

Ultra-sensitive variant detection with improved accuracy using variational inference for heterogeneous next-generation sequencing data

Fan Zhang

Worcester Polytechnic Institute, MA
fzhang@wpi.edu

Abstract

We propose a variational inference algorithm to estimate variant allele frequency (VAF) and identify single nucleotide variants in heterogeneous next-generation sequencing data. We demonstrate our variational algorithm with higher sensitivity and specificity than Markov Chain Monte Carlo (MCMC) sampling method on a synthetic data set, and more efficient at relative low median read depths of 10x and 100x. We apply our algorithm on a longitudinal anti-cancer drug resistance sequencing data set and identify XXX variants that XXX. We also show that our model with variational algorithm has an improved performance in XXX on a longitudinal clinical data set compared with the state of arts approaches.

1 Introduction

Next-generation sequencing (NGS) data and single nucleotide variants (SNVs). xxx

xxx
xxx
xxx
xxx
xxx
xxx
xxx

Variants detection methods and problems. xxx

xxx
xxx
xxx
xxx
xxx
xxx
xxx

Recently, some algorithms have been developed Bayesian models to identify SNVs in NGS data. The main computational problem for probabilistic Bayesian models is posterior inference. There are two popular ways to do posterior inference- Markov Chain Monte Carlo (MCMC) samplings, and variational algorithms. MCMC is a simple algorithm for

sampling, which is easy to understand. The main problem of MCMC is that it is hard to diagnose convergence and handle nonconjugate events. The main idea of variational inference is propose a variational distribution over latent variables and optimize the variational parameters to make the proposal distribution close to the true posterior distribution. So variational inference is more accurate to handle nonconjugate distributions and more efficient than MCMC. RVD2 uses a hierarchical Bayesian model to estimate allele frequency and call variants by deriving a Metropolis-within-Gibbs sampling inference algorithm [cite RVD2]. Generally speaking, this algorithm is able to accurately estimate variant allele frequencies (VAFs), but it shows a relatively low specificity when the variant allele frequency is low (less than 1.0%) but the median read depth is very high (40000x).

We show that we develop a variational expectation-maximization (EM) algorithm for our Bayesian statistical model to achieve sufficient accuracy and efficiency to identify variants in heterogeneous call samples. First, our variational EM algorithm is able to accurately approximate the posterior distribution of latent variables for a pair of samples- the sample of interest and a known reference sample. Then, a hypothesis test calls a variant by a significant difference between the key model parameters of this pair of samples. We compare the performance of variational inference algorithm to MCMC sampling method and several other variant detection methods. Finally, we demonstrate our variational algorithm on a longitudinal time-series DNA sequencing data to quantify the degree to which the variant modulates resistance to anti-cancer drug.

2 Model Structure

Our Bayesian statistical model is shown as a graphical representation in Figure 1. r_{ji} is the number of reads with a non-reference base at location j in experimental replicate i . n_{ji} is the total number of reads at location j in experimental replicate i . The model generative process is as follows:

1. For each location j :
 - (a) Draw an error rate $\mu_j \sim \text{Beta}(\mu_0, M_0)$
 - (b) For each replicate i :
 - i. Draw $\theta_{ji} \sim \text{Beta}(\mu_j, M_j)$
 - ii. Draw $r_{ji}|n_{ji} \sim \text{Binomial}(\theta_{ji}, n_{ji})$

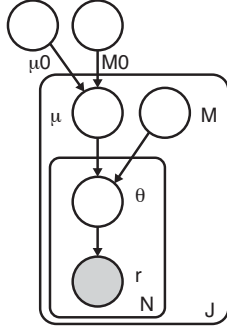


Figure 1: Graphical model representation of our Bayesian model. μ_0 , a global error rate; M_0 , a global precision; μ_j , a local error rate. M_j , a local precision. μ_0 , a global error rate to estimate the expected error rate across all locations. M_0 , a global precision, estimates the variation in the error rate across locations. The local error rate, μ_j , estimates the expected error rate across replicates at location j . The local precision, M_j , estimates the variation in the error rate across replicates at location j

3 Inference and Hypothesis Testing

3.1 Variational Expectation Maximization (EM) Inference

RVD3 improves RVD2 in the way of posterior distribution inference. We develop a non-conjugate variational inference algorithm to approximate the posterior distribution,

$$p(\mu, \theta | r, n; \phi) = \frac{p(\mu, \theta, r | n; \phi)}{p(r | n; \phi)}, \quad (1)$$

where the parameters are $\phi \triangleq \{\mu_0, M_0, M\}$.

Factorization

We propose the following factorized variational distribution to approximate the true posterior over latent variables μ_j and θ_{ij} . $q(\mu_j)$ approximates the variational posterior distribution of μ_j , which represents the local error rate distribution in position j across different replicates. $q(\theta_{ij})$ approximates the posterior distribution of θ_{ij} , which is the error rate distribution in position j replicate i .

$$q(\mu, \theta) = q(\mu)q(\theta) = \prod_{j=1}^J q(\mu_j) \prod_{i=1}^N q(\theta_{ji}). \quad (2)$$

Evidence Lower Bound (ELBO)

The log-likelihood of the data is lower-bounded according to Jensen's inequality:

$$\begin{aligned} \log p(r | \phi) &= \log \int_{\mu} \int_{\theta} p(r, \mu, \theta) d\theta d\mu \\ &= \log \int_{\mu} \int_{\theta} p(r, \mu, \theta) \frac{q(\mu, \theta)}{q(\mu, \theta)} d\theta d\mu \\ &\geq \int_{\mu} \int_{\theta} q(\mu, \theta) \log \frac{p(r, \mu, \theta)}{q(\mu, \theta)} d\theta d\mu \\ &= E_q[\log p(r, \mu, \theta)] - E_q[\log q(\mu, \theta)] \\ &\triangleq \mathcal{L}(q, \phi). \end{aligned} \quad (3)$$

where $\phi = (\mu_0, M_0, M)$.

The item $\mathcal{L}(q, \phi)$ is the evidence of lower bound (ELBO) of the log-likelihood of the data, which is the sum of q -expected complete log-likelihood and the entropy of the variational distribution q . The goal of variational inference is maximizing the ELBO. Equivalently, q is chosen by minimizing the Kullback-Liebler (KL) divergence between the variational distribution and the true posterior distribution.

Variational Distributions

The posterior distribution of θ_{ji} is a Beta distribution,

$$p(\theta_{ji} | r_{ji}, n_{ji}, \mu_j, M_j) \quad (4)$$

$$\sim \text{Beta}(r_{ji} + M_j \mu_j, n_{ji} - r_{ji} + M_j(1 - \mu_j)). \quad (5)$$

Therefore, we propose Beta distribution with parameter vector δ_{ji} as variational distribution,

$$\theta_{ji} \sim \text{Beta}(\delta_{ji})$$

The posterior distribution of μ_j is given by its Markov blanket

$$p(\mu_j | \theta_{ji}, M_j, \mu_0, M_0) \propto p(\mu_j | \mu_0, M_0) p(\theta_{ji} | \mu_j, M_j). \quad (6)$$

This is not in the form of any known distribution. Therefore, we propose Beta distribution with parameter vector γ_{ji} as variational distribution to simplify the variational derivation.

$$\mu_j \sim \text{Beta}(\gamma_j)$$

Variational Expectation Maximization (EM) Algorithm

Variational EM maximizes the ELBO on the true likelihood, by alternating between maximization over q (E-step) and maximization over ϕ (M-step).

Algorithm 1 RVD3 Variational Inference

- 1: Initialize $q(\theta, \mu)$ and $\hat{\phi}$
 - 2: **repeat**
 - 3: **repeat**
 - 4: **for** $j = 1$ to J **do**
 - 5: **for** $i = 1$ to N **do**
 - 6: Optimize $\mathcal{L}(q, \hat{\phi})$ over $q(\theta_{ji}; \delta_{ji}) = \text{Beta}(\delta_{ji})$
 - 7: **end for**
 - 8: **end for**
 - 9: **for** $j = 1$ to J **do**
 - 10: Optimize $\mathcal{L}(q, \hat{\phi})$ over $q(\mu_j; \gamma_j) = \text{Beta}(\gamma_j)$
 - 11: **end for**
 - 12: **until** change in $\mathcal{L}(q, \hat{\phi})$ is small
 - 13: Set $\hat{\phi} \leftarrow \arg \max_{\phi} \mathcal{L}(q, \phi)$
 - 14: **until** change in $\mathcal{L}(q, \hat{\phi})$ is small
-

4 Posterior Distribution Test

4.1 Z-test for Gaussian Distribution

Variational inference provides variational distributions for $q(\mu_j | r^{control})$ and $q(\mu_j | r^{case})$, which are approximated to

the posterior distributions for $p(\mu_j|r^{control})$ and $p(\mu_j|r^{case})$. We use Z -test based on Gaussian distribution.

$$\mu_j^\Delta = \mu_j|r^{case} - \mu_j|r^{control} \quad (7)$$

$$\sigma_j^\Delta = \sqrt{var_q[\mu_j|r^{case}] + var_q[\mu_j|r^{control}]} \quad (8)$$

$$Z_j = \frac{\tau - \mu_j^\Delta}{\sigma_j^\Delta} \quad (9)$$

$$Pr(Z_j) < \alpha \quad (10)$$

where τ is a detection threshold and α is a significance level. Here we set $\tau = 0$ and $\alpha = 0.05$. μ_j^{case} and $\mu_j^{control}$ are means of distributions. $var_q[\mu_j|r^{case}]$ and $var_q[\mu_j|r^{control}]$ are variances of distributions. A position is called as a variant when the p-value is less than the significant level α .

χ^2 goodness-of-fit test for non-uniform base distribution is also used to distinguish a scenario of a true variant from a scenario of a random sequencing error [cite RVD2].

5 Data Sets

5.1 Synthetic DNA Sequence Data

We synthesized two random 400bp DNA dequences. One sample with 14 loci variants is taken as the case and the other without variants is taken as the control. Case and control samples were mixed to yield defined VAFs of 0.1%, 0.3%, 1.0%, 10.0%, and 100.0%. The details of the experimental protocol are available in [cite RVD]. The synthetic DNA data were downsampled by 10x, 100x, 1,000x, and 10,000x using Picard. The final data set contains read pairs for 6 replicates for the control at different VAF levels.

5.2 Longitudinal Drug Resistance Data

xxx
xxx
xxx
xxx
xxx
xxx
xxx
xxx
xxx
xxx
xxx

6 Results

6.1 Performance on Synthetic DNA Data

Comparison of Sensitivity and Specificity

The posterior distribution over latent variables can be inferred using either MCMC sampling method or variational algorithm. We compare the performance of variational algorithm and MCMC sampling method in the performance of sensitivity and specificity (Figure 1). We use posterior distribution test with χ^2 test to detect variants for a broad range

of median read depths and different variant allele frequencies (VAFs). Variational algorithm works best compared with MCMC in sensitivity and specificity. The converging process for estimating variational posterior for VAF=100.0% at median read depth of 30590x is shown in (Figure 2). It takes 60 variational EM iterations to stop converging when the increased ELBO rate is less than 0.1%.

| VAF | Median Depth | Sensitivity | | Specificity | |
|--------|--------------|-------------|-------------|-------------|-------------|
| | | MCMC | Variational | MCMC | Variational |
| 0.1% | 39 | 0.00 | 0.00 | 1.00 | 1.00 |
| | 408 | 0.00 | 0.07 | 1.00 | 1.00 |
| | 4129 | 0.14 | 0.29 | 1.00 | 1.00 |
| | 41449 | 0.86 | 1.00 | 0.97 | 1.00 |
| 0.3% | 36 | 0.00 | 0.00 | 1.00 | 1.00 |
| | 410 | 0.00 | 0.00 | 1.00 | 1.00 |
| | 4156 | 1.00 | 1.00 | 0.99 | 0.98 |
| | 41472 | 1.00 | 0.93 | 0.85 | 0.91 |
| 1.0% | 53 | 0.00 | 0.00 | 1.00 | 1.00 |
| | 535 | 0.21 | 0.29 | 1.00 | 1.00 |
| | 5584 | 1.00 | 1.00 | 0.98 | 0.98 |
| | 55489 | 1.00 | 0.93 | 0.87 | 0.95 |
| 10.0% | 22 | 0.00 | 0.57 | 1.00 | 1.00 |
| | 260 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 2718 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 26959 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0% | 27 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 298 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 3089 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30590 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 1: Sensitivity/Specificity comparison of variational algorithm with MCMC on the synthetic DNA data set.

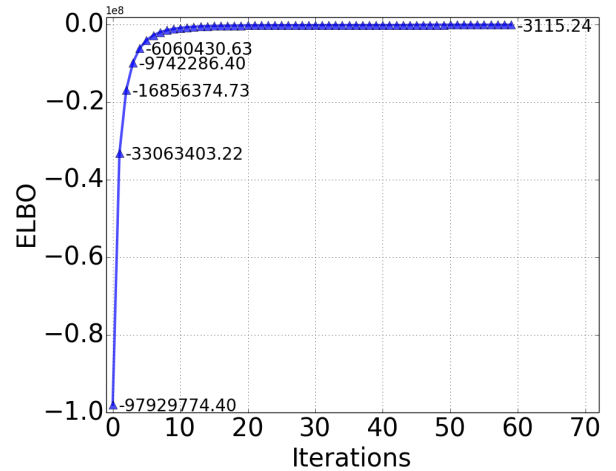


Figure 2: Converging ELBO for estimating variational posterior in the event of VAF=100.0% at median read depth of 30590x.

Comparison of Approximated Posterior Distribution

We show the approximated posterior distribution of variational algorithm and exact samples of MCMC. A true variant position 85 of sample of VAF=1.0% is taken as an example. Variational and MCMC both identify this position at median

Median depth is 5584

Case (Variational)
Control (Variational)
Case (MCMC)
Control (MCMC)

Median depth is 41449

Case (Variational)
Control (Variational)
Case (MCMC)
Control (MCMC)

Comparison of Timing

Timing profile for each parts of variational algorithm when median read depth is 3089x is also given in Figure 5. Optimizing γ function in E-step and optimizing M in M-step takes most of the time because an integration is needed.

Detected Drug Resistance Variants

A line graph comparing the convergence time (in minutes) of four variational methods (27x, 298x, 3089x, and 30590x Variational) against the length of the region of interest (in positions). The x-axis ranges from 100 to 400 positions, and the y-axis ranges from 0 to 120 minutes. The 30590x Variational method shows the highest convergence time, increasing from approximately 32 minutes at 100 positions to 110 minutes at 400 positions. The 3089x Variational method follows, increasing from about 27 minutes to 73 minutes. The 298x Variational method increases from about 4 minutes to 18 minutes. The 27x Variational method increases from about 3 minutes to 12 minutes. The MCMC method, represented by a dashed line with star markers, shows the lowest convergence time, increasing from about 6 minutes to 22 minutes.

| Length of region of interest (positions) | 27x Variational | 298x Variational | 3089x Variational | 30590x Variational | MCMC |
|--|-----------------|------------------|-------------------|--------------------|------|
| 100 | 3 | 4 | 27 | 32 | 6 |
| 200 | 5 | 8 | 36 | 40 | 13 |
| 300 | 9 | 14 | 52 | 73 | 16 |
| 400 | 12 | 18 | 73 | 110 | 22 |

Comparison to the State of Arts Approaches

6.3 Discussion

- ## Acknowledgments

References

b

| Computation resource | Region length | Load depth | Initialization | | E-step | | | M-step | | | | Save model | Total time (s) |
|----------------------|---------------|------------|----------------|--------|-------------------|-------------------|-------------|------------------|----------------|--------------|-------------|------------|----------------|
| | | | Parameters | ELBO | Optimize γ | Optimize δ | Update ELBO | Optimize μ_0 | Optimize M_0 | Optimize M | Update ELBO | | |
| single processor | 100 | 0.023 | 0.001 | 10.522 | 617.705 | 4.232 | 10.415 | 0.264 | 0.159 | 332.898 | 10.293 | 0.025 | 986.537 |
| | 200 | 0.059 | 0.001 | 20.961 | 1124.818 | 8.936 | 18.644 | 0.418 | 0.256 | 569.952 | 18.374 | 0.025 | 1762.444 |
| | 300 | 0.073 | 0.001 | 31.102 | 1728.444 | 13.265 | 27.807 | 0.445 | 0.400 | 851.467 | 27.650 | 0.025 | 2680.679 |
| | 400 | 0.127 | 0.001 | 42.727 | 2433.169 | 17.987 | 38.552 | 0.737 | 0.635 | 1176.250 | 38.171 | 0.034 | 3748.390 |
| 60 processors | 100 | 0.026 | 0.001 | 11.535 | 29.939 | 0.247 | 11.674 | 0.307 | 0.189 | 19.561 | 11.979 | 0.025 | 85.483 |
| | 200 | 0.036 | 0.001 | 24.184 | 44.697 | 0.417 | 22.139 | 0.523 | 0.304 | 24.036 | 22.235 | 0.026 | 138.598 |
| | 300 | 0.040 | 0.001 | 35.480 | 63.470 | 0.716 | 33.306 | 0.562 | 0.504 | 29.410 | 33.238 | 0.027 | 196.754 |
| | 400 | 0.058 | 0.001 | 46.325 | 94.660 | 0.727 | 42.783 | 0.820 | 0.706 | 35.742 | 44.283 | 0.027 | 266.132 |

Figure 5: Timing profile for one iteration of variational EM algorithm when median read depth is 3089x. Single and multiple processors are both tested for timing to estimate the model on the synthetic data set.