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 the test SNP

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

$$n_t^1 \sim Bin(n_t, X_t) \dot{\sim} N(n_t X_t, n_t X_t(1 - X_t))$$

where X_t is the true population frequency of the SNP t "1" allele.

$$\implies \frac{n_t^1}{n_t}|X_j \dot{\sim} N(X_t, \frac{\hat{X}_t}{\hat{X}_t(1 - \hat{X}_t)}) \tag{6}$$

where $\hat{X}_t = \frac{n_t^1}{n_t}$.

2.2 Measurement Error and Dispersion

We incorporate measurement error by introducing a single parameter (as in Wen & Stephens) ϵ^2 and assume

$$\vec{y^{obs}}|\vec{y^{true}} \sim N_p(\vec{y^{true}}, \epsilon^2 I) \tag{7}$$

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 **??** by introducing an over-dispersion paramters σ^2.

$$\vec{y^{true}} \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma}) \tag{8}$$

Combining both equations, we obtain,

$$\vec{y^{obs}} \sim N_p(\hat{\mu}, \sigma^2 \hat{\Sigma} + \epsilon^2 I) \tag{9}$$

where we can estimate σ^2 and ϵ^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).

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

$$\sim N_p\Big((\frac{\hat{\Sigma}^{-1}}{\sigma^2} + \frac{I}{\epsilon^2})^{-1}(\frac{\hat{\Sigma}^{-1}\hat{\mu}}{\sigma^2} + \frac{\vec{y^{obs}}}{\epsilon^2}), (\frac{\hat{\Sigma}^{-1}}{\sigma^2} + \frac{I}{\epsilon^2})^{-1}\Big) \tag{10}$$

2.3 Estimating the true frequency at SNP t

....

3 Phase II - estimating β

We can estimate the frequency of the test SNP in group j. Let the frequency of the test SNP in group j and replicate k be $f_{j,k}$. We can fit the following model for all groups j:

$$\begin{pmatrix} log(\frac{1-f_{j,1}}{f_{j,1}}) \\ \vdots \\ log(\frac{1-f_{j,n}}{f_{j,n}}) \end{pmatrix} = \beta_j \begin{pmatrix} log(\frac{1-f^A}{f^A}) \\ \vdots \\ log(\frac{1-f^A}{f^A}) \end{pmatrix} + \vec{\epsilon_j}$$

where $\vec{\epsilon_j} \sim N_n(0, \sigma_d^2 I)$, σ_d^2 is the variance due to drift and f_A is the frequency of the test SNP in the founding population.