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 - 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 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 - detect selection using improved frequency estimate: 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

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,

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

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

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

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

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

where $\hat{y_i} = \frac{n_i^1}{n_i}$

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

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

Letting $y_i^{true} = y_i$ and $y_i^{obs} = \hat{y_i}$, we see that (we don't have to assume independence here because the observed frequency are conditionally independent with each other given the true frequency, check with Matthew)

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

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

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.

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

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

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

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

$$\bar{\Sigma} = \big(\frac{\hat{\Sigma}^{-1}}{\sigma^2} + diag(\frac{1}{\epsilon_1}, ..., \frac{1}{\epsilon_p})\big)^{-1} \tag{11}$$

and,

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

Then since the normal is in the conjugate family,

$$\vec{y^{true}}|\vec{y^{obs}}, M \sim N_p(\bar{\theta}, \bar{\Sigma}) \tag{13}$$

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

3 Phase II - estimating β

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

$$log(\frac{1-f_{i,k,j}}{f_{i,k,j}}) = \mu_j + \beta_j g_i + \epsilon \tag{14}$$

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.