# GWAS on approximate phenotype

Hippolyte Verdier

September 2020

## 1 Introduction

### 1.1 Notations

- Phenotypes $P$
- Approximate / predicted phenotypes $\hat{P}$
- Genotypes $X$
- Features $F$

When needed, the subscript $i$ or $j$ is for the dimensions of these vectors, while the superscript $(k)$ refers to the $k$-th individual. Unless stated otherwise, $\langle . \rangle$ means average over individuals $\frac{1}{N}\sum_k .^{(k)}$

We often use the following projections - residuals decomposition :

$$\begin{aligned} P &= \sum_i \langle P, X_i \rangle + \epsilon = \sum_i \beta_i X_i + \epsilon \\ \hat{P} &= \sum_i \langle \hat{P}, X_i \rangle + \epsilon' = \sum_i \hat{\beta}_i X_i + \epsilon' \end{aligned} \tag{1}$$

This decomposition is such that $\forall i \langle X_i, \epsilon \rangle = \langle X_i, \epsilon' \rangle = 0$.

### 1.2 Assumptions

- For simplicity, we assume all genetic loci to be independant :

$$\langle X_i, X_j \rangle = \delta_{i,j} \langle X_i^2 \rangle \tag{2}$$

- For simplicity too, we assume all genetic loci to be centered and normed (by linearly rescaling the original vector of 0, 1 and 2s) :

$$\langle X_i^2 \rangle = 0 \text{ and } \langle X_i \rangle = 0 \tag{3}$$

- Phenotypes are centered and normed

$$\langle P^2 \rangle = 0 \text{ and } \langle P \rangle = 0 \text{ (for} \hat{P} \text{as well)} \tag{4}$$

## 2 How good an approximation do we need ?

The aim of this section is to derive a lower bound for the correlation we should expect between $\beta$ and $\hat{\beta}$, the coefficients of association between the true phenotype and the approximate phenotype, given the correlation between these phenotypes. This correlation is denoted $\rho$ and defined as follow :

$$\langle P, \hat{P} \rangle = \rho = \sum_{i,j} \beta_i \hat{\beta}_j \langle X_i X_j \rangle + \langle \epsilon, \epsilon' \rangle \tag{5}$$

We have $|\rho| < 1$ by definition, and we can assume without loss of generality that $\rho > 0$. We artificially decompose $\hat{P}$ as the sum of its projection on a genetic loci, its projection on $P$ and a residual signal $\epsilon_i''$ :

$$\begin{aligned} \hat{P} &= \langle \hat{P}, X_i \rangle X_i + \epsilon_i' \\ &= \hat{\beta}_i X_i + \langle \hat{P} - \hat{\beta}_i X_i, P \rangle P + \epsilon_i'' \\ \hat{P} &= \hat{\beta}_i X_i + (\rho - \hat{\beta}_i \beta_i) P + \epsilon_i'' \end{aligned} \tag{6}$$

Using the fact that $\langle \epsilon_i'', P \rangle = 0$, we have

$$\begin{aligned} \langle \hat{P}^2 \rangle = 1 &= \hat{\beta}_i^2 \langle X_i^2 \rangle + (\rho - \hat{\beta}_i \beta_i)^2 \langle P^2 \rangle + \langle \epsilon_i''^2 \rangle \\ &\quad + 2\hat{\beta}_i \beta_i (\rho - \beta_i \hat{\beta}_i) + 2\langle X_i, \epsilon_i'' \rangle \\ 1 &= (\rho - \hat{\beta}_i \beta_i)(\rho + \hat{\beta}_i \beta_i) + {\hat{\beta}_i}^2 \langle X_i^2 \rangle + 2\langle X_i, \epsilon_i'' \rangle + \langle \epsilon_i''^2 \rangle \\ 1 &= \rho^2 - \hat{\beta}_i^2 \beta_i^2 + \hat{\beta}_i^2 \langle X_i^2 \rangle + 2\hat{\beta}_i (1 - \langle X_i^2 \rangle) - 2\beta_i (\rho - \beta_i \hat{\beta}_i) + \langle \epsilon_i''^2 \rangle \\ 1 &= (\rho - \hat{\beta}_i \beta_i)^2 + \hat{\beta}_i^2 \langle X_i^2 \rangle + 2\hat{\beta}_i (1 - \langle X_i^2 \rangle) + \langle \epsilon_i''^2 \rangle \end{aligned}$$

Assuming that $\langle X_i^2 \rangle = 1$, and because $\langle \epsilon_i''^2 \rangle \geq 0$, we have

$$(\rho - \hat{\beta}_i \beta_i)^2 + \hat{\beta}_i^2 \leq 1 \tag{7}$$

This yields the following lower bound for $\beta_i \hat{\beta}_i$ :

$$\boxed{\beta_i \hat{\beta}_i \geq \rho - \sqrt{1 - \hat{\beta}_i^2}} \tag{8}$$

**Lower bound for one locus** Thus, $\beta_i$ and $\hat{\beta}_i$ are only guaranteed to be of the same size when

$$|\hat{\beta}_i| \geq \sqrt{1 - \rho} \tag{9}$$

In practice, effect sizes $\beta$ are rather small compared to 1 : assuming we have a correlation of $\rho = 0.5$, the effect size must be of the order of magnitude of $1/\sqrt{2} = 0.7$ to be preserved, which alost never happens in practice.

**Lower bound over all loci** However, assuming a distribution for $\langle \epsilon_i'^2 \rangle$ and averaging over all loci $i$ might give us a lower bound of the genome-wide correlation between effet sizes.

# 3 Genotype - features correlation

In practice, the predicted phenotype is predicted using a linear combination of features.

$$\hat{P} = \sum_i \lambda_i F_i \tag{10}$$

These features have a given correlation to the genotypes

$$F_i = \sum_j \langle F_i, X_j \rangle + \epsilon_i = \sum_j \alpha_{i,j} X_j + \epsilon_i \tag{11}$$

We have, like before, $\forall i, j, \langle X_j, \epsilon_i \rangle = 0$

Building on previous relations, we can write the following equation linking effect sizes of features to those of predicted phenotypes

$$\begin{aligned} \hat{P} = \sum_i \hat{\beta}_i X_i + \epsilon' &= \sum_i \lambda_i \left( \sum_j \alpha_{i,j} X_j + \epsilon_i \right) \\ &= \sum_i \left( \sum_j \lambda_j \alpha_{j,i} \right) X_i + \sum_j \lambda_j \epsilon_j \end{aligned} \tag{12}$$

Here, if we make the assumption that genotypes are independant ($\langle X_i, X_j \rangle = 0$), we can identify terms and write

$$\begin{aligned} \forall i, \hat{\beta}_i &= \sum_j \lambda_j \alpha_{j,i} \\ \epsilon' &= \sum_j \lambda_j \epsilon_j \end{aligned} \tag{13}$$

Hence, if a feature has a strong correlation with a genetic variant (i.e. if $|\alpha_{j,i}|$ is large) and if this feature has an important weight in the phenotype prediction ($|\lambda_j|$ is large), then it is likely that the predicted phenotype will have a strong correlation with this variant too.

**Problematic** The aim is to assess whether associations detected between predicted phenotype and genetic variants are residuals of the features' associations or "amplified" by the linear combination. In the first case, we can't provide much more evidence in favor of an association between our predicted phenotype and the locus than "it's predicted from this feature, and this feature is associated with the locus". In the second case, the fact that the coefficients of regression picked by the learning algorithm result in a "rare" association level suggests that the association is indeed linked to the phenotype which we aim to predict.

## 3.1 Spurious or amplified association ?

We'd like to know which genetic associations are amplified or muted by the linear regression predicting the phenotype. To proceed, we build "random" linear combinations of features and compute their associations with genotypes. We then compare the strengths of these associations with those of the actual phenotypes. Genetic variants which have a stronger association with the predicted phenotype than with a random one (*on average*, criterion to be determined) are likely to be associated with the phenotype of interest.

### 3.1.1 Random linear combination

Because features aren't independant one from the other, we can't just draw coefficients of the linear regression independantly : we need to do so in their principal components space. We thus decompose their correlation matrix in principal components

$$F^T F = MDM^{-1} \tag{14}$$

where $M$ is orthonormal and $D = \text{Diag}(\eta_1^2, \eta_2^2, \ldots, \eta_n^2)$.

We then sample $Z \sim \mathcal{N}(0, \sqrt{D})$ and build back a set of coefficients $L = MZ$. These correspond to the coefficients of a "random linear combination" of the features, such as one we'd had obtained after projection of the features on a random Gaussian vector $U$.

**Proof** : Given $U \sim \mathcal{N}(0, I_N)$

$$\begin{aligned}
\forall i, < U, F_i > &= L_i \\
\langle L_i, L_j \rangle_U &= \int dU \langle U, F_i \rangle \langle U, F_j \rangle \\
&= \int dU \sum_k U^{(k)} F_i^{(k)} \sum_l U^{(l)} F_j^{(l)} \\
&= \sum_{k,l} F_i^{(k)} F_j^{(l)} \int U^{(k)} U^{(l)} dU \\
&= \sum_k F_i^{(k)} F_j^{(k)} \\
\langle L_i, L_j \rangle_U &= \langle F_i, F_j \rangle
\end{aligned} \tag{15}$$

### 3.1.2 GWAS of random combinations of features

We perform a GWAS of the so obtained "random phenotypes", and look at which SNPs have an association with the phenotype of interest stronger than with random linear combinations.