Clustering SNVs for Tumor Heterogeneity

Math-CS Sc.B Thesis Proposal

David Liu

January 13, 2017

Problem statement

Cancer results from an evolutionary process where somatic mutations occur and accumulate in a population of cells. A mutation causes genetic variation at a genomic site called a single nucleotide variant (SNV) for the cell which obtained the mutation and all of its progeny. Thus, combinations of SNVs correspond to different subpopulations of cells (clones) which can be placed on a phylogenetic tree. This mixture of clones is a phenomenon known as intratumor hetereogeneity. Each clone is commonly modeled by a binary feature vector, where each feature is an SNV.

So suppose we take biopsy samples from a tumor seperated spatially or temporally. Each sample will have different proportions of clones whose relationships are unknown. Our goal is to characterize the evolutionary history of the clones (invariant across samples) and their mixing proportions (variant across samples).

The data typically used for tree inference is called the variant allele frequency (VAF), observed as follows. Suppose we bulk sequence each sample seperately, so that the reads are from a sample's mixture of clones. By comparing to a control sample, if a read contains the mutated allele at a SNV, it is called a variant read; otherwise it is called a reference read. The VAF is defined for each SNV in each sample, as, in each sample, the number of variant reads at an SNV divided by the number of total reads at that SNV.

There exist algorithms to infer a tree and its population mixing proportions given perfectly accurate VAFs. However in the real world, data is noisy and this isn't so simple. Two VAFs that look different may actually be from the same clone; likewise, two VAFs that look the same may be from different clones. Thus, it is common to perform clustering to assign mutations to clusters which correspond to clones. Tree inference is then performed on these clusters.

In this thesis, I propose and implement a method to cluster mutations using the Dirichlet Process and variational inference based on a binomial mixture model, which is suited for the clone mixing problem. It could also have applications in other clustering problems.

Model

Let $m = 1, \ldots, M$ index the samples and let $n = 1, ..., N$ index the SNVs. For each SNV n in sample m, we observe two quantities: the total number d_{mn} of reads and the number v_{mn} of variant reads. Let each data observation be a vector $\mathbf{x}_n$, a vector which encapsulates $d_{mn}, v_m n$ for all m and this n. Let $k = 1, \ldots, K$ index the clusters and let z_n denote the cluster assignment of SNV n, where z_n is a 1-of-K indicator vector. In addition, let $\mathbf{z}$ be the vector of all z_n.

Further suppose that the cluster memberships z_n and weights π_k are generated by a Dirichlet process prior. Thus K could be countably infinite, but with probability one, K is finite. Thus, we restrict K to a positive number, with all cluster indices greater than K being irrelevant.

Suppose that each clone emits variant reads according to a binomial distribution. Thus, for cluster k and some reads d_{mn}, we have variant reads distributed with $\text{Bin}(d_{mn}, \phi_{kz_{nk}})$. (**Note: May change this into a negative binomial model.**) We can consider the joint probability of a site by the product of each sample, since we assume they are independent. That is,

$$p(\mathbf{x_n}|\phi_\mathbf{n}) = \prod_{m=1}^{M} \text{Bin}(d_{mn}, \phi_{mn}) \tag{1}$$

As we can see by inspecting the graphical model, the likelihood of $\mathbf{x}$ depends on the latent variables in a straightforward way:

$$\begin{aligned} p(\mathbf{x_n}|\mathbf{z}, \phi) &= \prod_{k=1}^{K} p(\mathbf{x_n}|\phi_\mathbf{n})^{z_{nk}} \\ &= \prod_{k=1}^{K} \prod_{m=1}^{M} \text{Bin}(d_{mn}, \phi_{mn})^{z_{nk}} \end{aligned} \tag{2}$$

Now consider the joint likelihood of the observed data and latent variables, which follows from (2):

$$p(\mathbf{x_n}, \mathbf{z}|\pi, \phi_\mathbf{n}) = \prod_{k=1}^{K} \prod_{m=1}^{M} (\pi_k \text{Bin}(d_{mn}, \phi_{mn}))^{z_{nk}} \tag{3}$$

Figure 1 on the next page shows the graphical model.

$n = 1, \ldots, N$: SNVs
$m = 1, \ldots, M$: samples
$k = 1, \ldots, K$: clusters

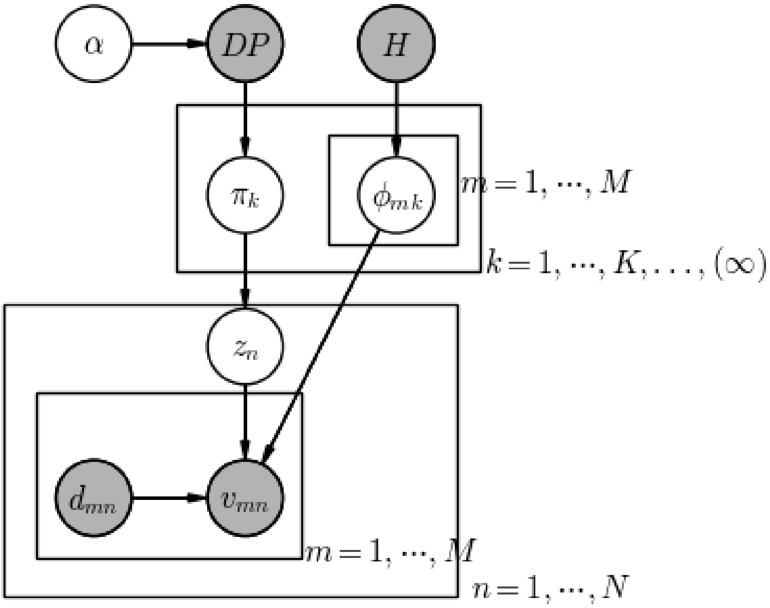

Figure 1: Graphical model for the VAFs.

DP = Dirichlet process RV (stick-breaking construction)
$H \sim U(0,1)$ = Base distribution for parameters ϕ_{nk}
π_k = Weights for categorical distribution
$z_n \in \{1, \cdots, K\} \sim \text{Categorical}(\pi_1, \cdots, \pi_K)$ = Cluster membership for SNV n
ϕ_{mk} = Cluster frequency
$v_{mn} \sim \text{Binom}(d_{m,n}, \phi_{m,c_n})$ = Observed variant reads
d_{mn} = Observed total reads

As described in (Blei 2006), the data arises in the following manner, with a stick-breaking construction given as:

$$\pi_i(\mathbf{w}) = w_i \prod_{j=1}^{i-1} (1 - v_j)$$

$$DP = \sum_{i=1}^{\infty} \pi_i(\mathbf{w}) \delta_{\phi_i}$$

1. Draw $W_k | \alpha \sim \text{Beta}(1, \alpha), \qquad k = \{1, 2, \ldots\}$
2. Draw $\phi_k | H \sim H^k, \qquad k = \{1, 2, \ldots\}$
3. For the nth data point:
   (a) Draw $z_n | \{\pi_1, \pi_2, \ldots\} \sim \text{Mult}(\pi \mathbf{w})$.
   (b) Draw $\mathbf{w_n} | z_n \sim p(x_n | \phi_n)$.

The full posterior

$$p(\mathbf{z}|\mathbf{x}, \alpha, H) = \int p(\mathbf{x}|\phi)p(\phi|\mathbf{x}, \alpha, H)\, d\phi \tag{4}$$

involves a Dirichlet Process and is thus analytically intractable. We must use some sort of computational technique to perform inference on this posterior.

Variational Inference

Variational inference is an alternative to MCMC-based inference methods. At a high level, variational inference factors a posterior using the mean-field approximation, which approximates the posterior in a higher-dimensional space using simpler independent functions. Then a simple coordinate ascent can be performed in order to infer the model parameters.

Details on Variational Inference

The ELBO

Let $\mathbf{z}$ denote the latent variables, and $\mathbf{x}$ denote the data. We seek to approximate the posterior $p(\mathbf{z}|\mathbf{x})$ from a family of distributions $\mathcal{D}$ by solving the following optimization problem:

$$q^*(z) = \arg\min_{q(\mathbf{z})\in\mathcal{D}} \textsc{kl}\left((q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))\right). \tag{5}$$

where $\textsc{kl}$ is the KL-divergence, which measures the "distance" between two distributions.

However, (4) requires us to compute the log evidence (which is intractable over the space of all $\mathbf{z}$) since

$$\textsc{kl}\left(q(\mathbf{z})\|p(\mathbf{z}\,|\,\mathbf{x})\right) = \mathrm{E}\left[\log q(\mathbf{z})\right] - \mathrm{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \log p(\mathbf{x}). \tag{6}$$

Instead, we optimize an objective function which is not dependent on $\log p(\mathbf{x})$. We call this the evidence lower bound (ELBO), which is equal to the negative KL-divergence plus the log evidence.

$$\textsc{elbo}(q) = \mathrm{E}\log p(\mathbf{z}, \mathbf{x}) - \mathrm{E}\log q(\mathbf{z}). \tag{7}$$

and thus we see that $\log p(\mathbf{x})$ is a constant with respect to q. The ELBO gets its name from the fact that it is a lower bound for the log evidence.

The mean-field variational family

And what family of distributions do we use for $\mathcal{D}$? The standard technique is to use a simple one from physics, the mean-field variational family. In this family, the latent variables $\mathbf{z}$ are mutually independent so that the joint distribution factorizes:

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j). \tag{8}$$

where q_j is a bounded variation dependent only on z_j. The structure of the model will dictate the optimal form of q_j.

Mean-field assumptions break dependency to give us coordinate ascent

The optimization is solved using a coordinate ascent algorithm, where the independence of the latent variables gives us orthogonality. Let z_{-j} denote the set of latent variables z_l such that $l \neq j$. Consider the complete conditional of z_j, which is a function of the other latent variables and the data, $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})$. Since the expectation in the ELBO is with respect to $q(\mathbf{z})$, which we have assumed factorizes, then we can dissect out the dependence with respect to $\mathbf{z}_j$ by using (6) and (7):

$$\begin{aligned}\textsc{elbo}(q) &= \int \prod q_i(\mathbf{z}_i) \left(\log p(\mathbf{z}, \mathbf{x}) - \sum_i \log q_i(\mathbf{z}_i) \right) d\mathbf{z} \\ &\propto \int q_j(z_j) \mathrm{E}_{-j} \left[\log p(\mathbf{x}, \mathbf{z}) \right] \, d\mathbf{z}_j - \int q_j(z_j) \log q_j(\mathbf{z}_j) \, d\mathbf{z}_j \end{aligned} \tag{9}$$

Now suppose that we fix z_{-j} and maximize the ELBO. Then the ELBO is maximized when $\log q_j(\mathbf{z}_j) \propto \mathrm{E}_{-j} \left[\log p(\mathbf{x}, \mathbf{z}) \right]$, by the positivity of the KL-divergence. Thus the optimal $q^*(\mathbf{z}_j)$ occurs when

$$q_j^*(\mathbf{z}_j) \propto \exp \left(\mathrm{E}_{-j} \left[\log p(\mathbf{x}, \mathbf{z}) \right] \right) \tag{10}$$

(9) underlies the coordinate-ascent variational inference algorithm. By iterating through each variational factor, fixing the others, and performing coordinate ascent (similar to Gibbs sampling), then we eventually reach a local optimum of the ELBO.

Algorithm 1: CAVI

```
Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while the ELBO has not converged do
    for $j \in \{1, \ldots, m\}$ do
        Set $q_j(z_j) \propto \exp\{\mathrm{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
    end
    Compute $\textsc{elbo}(q) = \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \mathbb{E}\left[\log q(\mathbf{z})\right]$
end
return $q(\mathbf{z})$
```

Variational Inference on Multi-sample Binomial Model

Lemma 1. *The product of two exponential family functions is also exponential.*

Proof. Exponential family functions can be parameterized by

$$f_X(x \mid \theta) = h(x) \exp \left(\theta^T \cdot T(x) - A(\theta) \right).$$

So suppose we have f_1, f_2 which are correspondingly parameterized. Note that $T(x)$ is the same because the sufficient statistic is invariant across a distribution. Then the product is

$$\begin{aligned} f_1 \times f_2 &= h_1(x) \exp\left(\theta_1^T \cdot T(x) - A(\theta_1)\right) \times h_2(x) \exp\left(\theta_2^T \cdot T(x) - A(\theta_2)\right) \\ &= \tilde{h}(x) \exp\Big((\theta_1 + \theta_2)^T \cdot T(x) - \tilde{A}(\theta_1, \theta + 2)\Big) \end{aligned}$$

which is also an exponential family distribution. □

Claim 1. *The (likelihood of* $\mathbf{x}|\mathbf{z}$*?) is also of the exponential family.*

Proof. Each binomial likelihood is exponential. Then the claim is trivial by lemma 1. □

Exponential factorizations

Exponential factorization of data model

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{z}_n) &= \prod_{m=1}^{M} \operatorname{Bin}(v_{mn}, \phi_{mn}; d_{mn}) \\ &= \prod_{m=1}^{M} \exp\left(v_{mn} \log\left(\frac{\phi_{mn}}{1-\phi_{mn}}\right) + d_{mn} \log(1-\phi_{mn})\right) \\ &= \left(\prod_{m=1}^{M} \binom{d_{mn}}{v_{mn}}\right) \exp \Bigg(\left[\log\left(\tfrac{\phi_{1n}}{1-\phi_{1n}}\right) \quad \cdots \quad \log\left(\tfrac{\phi_{Mn}}{1-\phi_{Mn}}\right)\right] \begin{bmatrix} v_{1n} \\ \vdots \\ v_{Mn} \end{bmatrix} + \left[\log\left(1-\phi_{1n}\right) \quad \cdots \quad \log\left(1-\phi_{Mn}\right)\right] \Bigg[\end{aligned} \tag{11}$$

so that the natural parameter is a vector

$$\boldsymbol{\eta} = \left[\log\left(\tfrac{\phi_{1n}}{1-\phi_{1n}}\right) \quad \cdots \quad \log\left(\tfrac{\phi_{Mn}}{1-\phi_{Mn}}\right)\right] \tag{12}$$

and sufficient statistics

$$T(\mathbf{x}_n) = \begin{bmatrix} v_{1n} \\ \vdots \\ v_{Mn} \end{bmatrix}$$

Prior exponential factorization

$$
\begin{aligned}
p(\boldsymbol{\phi}) &= \prod_{k=1}^{K} p(\boldsymbol{\phi_k}) \\
&= \prod_{k=1}^{K}\prod_{m=1}^{M} \mathrm{Beta}(a_0, b_0) && (a_0 = b_0 = \tfrac{1}{2}) \\
&= \prod_{k=1}^{K}\prod_{m=1}^{M} \frac{1}{\phi_{mk}(1-\phi_{mk})} \exp\left(\alpha_0 \log\phi_{mk} + \beta_0 \log(1-\phi_{mk}) + \log\Gamma(\alpha_0+\beta_0) - \log\Gamma(\alpha_0) - \log\Gamma(\beta_0)\right) \\
&= \left(\prod_{k=1}^{K}\prod_{m=1}^{M} \frac{1}{\phi_{mk}(1-\phi_{mk})}\right) \exp\left(\begin{bmatrix} [\alpha_0 & \beta_0] \\ \vdots & \vdots \\ [\alpha_0 & \beta_0] \end{bmatrix} \left[\begin{bmatrix} \log\phi_{1k} \\ \log(1-\phi_{1k} \end{bmatrix} \quad \cdots \quad \begin{bmatrix} \log\phi_{Mk} \\ \log(1-\phi_{Mk} \end{bmatrix}\right] + \text{the product of the be}
\end{aligned}
\tag{13}
$$

and here $a_0 = b_0 = \frac{1}{2}$ since we want the prior to be uniform over $[0, 1]$.

Coordinate ascent equations

(Note that λ is a hyperparameter such that λ_1, λ_2 chacterize the base distribution.)

We write the ELBO as:

$$
\begin{aligned}
\textsc{elbo}\,(q(\mathbf{x}, \mathbf{z}|\alpha, \lambda)) =& E_q[\log p(\mathbf{v}|\alpha)] + E_q[\log p(\phi|\lambda)] + \\
& \sum_{n=1}^{N} \left(E_q[\log p(z_n|\mathbf{v})] + E_q[\log p(x_n|z_n)]\right) \\
& - E_q[\log q(\mathbf{z}, \mathbf{v}, \phi)]
\end{aligned}
\tag{14}
$$

Suppose that the joint distribution for the last term in the ELBO factors as follows:

$$
q(\mathbf{z}, \mathbf{v}, \phi) = \underbrace{\prod_{k=1}^{K} q_{\tau_t}(\phi_{\mathbf{k}})}_{\text{Observation parameters}} \times \underbrace{\prod_{k=1}^{K} q_{\gamma_t}(\mathbf{v_k})}_{\text{Cluster proportions}} \times \underbrace{\prod_{n=1}^{N} q_{r_n}(z_n)}_{\text{Assignment of data to clusters}}
$$

where we have exponential family (uniform 0, 1), betas, and multinomials, respectively.

Note that $q(\phi_{\mathbf{k}})$ is a function of the **vector** ϕ_k, as it is the product of binomial distributions.

From here, we follow some derivations in (Blei 2006, p.129) to derive the coordinate ascent equations. The variational factor updates are actually pretty simple. For $k \in \{1, \dots, K\}$ and $n \in \{1, \dots, N\}$:

$$\gamma_{k,1} = 1 + \sum_n r_{n,k} \qquad \text{(These are the stick-breaking betas)}$$

$$\gamma_{k,2} = \alpha + \sum_n \sum_{j=k+1}^{K} r_{n,j} \qquad \text{(This is also for the stick-breaking betas)}$$

$$\tau_{k,1} = \lambda_1 + \sum_n r_{n,k} x_n$$

$$\text{(Look at exp defn on p125, this is for the cluster parameter from the base dist)}$$

$$\tau_{k,2} = \lambda_2 + \sum_n r_{n,k} \qquad \text{(Also for the cluster parameter)}$$

$$r_{n,t} \propto \exp(S_t) \qquad \text{(For the cluster assignments)}$$

where the details of calculating S_k are on page 129 of (Blei 2006).

Implementation

This will either be implemented in Python or R.

Evaluating results

The clusterings will be evaluated on simulated and real tumor data. It may also be used in conjunction with other algorithms which use such clusterings as input, to see if this improves their performance.