Identifying dropouts in Drop-seq data

Guilherme de Sena Brandine Chao Deng Andrew D. Smith

September 8, 2016

1 Introduction

Singe-cell RNA-Seq is a technique that allows transcriptome pro

les of tissues in a single-cell resolution. It has a wide variety of applications that range from analysis of tissue heterogeneity in complex systems like tumors and blood tissues to temporal reconstruction of difrentiation processes like embryogenesis. However, due to the small amount of starting material that is the set of messenger RNAs in a single cell, data generated by this procedure suers from large amounts of technical noise. One particular obstacle in analyzing such data are dropout events, in which the messenger RNA fails to be converted into cDNA, and as such it is not ampli

ed and in all posterior steps of analysis the gene's count will be set to zero, whereas in truth it is actually expressed. Since many genes in a single cell are not expressed for other biological reasons (eg, dierent stages of cell cycle or genes that simply aren't relevant to the cell's speciality), it is dicult to decide the actual origin of a zero.

2 Results

2.1 Model assumptions

We assume a gene expression profile is a vector of expression values, all of which sum to 1.

2.2 Complete data likelihood

In this section we give the complete data likelhood for the model outlined above.

2.3 Adapting the model to be more biological meaningful

Here we elaborate on the framework given in Section **??**. In particular,

- We move away from the assumption that dropouts happen uniformly between genes. We will allow gene-specific prior probabilities on dropouts.
- We assumed that a set of neighbors were known for each cell. Now we will use a set of other cells, and their contributions will be weighed based on their distance to the cell of interest.