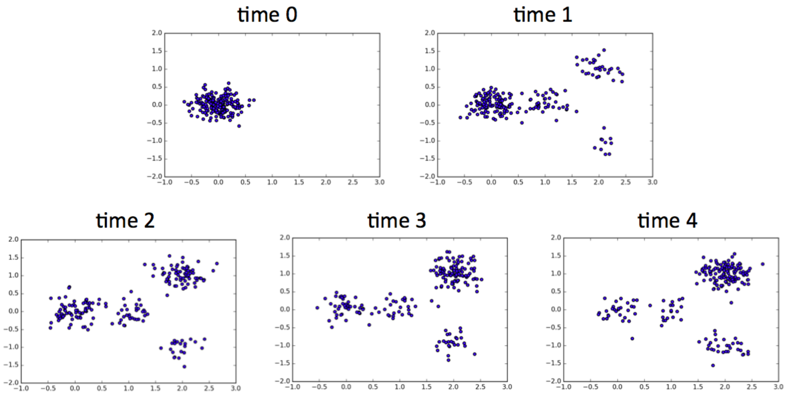
$$\begin{aligned}
P_t(A,B) &= \int_0^t P(t_A)P(t_B \geq t_1 - t_A)dt_A \\
&= \int_0^t \lambda_1 e^{-\lambda_1 t_A}(e^{-(\lambda_2+\lambda_3)(t-t_A)})dt_A \\
&= \frac{\lambda_1 e^{(\lambda_2+\lambda_3)t}}{-\lambda_1+\lambda_2+\lambda_3}\left(e^{t(-\lambda_1+\lambda_2+\lambda_3)} - 1\right)
\end{aligned}$$

Using $\hat{\lambda}_1$ and the ratio of number of B cells at time t to the number of A cells at time 0, we can estimate $\lambda_2 + \lambda_3$. Additionally, we know the ratio of λ_2 to λ_3 from the observations. We solve a system of equations to obtain estimates for λ_2 and λ_3. This approach seems to work reasonably well, and for the example from Figure 3 we obtain rate parameter estimates of $\hat{\lambda}_1 = 0.497, \hat{\lambda}_2 = 0.985$, and $\hat{\lambda}_3 = 0.281$.

Next we simulate noisy observations of each cell at each observation time. We assume that given states $S = \{A, B, C, D\}$ and $K = |S|$, observations are distributed $p_S(x) = \mathsf{N}(\mu_i, \sigma^2 I_K)$ where $\mu_i \in \mathbb{R}^K$. Using the states shown in Figure 3, we simulate observations at each of the five times by setting $\mu_A = (0,0), \mu_B =$

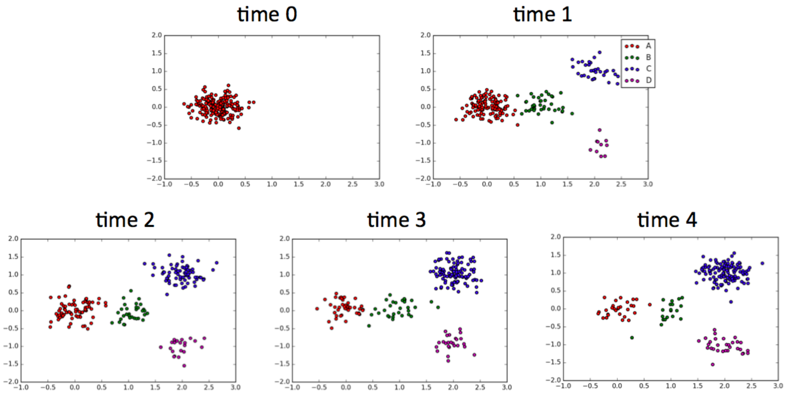
$(1,0), \mu_C = (2,1), \mu_D = (2,-1)$, and $\sigma^2 = 0.05$. The results of the simulation are presented in Figure 4.

Figure 4: Noisy observations are simulated from the cells of Figure 3, resulting in 200 noisy samples at each observation time. We see that the cells shift from the left (states A and B) to the right (end states C and D) as time increases.



Because we know both the number of clusters and that the observations are distributed iid Gaussian, at each observation time we can infer the states of the observed cells using a mixture of Gaussians model, which can be easily solved with a standard expectation-maximization (EM) algorithm. We use the Hungarian algorithm to match the labels output by EM with the true labels, resulting in the cluster assignments shown in Figure 5.

Figure 5: Cluster assignments of each observation generated in Figure 4 using EM.



After assigning points to states, we can estimate the exponential rates just as before: $\hat{\lambda}_1 = 0.501$, $\hat{\lambda}_2 = 0.997$, and $\hat{\lambda}_3 = 0.283$. We still seem to do reasonably well. The problem with this approach is that the estimators for λ_1, λ_2, and λ_3 are complicated and hard to analyze; extending these types of estimators to other tree-structures is not clear. Moreover, deepening our tree structure will involve longer sums of exponential random variables with different rate parameters, further complicating the estimators.

5 A simpler estimator

If we assume that we can track data points (i.e. we observe the state of each point at each sampling time), we can can derive a simple estimator for each sum of rate parameters leaving the same state. Let's say that at time t_j, we observe N_k points at state k, and at time t_{j+1}, we observe M_k of the N_k points remain in state k. Let $\lambda_1^{(k)}, \ldots, \lambda_{L_k}^{(k)}$ represent the rate parameters on the edges leaving state k. The rate of leaving state k is modeled by a race between L_k exponentials (or an exponential with rate $\Lambda = \lambda_1^{(k)} + \cdots + \lambda_{L_k}^{(k)}$). Let $T = t_{j+1} - t_j$. Then

$$\frac{M_k}{N_k} = P(\text{leave state } S_k \text{ by time } T) = P(t_{S_k} \geq T) = \exp\left(-T \sum_{i=1}^{L_k} \lambda_i^{(k)}\right) = \exp(-T\Lambda).$$

Assuming that our observation times are evenly spaced at T, we can use the memoryless property of exponential distributions to effectively increase our number of samples used for estimating Λ. Let $M_k(t)$ represent the number of cells in state S_k at both times $t - T$ and t, and let $N_k(t)$ represent the number of cells in state S_k at time t. If our last observation time is at $t'T$, then

$$\frac{\sum_{t=1}^{t'} M_k(t)}{\sum_{t=0}^{t'} N_k(t)} = \exp(-T\Lambda),$$

which we will call $\hat{p}$ for convenience. Intuitively, $\hat{p}$ is simply ratio of the total number of samples that have ever moved out of state S_k to the total number of samples that were ever in state S_k. Our estimator for Λ is

$$\hat{\Lambda} = -\frac{1}{T}\log(\hat{p}).$$

Both estimators are unbiased, and

$$\begin{aligned} \mathbb{E}\hat{\Lambda} &= \Lambda \\ \mathbb{E}\hat{p} &= p = \exp(-T\Lambda) \end{aligned}$$

We can bound our estimator for $\exp(-T\Lambda)$ using the Chernoff Bound:

$$\mathbb{P}\left(|\hat{p} - \exp(-T\Lambda)| \geq \epsilon\right) \leq \exp\left(-2\epsilon \sum_{t=0}^{t'} N_k(t)\right) = \delta.$$

From this bound we can construct 1-δ confidence intervals for Λ:

$$\begin{aligned} &\exp(-\hat{\Lambda}T) \in p \pm \epsilon \\ \implies &\hat{\Lambda} \in -\frac{1}{T}\log(p \pm \epsilon) \\ \implies &\hat{\Lambda} \in \lambda - \frac{1}{T}\log\left(1 \pm \frac{\epsilon}{p}\right) \\ \implies &\hat{\Lambda} \in \Lambda \pm \frac{1}{T}\frac{\epsilon}{\exp(-\Lambda T)}. \end{aligned}$$

Solving for the ϵ in the Chernoff bound, we can express this 1-δ confidence interval as:

$$\hat{\Lambda} \in \Lambda \pm \frac{1}{T\exp(-\Lambda T)}\sqrt{\frac{\log\frac{1}{\delta} + \log 2}{\sum_{t=0}^{t'} N_k(t)}}$$

After obtaining an estimate for Λ, we can estimate the individual $\lambda_1^{(k)}, \ldots, \lambda_{L_k}^{(k)}$ using the observed ratios of the numbers of samples that end up in each possible descendant branch. For example, if three edges leave S_k with rates $\lambda_1^{(k)}, \lambda_2^{(k)}, \lambda_3^{(k)}$ leading a sample to states $S_{k+1}, S_{k+2}, S_{k+3}$, respectively, then the ratio of the number of samples that move from S_k to S_{k+1} or any of S_{k+1}'s descendants to the number of samples that move from S_k to S_{k+2} or any of S_{k+1}'s descendants is equal to $\lambda_1^{(k)}/\lambda_2^{(k)}$. We can solve a system of equations to obtain values for each $\lambda_i^{(k)}$.

A possible problem we could have is that $\sum_{t=0}^{t'} N_k(t)$ could be small. This is especially true when Λ is large. If we assume that $\Lambda < \frac{1}{T}$, then on sampling enough time points we have that

$$\sum_{t=0}^{t'} N_k(t) \geq n(1 - \frac{1}{e}),$$

where n is the total number of cells sampled. In general as $\Lambda \rightarrow \infty$, the number of cells sampled in a state goes to 0. We can make Λ large enough so that for a state, arrivals into a state have to occur in a limited time interval to be able to captured ($\Lambda = O(n^2)$ for instance). However deriving conditions in general is hard, because we can have periodicity.

Given enough samples, a tree-structured CTMC, a small enough sampling period T, and the state of each cell at each sampling time, we can estimate all rate parameters of the CTMC. We tested this estimator on the tree structure example from Figure 2, and we obtained reasonable estimates of $\lambda_1 = 0.529, \lambda_2 = 1.07, \lambda_3 = 0.202$.

6 Future steps

Initial simulations seem to generate promising results because we have made several strong simplifying assumptions. In the case of the dataset presented by [1], individual data points are high-dimensional vectors (length 47192), making EM much more difficult to compute. Determining the correct method for reducing the dimensionality of the data proves to be a challenge. Additionally, we have no reason to believe that observations come from one of K Gaussian distributions. We have no reason to believe that the Gaussian, equal-variance, or independence assumptions are true. In fact, we intuitively expect the evolution of a cell to be a gradual process, and therefore μ should be a function of some latent time variable. Finally, the set of cells sampled using real data is completely different at each observation time. Ideally, we would be able to account for the the fact that we do not observe time 0 as the first observation does not show a homogenous population of cells all of one type.

If our model is correct, we would like to say something about the guarantees we can make as a function of the CTMC complexity (number of nodes and edges), the number of samples, and the number of observation times. Another aspect would be estimating the tree structure from the data. This seems to be an interesting model selection problem which gives us constraints on the number of time points we sample.

7 References

1. Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. ”The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.” Nature biotechnology 32, no. 4 (2014): 381-386.

2. Mueller, Jonas, Tommi Jaakkola, and David Gifford. ”Modeling Trends in Distributions.” arXiv preprint arXiv:1511.04486 (2015).

3. Ntranos, Vasilis, Govinda M. Kamath, Jesse Zhang, Lior Pachter, and N. Tse David. ”Fast and accurate single-cell RNA-Seq analysis by clustering of transcript-compatibility counts.” bioRxiv (2016): 036863.