Estimation of rate parameters in a tree-structured CTMC

Govinda Kamath and Jesse Zhang

Problem Statement

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, and the authors use a heuristic for recovering a tree structure describing how a cell evolves into one of two end types. For this project, we hope to create and analyze a more statistically sound method of estimating the underlying tree structure. The success of this project could lead to meaningful insights in modeling cell differentiation.

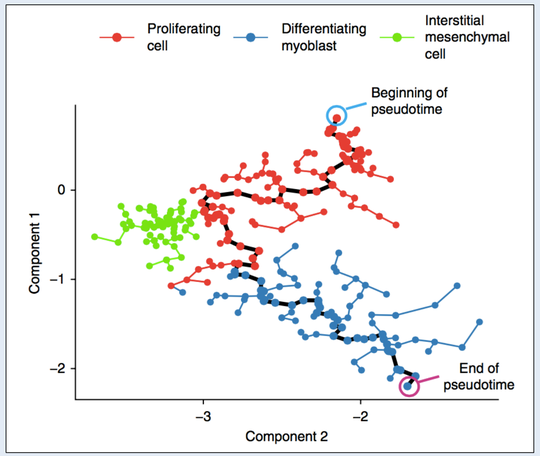

Figure 1. Differential pattern uncovered heuristically by [1].

We propose a continuous-time Markov Chain to describe the differentiation process. We estimate the rate parameters for the Markov chain using observations obtained at multiple sampling points.

References

1. Trapnell, Cole *et al.* "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 accu- rate single-cell RNA-Seq analysis by clustering of transcript-compatibility counts." bioRxiv (2016): 036863.

Identifiability

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.

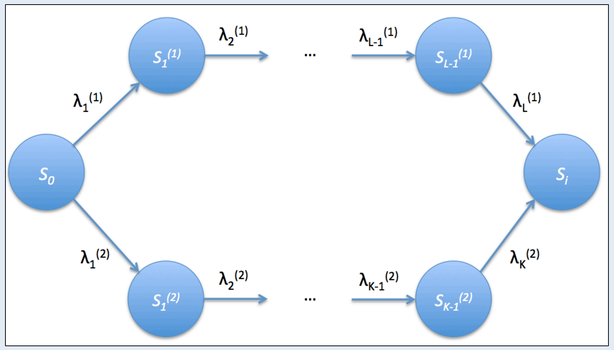

Figure 2. Multiple paths from starting node S_0 to some interior node S_i results in identifiability issues.

Initial simulations

We simulated cell state observations using a CTMC inspired by [1]. The simulations were done using 200 samples, 4 cell types, 5 observation times, and rate parameters of λ_1= 0.5, λ_1= 1, and λ_1= 0.25. The recovered rates were λ_1= 0.501, λ_1= 0.997, and λ_1= 0.283.

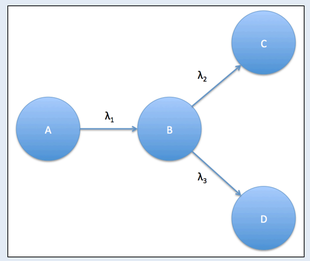

Figure 3. CTMC used to generate simulation points

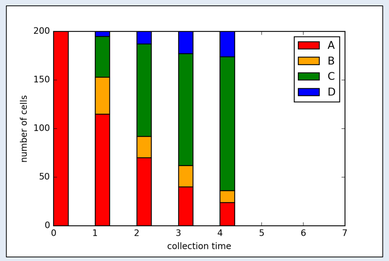

Figure 4. Proportion of each cell type at 5 simulated observation times.

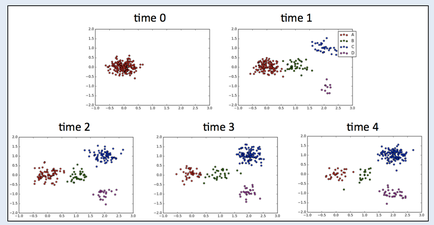

Figure 6. Recovered cell states using expectation maximization of Gaussian mixture.

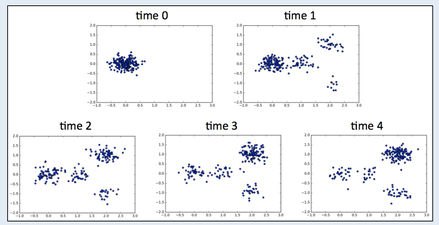

Figure 5. Gaussian noise added to observed expression values for each cell in each observation.

Rate estimators

To estimate the rates of the CTMC in Figure 3, use the following equations to estimate the values of the three rate parameters.

$$\hat{\lambda}_1 = -\frac{1}{t}\log\frac{N_A(t)}{N_A(0)} \qquad \frac{\lambda_1 e^{(\lambda_2+\lambda_3)t}}{-\lambda_1+\lambda_2+\lambda_3}(e^{t(-\lambda_1+\lambda_2+\lambda_3)}-1) = \frac{N_B(t)}{N_A(0)}$$

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.

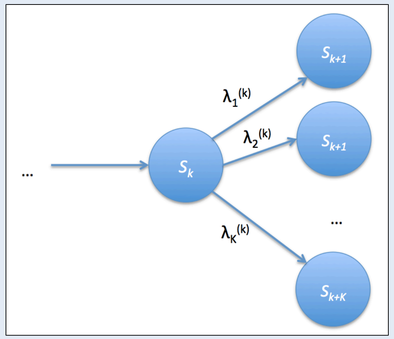

Figure 7. Example of a state node in a general tree-like CTMC. The time a sample spends as state S_k is dictated by a race between K exponential distributions with different rate parameters.

The expressions for the sum of rate parameters leaving the same state and its estimator are:

$$\Lambda = \sum_{i=1}^{K}\lambda_i(k) \qquad \hat{\Lambda} = -\frac{1}{T}\log\frac{\sum_{t=1}^{t'}M_k(t)}{\sum_{t=0}^{t'}N_k(t)}$$

Using the Chernoff Bound, we construct a 1-δ confidence interval for our estimator:

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

Future Work

The relaxation of several assumptions made in this project could lead to interesting results. The problem in reality is significantly harder: each data point is only observed once, the data points have very high dimensions, the underlying CTMC is not known a priori, and both the number of samples and the number of observations are small.