LDSP- Linear Detection of Selection in Pooled sequence data

Matthew Stephens & Hussein Al-Asadi

1 Introduction

We break up the process into two phases.

Phase I: The Intuition is that data at each SNP are binomial counts, which help estimate the frequency of a SNP in a pool, but they don't tell you the frequency exactly, they are noisy. But by combining information across multiple corrected SNPs, you can improve the estimated frequency of the test SNP

Phase II: After we estimate the frequency of the putatively selected SNP in each replicate population, we estimate the group effect using a linear model which also allows us to model genetic drift with a normal error term.. The idea is that in the positively selected population, the group effect will be positive while negative in the negatively selected population, and 0 in the neutrally evolving population.

2 Phase I

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 . As in (Wen & Stephens, 2010), we assume

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

where μ and Σ is calculated from a reference panel consisting of $2m$ haplotypes and p SNPs. (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,

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

and,

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

2.1 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

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

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

$$\implies \hat{X}_i | X_t i \ \sim \ N(X_i, \frac{X_i(1 - X_i)}{n_i}) \tag{6}$$

where $\hat{X}_i = \frac{n_i^1}{n_i}$

Next we replace X_i by $\hat{X}_i$ in the variance for tractibility issues. Therefore,

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

2.2 Incorporating Dispersion

Letting $y_i^{obs} = \hat{X}_i$ from equation 7, we see that

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

where $\epsilon_i = \frac{y_i^{obs}(1 - y_i^{obs})}{n}$ and n is the total coverage for SNP i

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.

$$\vec{y^{true}} \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma}) \tag{9}$$

Combining both equations, we obtain,

$$\vec{y^{obs}} \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma} + diag(\epsilon_1, ..., \epsilon_p)) \tag{10}$$

where we can estimate σ^2 by maximum likelihood.

We use Bayes theorem to obtain the distribution for the true frequencies conditional on the observed data (as derived in Wen & Stephens).

$$P(\vec{y^{true}} | \vec{y^{obs}}) = \frac{P(\vec{y^{obs}} | \vec{y^{true}}) P(\vec{y^{true}})}{P(\vec{y^{obs}})}$$

2.3 Estimating the true frequency at SNP t

....

3 Phase II - estimating β

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

$$log(\frac{1-f_{i,k,j}}{f_{i,k,j}}) = \mu_j + \beta_j g_i + \epsilon \tag{11}$$

where $\epsilon \sim N_(0, \sigma_d^2)$, σ_d^2 is the variance due to drift, μ_j is the frequency of the $j$the 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.