Mathematical Notes on Mutect

David Benjamin*
Broad Institute, 75 Ames Street, Cambridge, MA 02142
(Dated: January 25, 2017)

I. INTRODUCTION

We have a set of potential somatic alleles and read-allele likelihoods $\ell_{ra} \equiv P(\text{read } r|\text{allele } a)$. We don't know which alleles are real somatic alleles and so we must compute, for each subset $\mathbb{A}$ of alleles, the likelihood that the reads come from $\mathbb{A}$. A simple model for this likelihood is as follows: each read r is associated with a latent indicator vector $\mathbf{z}_r$ with one-hot encoding $z_{ra} = 1$ iff read r came from allele $a \in \mathbb{A}$. The conditional probability of the reads $\mathbb{R}$ given their allele assignments is

$$P(\mathbb{R}|\mathbf{z}, \mathbb{A}) = \prod_{r \in \mathbb{R}} \prod_a \ell_{ra}^{z_{ra}}. \tag{1}$$

The alleles are not equally likely because there is a latent vector $\mathbf{f}$ of allele fractions – f_a is the allele fraction of allele a. Since the components of $\mathbf{f}$ sum to one it is a categorical distribution and can be given a Dirichlet prior,

$$P(\mathbf{f}) = \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}). \tag{2}$$

Then f_a is the prior probability that a read comes from allele a and thus the conditional probability of the indicators $\mathbf{z}$ given the allele fractions $\mathbf{f}$ is

$$P(\mathbf{z}|\mathbf{f}) = \prod_r \prod_a f_a^{z_{ra}}. \tag{3}$$

The full-model likelihood is therefore

$$\mathbb{L}(\mathbb{A}) = P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A}) = \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r \left(f_a \ell_{ra}\right)^{z_{ra}}. \tag{4}$$

And the marginalized likelihood of $\mathbb{A}$, that is, the model evidence for allele subset $\mathbb{A}$, is

$$P(\mathbb{R}|\mathbb{A}) = \sum_{\mathbf{z}} \int d\mathbf{f}\, \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r \left(f_a \ell_{ra}\right)^{z_{ra}}, \tag{5}$$

where the integral is over the probability simplex $\sum_a f_a = 1$.

The integral over $\mathbf{f}$ is the normalization constant of a Dirichlet distribution and as such we can simply look up its formula. However, the sum over all values of $\mathbf{z}$ for all reads has exponentially many terms. We will get around this difficulty by handling $\mathbf{z}$ with a mean-field approximation in which we factorize the likelihood as $\mathbb{L} \approx q(\mathbf{z})q(\mathbf{f})$. This approximation is exact in two limits: first, if there are many reads, each allele is associated with many reads and therefore the Law of Large Numbers causes $\mathbf{f}$ and $\mathbf{z}$ to become uncorrelated. Second, if the allele assignments of reads are obvious $\mathbf{z}_r$ is effectively not a random variable at all (there is no uncertainty as to which of component is non-zero) and also becomes uncorrelated with $\mathbf{f}$.

In the variational Bayesian mean-field formalism the value of $\mathbf{f}$ that $\mathbf{z}$ "sees" is the expectation of $\log \mathbb{L}$ with respect to $q(\mathbf{f})$ and vice versa. That is,

$$q(\mathbf{f}) \propto \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r f_a^{\bar{z}_{ra}} \propto \text{Dir}(\mathbf{f}|\boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r) \tag{6}$$

and

$$q(\mathbf{z}_r) = \prod_a \left(\tilde{f}_a \ell_{ra}\right)^{z_{ra}}, \tilde{f}_a = \exp E[\ln f_a] \tag{7}$$

*Electronic address: davidben@broadinstitute.org

Because $q(\mathbf{z})$ is categorical and $q(\mathbf{f})$ is Dirichlet[1] the necessary mean fields are easily obtained and we have

$$\bar{z}_{ra} = \frac{\tilde{f}_a \ell_{ra}}{\sum_{a'} \tilde{f}_{a'} \ell_{ra'}} \tag{8}$$

and

$$\ln \tilde{f}_a = \psi(\alpha_a + \sum_r \bar{z}_{ra}) - \psi(\sum_{a'} \alpha_{a'} + N) \tag{9}$$

where ψ is the digamma function and N is the number of reads. To obtain $q(\mathbf{z})$ and $q(\mathbf{f})$ we iterate Equations 8 and 9 until convergence. A very reasonable initialization is to set $\bar{z}_{ra} = 1$ if a is the most likely allele for read r, 0 otherwise. Having obtained the mean field of $\mathbf{z}$, we plug it into Eq 5 to obtain[2]

$$P(\mathbb{R}|\mathbb{A}) = \int d\mathbf{f}\, \mathrm{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r (f_a \ell_{ra})^{\bar{z}_{ra}} \tag{10}$$

$$= \frac{\Gamma(\sum_a \alpha_a)}{\prod_a \Gamma(\alpha_a)} \prod_{ra} \ell_{ra}^{\bar{z}_{ra}} \int d\mathbf{f} \prod_a f_a^{\alpha_a + \sum_r \bar{z}_{ra} - 1} \tag{11}$$

$$= \frac{\Gamma(\sum_a \alpha_a)}{\prod_a \Gamma(\alpha_a)} \prod_{ra} \ell_{ra}^{\bar{z}_{ra}} \frac{\prod_a \Gamma(\alpha_a + \sum_r \bar{z}_{ra})}{\Gamma(\sum_a \alpha_a + N)} \tag{12}$$

We now have the model evidence for allele subset $\mathbb{A}$. This lets us choose which alleles are true somatic variants. It also lets us make calls on somatic loss of heterozygosity events. Furthermore, instead of reporting max-likelihood allele fractions as before, we may emit the parameters of the Dirichlet posterior $q(\mathbf{f})$, which encode both the maximum likelihood allele fractions and their uncertainty.

[1] Note that we didn't *impose* this in any way. It simply falls out of the mean field equations.
[2] Because we are doing model comparison of different $\mathbb{A}$, we do not drop constant multiplicative factors below.