Clustering of high-throughput sequencing data

1 Introduction

Clustering techniques have often been used in the analysis of genetic data as a powerful exploratory tool. In particular, the STRUCTURE model [**?**] (as known as the admixture/grade-of-membership/latent Dirichlet allocation models) was proposed as a "soft" clustering method for genotype data to infer population structure. Other clustering tools such as k-means and hierarchical clustering have also been used in the analysis of micro-array data, to cluster both genes and samples ([]), as well as sequencing data ([]). More recently the same admixture model was proposed by [**?**] to cluster samples from RNA-seq data, motivated by the argument that each sample could be a mixture of different cell types, so that the membership profile of the sample would indicate the proportions of cell types present.

Motivated by the results from the admixture models in both [**?**] and [**?**], we propose an extension of the sample model to analyze sequencing data at the base pair resolution (RNA-seq reads were summed up within each gene in [**?**]). The goal of such a model is to capture clusters at a finer scale compared to gene-level analysis. For example, certain samples might share similar expression patterns for a portion of a gene, but reveal different patterns in another portion of the same gene. One possible explanation for this might be variations in splicing patterns, and new clusters could potentially reveal transcripts not previously found.

Due to the typically small number of counts when analyzing sequencing data at the base pair resolution, certain assumptions are often required for any given model. Here we make the "smoothness" assumption ([**?**], [**?**], [**?**]), and add an additional "smoothing" step when fitting the admixture model. As we will show, this additional step greatly increases the accuracy of the model when compared to the standard admixture model.

2 Method

We first describe the basic model (without smoothing) for sequencing data in the context of topic models for easier conceptual understanding. We first assume that there is a number of underlying (normalized) intensity profiles (eg expression profiles in the context of RNA-seq), which are the latent "topics", or clusters. Each profile determines the distribution of the read counts (ie the mean at each base pair is the probability that a DNA fragment (which we assume to be of length 1 for simplicity) maps to that base). This is essentially the "vocabulary matrix" in a topic model, where each base pair is a "word". Given the cluster profiles, the generative model can be specified as

1. For the j-th DNA fragment (of length 1) in sample i, first choose cluster Z_{ij} according to $P(Z_{ij} = k|\boldsymbol{\pi}) = \pi_{ik}$, where $\sum_k \pi_{ik} = 1$.

2. Given Z_{ij}, choose the base R_{ij} it maps to according to $P(R_{ij} = b|Z_{ij}, \boldsymbol{\phi}) = \phi_{Z_{ij},b}$, where $\sum_b \phi_{kb} = 1$.

Given this generative model, the likelihood $P(D|\boldsymbol{\phi}, \boldsymbol{\pi})$ is given by

$$\begin{aligned} P(D|\boldsymbol{\phi}, \boldsymbol{\pi}) &= \prod_i \prod_j \sum_k P(R_{ij}|Z_{ij} = k, \boldsymbol{\phi}) P(Z_{ij} = k|\boldsymbol{\pi}) \\ &= \prod_i \prod_b \left(\sum_k P(R_{ij} = b|Z_{ij} = k, \boldsymbol{\phi}) P(Z_{ij} = k|\boldsymbol{\pi}) \right)^{Y_{ib}} \qquad \text{where } Y_{ib} = \textstyle\sum_j I_{\{R_{ij}=b\}} \\ &= \prod_i \prod_b \left(\sum_k \pi_{ik} \phi_{kb} \right)^{Y_i b} \qquad (1) \\ & \qquad (2) \end{aligned}$$

Note that we have re-expressed the data using the sum of the reads at a given base pair instead of the read position themselves, as this is the actual input data that we have. We use the EM algorithm to estimate the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$, resulting in the updates

$$\pi_{ik}^{(t)} = \frac{\sum_b Y_{ib} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b}^{(t-1)}} \right)}{\sum_{b'} Y_{ib'}} \qquad (3)$$

$$\phi_{kb}^{(t)} = \frac{\sum_i Y_{ib} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b}^{(t-1)}} \right)}{\sum_i \sum_{b'} Y_{ib'} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb'}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b'}^{(t-1)}} \right)} \qquad (4)$$

, which is iterated until convergence. However, we found (see ??) that the standard admixture model does not perform well when applied directly to sequencing data. The low (possibly 0) counts are problematic when trying to estimate the parameters, so we perform an additional "regularization" step during the each update 3 through the usage of a smoothing procedure designed specifically to smooth Poisson data, SMASH. Hence, the updates in the regularized model is given by

$$\pi_{ik}^{(t)} = \frac{\sum_b Y_{ib} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b}^{(t-1)}} \right)}{\sum_{b'} Y_{ib'}} \qquad (5)$$

$$\phi_{kb}^{(t)} = f \left(\frac{\sum_i Y_{ib} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b}^{(t-1)}} \right)}{\sum_i \sum_{b'} Y_{ib'} \left(\frac{\pi_{ik}^{(t-1)} \phi_{kb'}^{(t-1)}}{\sum_{k'} \pi_{ik'}^{(t-1)} \phi_{k'b'}^{(t-1)}} \right)} \right) \qquad (6)$$

Specifically, the function f consists of the following steps:

1. Rescale the term inside f, which we call ϕ_{kb}^o, to approximately the original scale.
2. Apply SMASH to the rescaled ϕ_{kb}^o for each k, resulting in ϕ_{kb}.
3. Scale ϕ_{kb} back so that $\sum_b \phi_{kb} = 1$.

The rationale for the scaling is because SMASH is designed to smooth Poisson data, and we can think of the ϕ_{kb}'s as a "noisy" version of the of the "true" ϕ_{kb}'s, so that scaling them gives rise to estimates that resemble a "Poisson" process. We then return the final smoothed $\boldsymbol{\phi}$'s as well as the $\boldsymbol{\pi}$'s as the parameter estimates.