LDSP - Linear Detection of Selection in Pooled data

Hussein Al-Asadi, Matthew Stephens

1 Introduction

Phase I - Calculate effective coverage 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 estimate.

Phase II - detect selection using effective coverage : fit a linear model.

2 Phase I - calculating effective coverage

2.1 The Prior

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 M denote the $(2m)$xp panel (i.e. $2m$ haplotypes and p SNPs). As in (Wen & Stephens, 2010), we assume

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

(Wen & Stephens, 2010) derive estimates for μ and Σ:

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

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

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 The likelihood

Let (n_i^0, n_i^1) denote the counts of 0 and 1 alleles at SNP i, $n_i = n_i^0 + n_i^1$, and y_i is the true population frequency of the SNP i 1 allele.

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

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

let $y_i^{obs} = \frac{n_i^1}{n_i}$ and replace y_i^{true} by y_i^{obs} in the variance (a common simplification). Therefore our equation becomes,

$$y_i^{obs} | y_i^{true} \dot{\sim} N(y_i^{true}, \frac{y_i^{obs}(1 - y_i^{obs})}{n_i}) \tag{7}$$

$y_1^{obs}|y_1^{true}, y_2^{obs}|y_2^{true}, \ldots, y_p^{obs}|y_p^{true}$ are independent therefore we can write,

$$y^{obs} | y^{true} \dot{\sim} N_p(y^{true},\ diag(\epsilon_1, ..., \epsilon_p)) \tag{8}$$

where $\epsilon_i = \frac{y_i^{obs}(1-y_i^{obs})}{n_i}$

2.2.1 Avoiding $y_i^{obs} = 0$

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

$$y_i^{obs} = \frac{n_i^1 + \frac{1}{2}}{n_i + 1} \tag{9}$$

2.2.2 Incorporating base quality scores

2.3 The Posterior

In the distribution of y^{true}, 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.

$$y^{true} | M \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma}) \tag{10}$$

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

$$y^{obs} | M \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma} + diag(\epsilon_1, ..., \epsilon_p)) \tag{11}$$

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

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

Let,

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^{-1}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{12}$$

and,

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

Then since the normal is in the conjugate family,

$$y^{true}|y^{obs}, M \sim N_p(\bar{\mu}, \bar{\Sigma}) \tag{14}$$

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

2.3.1 Avoiding prior mean bias

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 use the information from SNP correlations. $L(y_i^{true})$ will do the job.

$$L(y_i^{true}) = P(y^{obs}|y_i^{true}, M) \propto \frac{P(y_i^{true}|y^{obs}, M)}{P(y_i^{true}|M)} \tag{15}$$

We showed above,

$$y_i^{true}|y_i^{obs}, M \sim N(\bar{\mu_i}, \bar{\Sigma}_{ii}) \tag{16}$$

and,

$$y_i^{true}|M \sim N(\hat{\mu_i}, \hat{\Sigma_{ii}}) \tag{17}$$

Thus,

$$L(y_i^{true}) \propto e^{\frac{-(y_i^{true}-\bar{\mu_i})^2}{2\bar{\Sigma}_{ii}} + \frac{(y_i^{true}-\hat{\mu_i})^2}{2\hat{\Sigma}_{ii}}} \tag{18}$$

Completing the square, we can see

$$\frac{-(y_i^{true}-\bar{\mu_i})^2}{2\bar{\Sigma}_{ii}} + \frac{(y_i^{true}-\hat{\mu_i})^2}{2\hat{\Sigma}_{ii}} = \frac{1}{2}(\frac{1}{\hat{\Sigma}_{ii}} - \frac{1}{\bar{\Sigma}_{ii}})(y_i^{true} - \frac{\hat{\mu_i}\bar{\Sigma}_{ii} - \bar{\mu_i}\hat{\Sigma}_{ii}}{\bar{\Sigma}_{ii} - \hat{\Sigma}_{ii}})^2 + K \tag{19}$$

Let,

$$\tilde{\mu_i} = \frac{\hat{\mu_i}\bar{\Sigma}_{ii} - \bar{\mu_i}\hat{\Sigma}_{ii}}{\bar{\Sigma}_{ii} - \hat{\Sigma}_{ii}} \tag{20}$$

and,

$$\tilde{\sigma}_i^2 = \frac{1}{-\frac{1}{\hat{\Sigma}_{ii}} + \frac{1}{\Sigma_{ii}}} \tag{21}$$

then,

$$L(y_i^{true}) = \frac{1}{\tilde{\sigma}_i\sqrt{2\pi}} e^{-\left(y_i^{true}-\tilde{\mu_i}\right)^2 / 2\tilde{\sigma}_i^2} \tag{22}$$

2.4 Computing the Effective Coverage

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

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

$$p_e = \tilde{\mu}_i \tag{23}$$

and,

$$\frac{p_e(1-p_e)}{n_e} = \tilde{\sigma}_i^2 \tag{24}$$

Then,

$$n_e = \frac{\tilde{\mu}_i(1-\tilde{\mu}_i)}{\tilde{\sigma}_i^2} \tag{25}$$

2.4.2 Using the Taylor expansion

Our goal is to approximate a normal with a binomial. Let,

$$f(p) = logl(p) = log(p^{n_1}(1-p)^{n-n_1}) \tag{26}$$

Taking the taylor expansion of $f(p)$ around its maximum we get,

$$f(p) \approx f(\hat{p}) + \frac{(p-\hat{p})^2}{2} f''(\hat{p}) \tag{27}$$

Therefore,

$$e^{f(p)} = p^{n_1}(1-p)^{n-n_1} \approx Ce^{\frac{-(p-\hat{p})^2}{2\frac{-1}{f''(\hat{p})}}} \tag{28}$$

let,

$$\tilde{\mu_i} = \hat{p} \tag{29}$$

and,

$$\tilde{\sigma}_i^2 = \frac{-1}{f''(\hat{p})} = \frac{(1-\hat{p})\hat{p}^2}{n_1} \tag{30}$$

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

$$n_e = \frac{\tilde{\mu_i}(1-\tilde{\mu_i})}{\tilde{\sigma}_i^2} \tag{31}$$

which is the same as above!

3 Phase II - estimating β

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

$$log(\frac{1-f_{i,k,j}}{f_{i,k,j}}) = \mu_j + \beta_j g_i + \epsilon \tag{32}$$

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.

4 Computational Issues

4.1 Calculating the inverse of the covariance matrix

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

$$\hat{\Sigma} = U \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{33}$$

where the pseudo-inverse is,

$$\hat{\Sigma}^+ = V \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T \tag{34}$$

We modify 12 which is now probably non-singular?

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^+}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{35}$$