Non-parametric inference of steady state RNA distributions from single cell transcriptomic data

Joshua G. Schraiber

Started on March 17, 2017. Compiled on March 17, 2017

1 Basic model

Suppose we have a genome with G genes in it, each of which produces mRNA transcripts in each of C different cells. We assume that each gene j is expressed at some level e_{ij} in cell i, with $e_{ij} \in \mathbb{N}$. We further assume that there exists a steady-state distribution of expression levels for each gene j, $\mathbb{R}_j$. Because , we can represent $\mathbb{R}_j$ as a sum of delta masses, say

$$\mathbb{R}_j = \sum_{k=0}^{\infty} \pi_{jk} \delta_k$$

where δ_k is the delta mass at k and π_{jk} is the probability that there are k transcripts of gene j in a cell.

We assume we have performed single-cell transcriptomics on each cell. Thus, for each cell i, we have N_i total sequencing reads, and for each gene j, we observe the counts r_{ij} of reads of that gene in that cell. Note that r_{ij} can be thought of as a noisy proxy for e_{ij}, and we specifically assume that the the r_{ij} are obtained by multinomial sampling from the e_{ij}. Thus, our full model is

$$\begin{aligned} e_{ij} &\sim \mathbb{R}_j \\ r_{ij} | e_{ij} &\sim \text{Multinomial}\left(N_i; \frac{e_{i1}}{\sum_j e_{ij}}, \frac{e_{i2}}{\sum_j e_{ij}}, \ldots, \frac{e_{iG}}{\sum_j e_{ij}} \right). \end{aligned}$$

In essence, we would like to infer the π_{jk} from the r_{ij}. Note that if we had direct access to the e_{ij} that would be easy: you can simply estimate

$$\hat{\pi}_{jk} = \frac{\sum_{i=1}^{C} \mathbb{I}\{e_{ij} = k\}}{C}.$$

This suggests an EM algorithm. However, an proper EM algorithm would be very difficult, because of the fact that the read counts of every gene in a cell depends on the read count of every other gene in that cell. Instead, we propose an approximate EM algorithm as follows.