Allele frequency estimation under alignment and typing uncertainty
A general latent variable framework for genotyping, de novo and somatic mutation calling

Louis Dijkstra*

November 17, 2014

Abstract

1 Introduction

1.1 (Next-generation) sequence data

1.2 Notation and nomenclature

Manifest and latent variables are denoted with, respectively, Latin capital letters and Greek symbols. Observations/realizations of manifest variables are represented by a small letter.

2 The general model

In this section we present a (Bayesian) latent variable model that links (next-generation sequence) reads and their alignments with the allele frequency of a variant of interest. We made the choice for a latent variable model for mainly two reasons:

1. the allele frequency is a quantity that cannot be observed directly and representing it, therefore, as a hidden/latent variable seems to be natural;

2. a latent variable model allows for incorporating measurement uncertainties in a clear way [REF].

*E-mail: dijkstra@cwi.nl

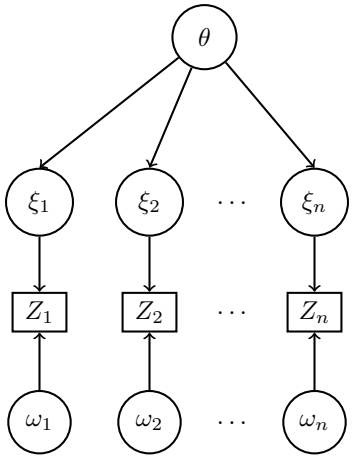

Figure 1: The general latent variable model. Latent variables are shown with circles, while manifest variables are represented by boxes.

The variant allele frequency (VAF) Let $\theta \in \Theta \subset [0, 1]$ be a latent random variable representing the main quantity of interest: the variant allele frequency (VAF) of a particular indel. We will assume that the locus of the indel is given beforehand and that its length, denoted by δ, is known. We will for the moment assume a little as possible about θ: its supports can either be discrete[1] or continuous (which will depend on the application at hand, see Sect. XXX). In addition, we assume a uniform (hyper-)prior:

$$\theta \sim \text{uniform over } \Theta. \tag{2.1}$$

Alignments Let $R = \{R_1, R_2, \ldots, R_n\}$ be a sequence of reads and $A = \{A_1, \ldots, A_n\}$ be the corresponding set of alignments to the locus of interest provided by an aligner. Each alignment is associated with an *alignment probability* denoted with π_i, i.e., the chance that alignment A_i is the origin of the i-th read.

Let $Z' = (Z_1, Z_2, \ldots, Z_n) \in \mathbb{R}^n$ be a vector of manifest variables representing observed quantities derived from n different alignments, i.e.,

$$Z_i := w_i(A_i) \qquad \text{for } i = 1, 2 \ldots, n. \tag{2.2}$$

We will for the moment not specify the functions $w_i(\cdot)$; this depends on the type of read and how it is aligned to the locus of interest. For now we will limit ourselves to describe how these observations can be linked to our original quantity of interest, the VAF θ. We do this to show how the model presented here can be easily adapted to accomodate a wide variety of reads/alignments and observations. Later on we will show how the measurement model can applied specifically to internal segment lengths and overlapping alignments.

[1] In the case of a diploid individual, for example, the support of the VAF is $\Theta = \{0, 1/2, 1\}$, since none, only one or both of the chromosomal copies can harbour the variant allele.

As mentioned before, any observation based on an alignment is commonly associated with two types of uncertainty: 1) alignment uncertainty (the alignment could be placed erroneously), and 2) typing uncertainty (the alignment can only support the presence/absence of a genetic variant up to a certain level). These two types of uncertainties are represented in our model by associating every observation Z_i with two Bernoulli distributed latent variables: ω_i and ξ_i. The latent variable ω_i is 1 iff the i-th alignment is placed correctly and 0 otherwise. The probability of alignment i to be correctly aligned is commonly provided by the aligner and denoted here with π_i:

$$\omega_i \sim \text{Bernoulli}\,(\pi_i) \qquad \text{for } i = 1, 2 \ldots, n. \tag{2.3}$$

Note that when an alignment Z_i only provides information about θ when it stems from the locus of interest, i.e., $\omega_i = 1$. The binary variable ξ_i is 1 iff the i-th alignment stems from a chromosome that harbours the indel of interest and 0 otherwise. The probability that a read stems from a variant-allele-affected chromosome is equal to the variant allele frequency[2], θ, i.e.,

$$\xi_i \mid \theta \sim \text{Bernoulli}\,(\theta) \qquad \text{for } i = 1, 2 \ldots, n. \tag{2.4}$$

Whether ξ_i is 1 or 0 is often not evident from the value of Z_i, although the values Z_i is likely to take will vary when the indel is present or not. More formally,

$$\begin{aligned} Z_i \mid \omega_i = 1, \xi_i = 0 &\sim f_i^{(\text{absent})} \\ Z_i \mid \omega_i = 1, \xi_i = 1 &\sim f_i^{(\text{present})} \end{aligned} \tag{2.5}$$

for $i = 1, 2, \ldots, n$. The probability distributions $f_i^{(\text{absent})}$ and $f_i^{(\text{present})}$ returns the probability of observing a certain realization Z_i when the indel of interest is either absent or present. When the i-th alignment does not stem from the locus of interest, i.e., $\omega_i = 0$:

$$Z_i \mid \omega_i = 0 \sim u_i \qquad \text{for } i = 1, 2 \ldots, n. \tag{2.6}$$

The probability distribution u_i is arbitrary and should be chosen such that the posterior distribution of the VAF θ is not influenced. We will return to this point later.

The probability distribution of Z_i given the VAF θ, $g_i(z_i \mid \theta)$, can easily be found by applying the law of total probability repeatedly:

$$\begin{aligned} g_i(z_i \mid \theta) &= g_i(z_i \mid \theta, \omega_i = 1)\mathbb{P}\left\{\omega_i = 1\right\} + g_i(z_i \mid \theta, \omega_i = 0)\mathbb{P}\left\{\omega_i = 0\right\} \\ &= \pi_i g_i(z_i \mid \theta, \omega_i = 1) + (1 - \pi_i)\, u_i(z_i). \end{aligned} \tag{2.7}$$

The probability distribution $g_i(z_i \mid \theta, \omega_i = 1)$ is given by

$$\begin{aligned} g_i(z_i \mid \theta, \omega_i = 1) &= g_i(z_i \mid \theta, \omega_i = 1, \xi_i = 1)\mathbb{P}\left\{\xi_i = 1 \mid \theta\right\} \\ &\quad + g_i(z_i \mid \theta, \omega_i = 1, \xi_i = 0)\mathbb{P}\left\{\xi_i = 0 \mid \theta\right\} \\ &= \theta f_i^{(\text{present})}(z_i) + (1 - \theta) f_i^{(\text{absent})}(z_i). \end{aligned} \tag{2.8}$$

[2]Or at least, under the assumption that a chromosome carrying the indel is equally likely to bring forth a read as a chromosome not harboring the indel.

In conclusion,

$$g_i(z_i \mid \theta) = \pi_i \left[\theta f_i^{(\text{present})}(z_i) + (1-\theta)\, f_i^{(\text{absent})}(z_i)\right] + (1-\pi_i)\, u_i(z_i). \tag{2.9}$$

The posterior distribution of the VAF θ The posterior distribution of the VAF θ is given a vector of observations z:

$$h(\theta \mid Z) = \frac{f(z \mid \theta) h(\theta)}{f(z)} \qquad \text{(Bayes' rule)} \tag{2.10}$$

where $f(z \mid \theta)$ is the likelihood of the data given θ, $h(\cdot)$ is the VAF's (uniform) prior and $f(z)$ is $\sum_{\theta \in \Theta} f(z \mid \theta) h(\theta)$ and $\int_{\theta \in \Theta} f(z \mid \theta) h(\theta) d\theta$, when the support of θ is, respectively, discrete and continuous. Under the assupmtion of local/conditional independence of the manifest variables $Z_1, Z_2, \ldots, Z_n$, the likelihood is linear in the number of aligments:

$$\begin{aligned} L(\theta \mid z) &= f(z_1, z_2, \ldots, z_n \mid \theta) \\ &= \prod_{i=1}^{n} g_i(z_i \mid \theta) \\ &= \prod_{i=1}^{n} \left\{ \pi_i \left[\theta f_i^{(\text{present})}(z_i) + (1-\theta)\, f_i^{(\text{absent})}(z_i)\right] + (1-\pi_i)\, u_i(z_i) \right\}. \end{aligned} \tag{2.11}$$

Note that the likelihood function is a n-degree polynomial in θ. Determining its maximum analytically when θ's support is continuous, is, therefore, impossible.

Theorem 1. *The likelihood function $L(\theta \mid z_1, \ldots, z_n)$ attains a unique global maximum $\widehat{\theta}$ on the unit interval $\Theta = [0, 1]$ when (1) the subset*

$$I := \{\theta \in \Theta : g_i(z_i \mid \theta) > 0 \text{ for } i = 1, \ldots, n\} \tag{2.12}$$

is connected and non-empty, and, (2) there exists an observation z_i for which the alignment probability π_i is strictly larger than zero and $f_i^{(present)}(z_i) \neq f_i^{(absent)}(z_i)$.

Proof. Note that $L(\theta \mid z_1, \ldots, z_n) = \prod_i g_i(z_i \mid \theta) = 0$ when $\theta \notin I$, since for those θ's there exists an observation z_i for which the likelihood $g_i(z_i \mid \theta)$ is equal to zero. The likelihood $L(\theta \mid z_1, \ldots, z_n)$ is strictly larger than zero by definition when $\theta \in I$. Since the likelihood function is a n-th order polynomial and, therefore, continuous, it must attain a global maximum on the interval I.

Suppose condition *(1)* is met. The point $\widehat{\theta} \in I$ is a maximum of $L(\cdot \mid z_1, \ldots, z_n)$ if and only if it is a maximum of the loglikelihood function

$$\ell(\theta \mid z_1, \ldots, z_n) := \log L(\theta \mid z_1, \ldots, z_n) = \sum_{i=1}^{n} \log g_i(z_i \mid \theta) \qquad (\theta \in I) \tag{2.13}$$

since the logarithm is a monotonic transformation. Note that the loglikelihood function is only defined on the subset I. The second order derivative of the loglikelihood with respect to the VAF is easily found to be

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\sum_{i=1}^{n} \left[\frac{\partial g_i(z_i \mid \theta)/\partial \theta}{g_i(z_i \mid \theta)}\right]^2 \leq 0 \tag{2.14}$$

indicating that the loglikelihood function is concave. Note that it is strictly concave, i.e., $\partial^2\ell/\partial\theta^2 < 0$, iff there exists an observation z_i for which

$$\frac{\partial g_i(z_i \mid \theta)}{\partial\theta} = \pi_i \left[f_i^{(\text{present})}(z_i) - f_i^{(\text{absent})}(z_i) \right] \neq 0. \tag{2.15}$$

This inequality holds only when $\pi_i \neq 0$ and $f_i^{(\text{present})}(z_i) \neq f_i^{(\text{absent})}(z_i)$ which constitutes condition *(2)*.

Suppose I is a non-empty closed set on the unit interval, i.e., $I = [a, b]$. Since the loglikelihood is strictly concave when condition *(2)* is met, it attains a unique global maximum $\widehat{\theta}$ on I. Because the logarithm is a monotonic transform, $\widehat{\theta}$ must be the unique global maximum of the likelihood function as well.

A similar reasoning holds up when I is open or half-open. The maximum must lie in the interior of I, since the likelihood function is zero for those endpoints not in I, e.g., when I is the open interval (a, b), $L(a \mid z_1, \ldots, z_n) = L(b \mid z_1, \ldots, z_n) = 0$ while $L(\theta \mid z_1, \ldots, z_n)$ is strictly positive on I. The loglikelihood function is strictly concave on I, therefore, the likelihood function attains a unqiue global maximum. □

The functions $u_i(\cdot)$ must be chosen such that

1. when $\pi_i = 0$, i.e., the read in question does with 100% certainty not align to the locus of interest, then the observation z_i should not provide any information on the VAF θ, i.e., $h(\theta \mid z_i) = h(\theta)$.

The first condition is met for any probability distribution $u_i(\cdot)$ that does not depend on θ, since the posterior distribution $h(\theta \mid z_i)$ when $\pi_i = 0$ is equal to

$$h(\theta \mid z_i) = \frac{u_i(z_i)h(\theta)}{\int_\Theta u_i(z_i)h(\theta)d\theta} = \frac{u_i(z_i)h(\theta)}{u_i(z_i)\int_\Theta h(\theta)d\theta} = h(\theta). \tag{2.16}$$

(The derivation is similar when the support of θ is discrete).

2.1 Alignments

Let $Z \in \mathbb{R}$ be a manifest variable denoting an observation based on an alignment. Every alignment is surrounded with uncertainty.