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

Joshua G. Schraiber

Started on March 17, 2017. Compiled on March 23, 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

$$e_{ij} \sim \mathbb{R}_j$$

$$\{r_{ij}, 1 \leq j \leq G\} | \{e_{ij}, 1 \leq j \leq G\} \sim \text{Multinomial}\left(N_i; \frac{e_{i1}}{\sum_j e_{ij}}, \frac{e_{i2}}{\sum_j e_{ij}}, \ldots, \frac{e_{iG}}{\sum_j e_{ij}}\right).$$

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.

We need to compute the posterior probability, given the current estimates of π_{jk}, that $e_{ig} = k$, this is given by

$$\mathbb{P}(e_{ij} = k | r_{ij}, \{\pi_{jk}\}) \propto \mathbb{P}(r_{ij} | e_{ij} = k) \pi_{jk}$$

To do so, we first approximate the distribution of read counts given all of the e_{ij} by the marginal binomial distributions, i.e.

$$\mathbb{P}(r_{ij} | e_{ij} = k, \{e_{ig}, g \neq j\}) = \binom{N_i}{r_{ij}} \left(\frac{k}{\sum_{g \neq j} e_{ig} + k} \right)^{r_{ij}} \left(1 - \frac{k}{\sum_{g \neq j} e_{ig} + k} \right)^{N_i - r_{ij}}.$$

We then marginalize over all the e_{ig} for $g \neq j$. Doing that exactly would be difficult. Thus, we proceed with an approximation. Letting $T_{jk} = \sum_{g \neq j} e_{ig} + k$, we can Taylor exapnd around $\mathbb{E}(T_{jk})$,

$$\mathbb{E}\left(\binom{N_i}{r_{ij}} \left(\frac{k}{T_{jk}} \right)^{r_{ij}} \left(1 - \frac{k}{T_{jk}} \right)^{N_i - r_{ij}} \right) \approx \left(\frac{k}{\mathbb{E}(T_{jk})} \right)^{r_{ij}} \left(1 - \frac{k}{\mathbb{E}(T_{jk})} \right)^{N_i - r_{ij}} (1 + C_k \mathrm{Var}(T_{jk}))$$

where C_k is a really ugly constant. Note that given the π_{jk}, we can compute $\mathbb{E}(e_{ij}) = \sum_{k=0}^{\infty} k \pi_{jk}$ and $\mathbb{E}(e_{ij}^2) = \sum_{k=0}^{\infty} k^2 \pi_{jk}$. Thus, if we let $T = \sum_g e_{ig}$ (i.e. without fixing any genes to a specific value), we can compute $\mathbb{E}(T_{jk}) = \mathbb{E}(T) - \mathbb{E}(e_{ij}) + k$ and $\mathrm{var}(T_{jk}) = \mathrm{var}(T) - \mathrm{var}(e_{ij})$

Then, we can re-estimate the π_{jk} by updating

$$\hat{\pi}_{jk}^{(n+1)} = \frac{\sum_{i=1}^{C} \mathbb{P}(e_{ij} = k | r_{ij}, \{\pi_{jk}^n\})}{C}$$

TODO: incorporate gene length. This might change the C_k constant, so check that...