LDSP - Linear Detection of Selection in Pooled data

Hussein Al-Asadi, Matthew Stephens

1 Introduction

Phase I - better estimate frequency using haplotypic information: The Intuition is that data at each SNP are binomial counts, which help estimate the frequency of a SNP in a pool. But by combining information across multiple corrected SNPs, you can improve the estimated frequency of the test SNP.

We demonstrate the feasibility of such an endeavor by a simple probability calculation: Say the objective is to estimate the frequency of SNP 1 in a pool. Denote SNP 1 as S_1 which can take values from $\{0, 1\}$.

Consider another SNP 2 (S_2). For simplicity, we suppose perfect correlation between the two SNPs (e.g. $P(S_1 = 1|S_2 = 1) = 1$ & $P(S_1 = 0|S_2 = 0) = 1$)

one estimate: $P(S_1 = 1) \approx \frac{n_1^1}{n_1}$
second estimate: $P(S_1 = 1) = P(S_1 = 1|S_2 = 0)P(S_2 = 0) + P(S_1 = 1|S_2 = 1)P(S_2 = 1) = P(S_1 = 1|S_2 = 1)P(S_2 = 1) = P(S_2 = 1) \approx \frac{n_2^1}{n_2}$

where n_j^1 is the number of 1 allele reads at SNP j and n_j^0 is the number of 0 allele reads at SNP j and $n_j = n_j^1 + n_j^0$.

We now have two estimates of $P(S_1 = 1)$ using two different pieces of data. Therefore, effectively we have doubled our coverage for SNP 1. If instead of only one perfectly correlated SNP, we have a 1000 then sequencing only at 1x coverage will be like sequencing at 1000x coverage!

Phase II - detect selection using improved estimates: To detect selection, we find sites that have had significant changes in their frequency compared to the founding population. We can do this by using a linear model which also allows us to model genetic drift with a normal error term.

2 Phase I - calculating effective coverage

2.1 Prior from Li & Stephens

Consider one lineage for now.

Let $y = (y_1, y_2, ..., y_p)'$ denote the vector of allele frequencies in the study sample. Let $E[y_i] = \mu_i$ and the frequency of the test SNP be y_t and M denote the $2mxp$ panel (i.e. $2m$ haplotypes and p SNPs). As in (Wen & Stephens, 2010), we assume

$$\vec{y}|M \sim N_p(\mu, \Sigma) \tag{1}$$

(Wen & Stephens, 2010) derived the estimates for μ and Σ from the haplotype copying model presented in (Li & Stephens, 2003).

$$\hat{\mu} = (1-\theta)f^{panel} + \frac{\theta}{2}1 \tag{2}$$

$$\hat{\Sigma} = (1-\theta)^2 S + \frac{\theta}{2}(1-\frac{\theta}{2})I \tag{3}$$

and S is obtained from Σ^{panel}, specifically,

$$S_{i,j} = \begin{cases} \Sigma^{panel}_{i,j} & i = j \\ e^{-\frac{-\rho_{i,j}}{2m}}\Sigma^{panel}_{i,j} & i \neq j \end{cases} \tag{4}$$

$\rho_{i,j} = -4Nc_{i,j}d_{i,j}$ where $d_{i,j}$ is the physical distance between markers i and j, N is the effective diploid population size, $c_{i,j}$ is the average rate of crossover per unit physical distance, per meiosis, between sites i and j (so that $c_{i,j}d_{i,j}$ is the genetic distance between sites i and j).

and,

$$\theta = \frac{(\sum_{i=1}^{2m-1}\frac{1}{i})^{-1}}{2m + (\sum_{i=1}^{2m-1}\frac{1}{i})^{-1}} \tag{5}$$

2.2 Data at SNP i

Let (n_i^0, n_i^1) denote the counts of 0 and 1 alleles at SNP i and $n_i = n_i^0 + n_i^1$. Then

$$n_i^1|y_i \sim Bin(n_i, y_i) \dot{\sim} N(n_i y_i, n_i y_i(1-y_i))$$

where y_i is the true population frequency of the SNP i 1 allele.

$$\implies \frac{n_i^1}{n_i}|y_i \dot{\sim} N(y_i, \frac{y_i(1-y_i)}{n_i}) \tag{6}$$

let $\hat{y}_i = \frac{n_i^1}{n_i}$
Next we replace y_i by $\hat{y}_i$ in the variance for tractibility issues. Therefore,

$$\hat{y}_i|y_i \dot{\sim} N(y_i, \frac{\hat{y}_i(1-\hat{y}_i)}{n_i}) \tag{7}$$

We can expand equation 7 to p-dimensions (we can since the $\hat{y}_i|y_i$ are independent)

$$\hat{\vec{y}}|\vec{y} \dot{\sim} N_p(\vec{y},\ diag(\epsilon_1, ..., \epsilon_p)) \tag{8}$$

where $\epsilon_i = \frac{\hat{y}_i(1-\hat{y}_i)}{n_i}$

We re-name the variables such that $y_i^{obs} = \hat{y}_i$, $y_i^{true} = y_i$ and make the approximation exact.

$$\vec{y^{obs}}|\vec{y^{true}} \sim N_p(\vec{y^{true}},\ diag(\epsilon_1, ..., \epsilon_p)) \tag{9}$$

We assume that, given y^{true}, the observations y^{obs} are conditionally independent of the panel data (M).

If the coverage is low, then the estimate of the frequency of a SNP can be 0 (i.e. $\frac{n_i^1}{n_i} = 0$) which will introduce complications when we must invert matrices. Therefore we make the following modification,

$$y_i^{obs} = \frac{n_i^1 + \frac{1}{2}}{n_i + 1} \tag{10}$$

2.3 Incorporating Dispersion

In the distribution of $\vec{y}$, we assumed that the panel and study individuals are from the sample population, and the parameters θ and ρ are estimated without error. Deviations from these assumptions will cause over-dispersion: the true allele frequencies will lie further from their expected values than the model predicts. To allow this, we modify equation 1 by introducing an over-dispersion parameter σ^2.

$$\vec{y^{true}}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma}) \tag{11}$$

We estimate σ^2 by maximizing the multivariate normal likelihood:

$$\vec{y^{obs}}|M \sim N_p(\hat{\mu}, \sigma^2\hat{\Sigma} + diag(\epsilon_1, ..., \epsilon_p)) \tag{12}$$

2.4 Calculating the Posterior

To obtain the distribution for the true frequencies conditional on the observed data, we use Bayes theorem

$$P(\vec{y^{true}}|\vec{y^{obs}}, M) \propto P(\vec{y^{obs}}|\vec{y^{true}})P(\vec{y^{true}}|M)$$

Let,

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^{-1}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{13}$$

and,

$$\bar{\theta} = \bar{\Sigma}\ (\frac{\hat{\Sigma}^{-1}}{\sigma^2}\hat{\mu} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\ \vec{y^{obs}}) \tag{14}$$

Then since the normal is in the conjugate family,

$$\vec{y^{true}}|\vec{y^{obs}}, M \sim N_p(\bar{\theta}, \bar{\Sigma}) \tag{15}$$

Therefore a natural point estimate for $\vec{y^{true}}$ is $\bar{\theta}$.

2.5 Calculating the likelihood to avoid bias of the prior mean

As mentioned above, we assume the the panel and sample individuals are drawn from the same population. This is the never the case in reality but in some applications, the frequencies of alleles of interest have changed significantly but the correlation structure has changed slightly (i.e. very little recombination between nearby SNPs). Therefore we would just like to erase the influence of the prior mean. To do this, we calculate $L(y_t^{true})$.

Denote the test SNP by the subscript t. Then,

$$L(y_t^{true}) = P(y^{obs}|y_t^{true}, M) \propto \frac{P(y_t^{true}|y^{obs}, M)}{P(y_t^{true}|M)} \tag{16}$$

From the above equations, we can see:

$$y_t^{true}|y_t^{obs}, M \sim N(\bar{\theta}_t, \bar{\Sigma}_{tt}) \tag{17}$$

and,

$$y_t^{true}|M \sim N(\theta_t, \Sigma_{tt}) \tag{18}$$

Thus,

$$L(y_t^{true}) \propto e^{\frac{-(y_t^{true}-\bar{\theta}_t)^2}{2\bar{\Sigma}_{tt}} + \frac{(y_t^{true}-\theta_t)^2}{2\Sigma_{tt}}} \tag{19}$$

Completing the square, we can see

$$\frac{-(y_t^{true}-\bar{\theta}_t)^2}{2\bar{\Sigma}_{tt}} + \frac{(y_t^{true}-\theta_t)^2}{2\Sigma_{tt}} = \frac{1}{2}(\frac{1}{\Sigma_{tt}} - \frac{1}{\bar{\Sigma}_{tt}})(y_t^{true} - \frac{\theta_t\bar{\Sigma}_{tt} - \bar{\theta}_t\Sigma_{tt}}{\bar{\Sigma}_{tt} - \Sigma_{tt}})^2 + K \tag{20}$$

Let,

$$\mu_t = \frac{\theta_t\bar{\Sigma}_{tt} - \bar{\theta}_t\Sigma_{tt}}{\bar{\Sigma}_{tt} - \Sigma_{tt}} \tag{21}$$

and,

$$\sigma_t^2 = \frac{1}{-\frac{1}{\Sigma_{tt}} + \frac{1}{\bar{\Sigma}_{tt}}} \tag{22}$$

Because the normal is in the conjugate family,

$$L(y_t^{true}) = pdf\ of\ N(\mu_t, \sigma_t^2) \tag{23}$$

Therefore the MLE of y_t^{true} is μ_t.

2.6 Calculating the inverse of the covariance matrix

When $\hat{\Sigma}$ is singular, we decompose using SVD and calculate the pseudo-inverse,

$$\hat{\Sigma} = U \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{24}$$

where the pseudo-inverse is,

$$\hat{\Sigma}^{+} = V \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^{T} \tag{25}$$

We modify 13 which is now probably non-singular?

$$\bar{\Sigma} = \left(\frac{\hat{\Sigma}^{+}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\right)^{-1} \tag{26}$$

2.7 Computing the Effective Coverage

To calculate effective coverage (n_e) and the effective proportion (p_e), we approximate the normal with a binomial.

2.7.1 Simply taking the reverse mapping of the well known binomial to normal transformation

$$p_e = \mu_t \tag{27}$$

and,

$$\frac{p_e(1-p_e)}{n_e} = \sigma_t^2 \tag{28}$$

Then,

$$n_e = \frac{\mu_t(1-\mu_t)}{\sigma_t^2} \tag{29}$$

2.7.2 Using the Taylor expansion

Our goal is to approximate a normal with a binomial. Let's try this,

$$f(p) = logl(p) = log(p^{n_1}(1-p)^{n-n_1}) \tag{30}$$

Taylor expanding $f(p)$ around its maximum we get,

$$f(p) \approx f(\hat{p}) + \frac{(p-\hat{p})^2}{2} f''(\hat{p}) \tag{31}$$

Therefore,

$$e^{f(p)} = p^{n_1}(1-p)^{n-n_1} \approx Ce^{\frac{-(p-\hat{p})^2}{2\frac{-1}{f''(\hat{p})}}} \tag{32}$$

We can then let,

$$\mu_t = \hat{p} \tag{33}$$

and,

$$\sigma_t^2 = \frac{-1}{f''(\hat{p})} = \frac{(1-\hat{p})\hat{p}^2}{n_1} \tag{34}$$

where $\hat{p} = \frac{n_1}{n}$. Solving the equations for n_e (i.e. n):

$$n_e = \frac{\mu_t(1-\mu_t)}{\sigma_t^2} \tag{35}$$

which is the same as above!

2.8 large-scale simulation

use MS to generate haplotypes and then evolve population with recombination - can use Forqs here or maybe better to write my own code.

3 Phase II - estimating β

Let $f_{i,k,j}$ denote the frequency of the jth SNP in population i and replicate k. Then,

$$log(\frac{1 - f_{i,k,j}}{f_{i,k,j}}) = \mu_j + \beta_j g_i + \epsilon \tag{36}$$

where $\epsilon \sim N_(0, \sigma_d^2)$, σ_d^2 is the variance due to drift, μ_j is the frequency of the jth SNP in the founding population and

$$g_i = \begin{cases} -1 & i = 0 \\ 0 & i = 1 \\ 1 & i = 2 \end{cases}$$

The intuition here is that sites with large β coefficients are under selection.