# EigenGWAS theory and application

*Guo-Bo Chen [chen.guobo@foxmail.com]*

*2018-11-30*

# Contents

# Chapter 1

# EigenGWAS basis

This project is dedicated to `EigenGWAS`, a linear model analysis approach for eigenvectors on genomic data, which can be represented as $\mathbf{X}$ the $n \times m$ genotype matrix. Without loss of generality, $x_{jl}$ is a genotype code for the $i^{th}$ sample at the $l^{th}$ biallic locus. The data matrix $\mathbf{X}$ can be generated from genotyping chips, NGS, or GBS.

## 1.1 Genetic relatedness matrix G

We can construct the $n \times n$ genetic relatedness matrix as

$$\mathbf{G} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$$

in which $\tilde{\mathbf{X}}$ is the scaled form of $\mathbf{X}$. However, upon the mating type of the species, $\mathbf{G}$ should be constructed accordingly. For a random mating population, $x_l$ is scaled as $\tilde{x}_l = \frac{x_l - 2p_l}{\sqrt{2p_l q_l}}$, whereas for inbred population, $\tilde{x}_l = \frac{x_l - 2p_l}{\sqrt{4p_l q_l}}$, and $q_l = 1 - p_l$ the frequency for the alternative allele.

So for individual $i$ and $j$,

$$G_{ij} = \frac{1}{\tilde{m}} \sum_{l}^{\tilde{m}} \frac{(x_{il} - 2p_l)(x_{jl} - 2p_l)}{2(1+F)p_l q_l} \tag{1.1}$$

in which $\tilde{m}$ is the number of genotyped loci at both individual $i$ and $j$, and $F$ the inbreeding coefficient takes value of 0 for random mating population and 1 for inbred population.

### 1.1.1 Statistical properties of G

Given $\mathbf{G}$, we can define two population parameters, $n_e$, the `effective population size`, and $m_e$, the `effective number of markers`.

Let $\mathbf{G}_o$ denote the off diagonal elements of $\mathbf{G}$, then we have

$$n_e = \frac{-1}{mean(\mathbf{G}_o)} \tag{1.2}$$

$n_e$ reflects true relatedness between any pair of samples;

$$m_e = \frac{1}{Var(\mathbf{G}_o)} \tag{1.3}$$

The ratio between $\frac{m_e}{m}$ reflects the average linkage disequilibrium between the any pair of markers, and alternatively $m_e$ can be expressed as

$$m_e = \frac{\sum_{l_1=1}^{m}\sum_{l_2=1}^{m}\rho_{l_1 l_2}^2}{m^2} = \bar{\rho}^2 \tag{1.4}$$

in which $\rho_{l_1 l_2}$ is Pearson's correlation between a pair of SNPs, see Appendix **??**. It is an important parameter to describe the evolutionary process of a population of study.

## 1.2 EigenGWAS linear model

Given eigenanalysis (see more its details in wikipedia) of $\mathbf{X}$, we have $\mathbf{E}$ and $\mathbf{\Lambda}$, in which $\mathbf{\Lambda}$ is an $n \times n$ diagonal matrix for eigenvalues and $\mathbf{E}$ is an $n \times n$ matrix for the eigenvectors. $\mathbf{E}_k$ is the $k^{th}$ eigenvector associated with the $k^{th}$ largest eigenvalue. Regressing $\mathbf{E}_k$ against the $l^{th}$ marker, we have the model below

$$\mathbf{E}_k = a + \beta_l \mathbf{x}_l + e \tag{1.5}$$

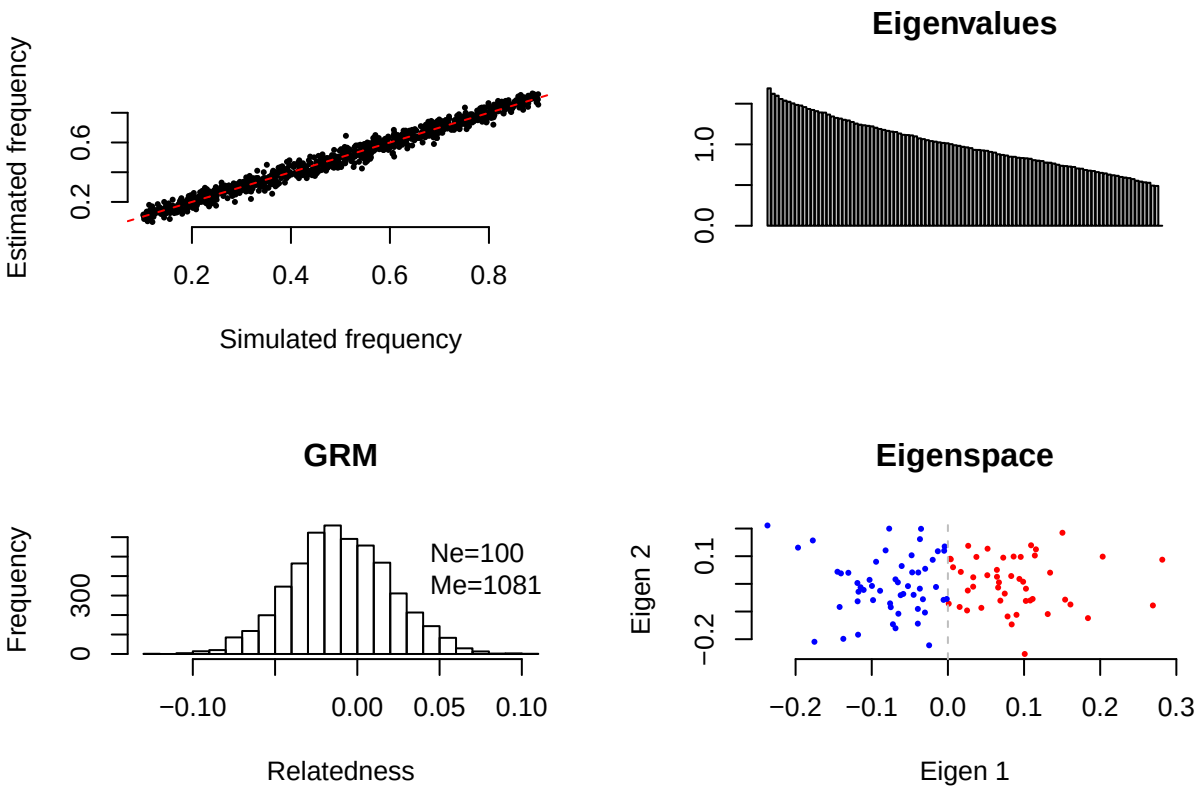
It consequently generates $m$ estimates of $\hat{\beta}$, $\hat{\sigma}_\beta$, and their corresponding $p$ values.

In particular, the one-degree-of-freedom $\chi_1^2$ has approximation as

$$4\frac{n_1 n_2}{n}\frac{(p_{1,l} - p_{2,l})^2}{2p_l q_l} = 4n\omega_1\omega_2 F_{st}^N = nF_{st}^W \tag{1.6}$$

in which $n_1$ and $n_2$ are the numbers of samples at the left and right side of "0" on the eigenvector, see the figure below, and $\omega_g = \frac{n_g}{n}$ the proportion of a subgroup in the sample. $g = 2$ in EigenGWAS analysis. $p_{1,l}$ and $p_{2,l}$ are the frequencies of the reference allele in two subgroups, respectively. $F_{st}^N = \frac{(p_{1,l}-p_{2,l})^2}{2p_l q_l}$ and $F_{st}^W = 2\frac{\sum_{g=1}^{2}\omega_g(p_{g,l}-p_l)^2}{p_l q_l}$.

```
## [1]  100 1000
## [1] "Ne= 100 Me= 1080.91076380488 given N= 100 and M= 1000"
```




### 1.2.1 $\lambda_{GC}$ correction

We can define $\lambda_{GC} = \chi^2_{1,median(p)}/\chi^2_{1,0.5}$, in which $\chi^2_{1,0.5} = 0.455$. We further use subscript $k$ to denote $\lambda_{GC_k}$ the one that is estimated from the EigenGWAS analysis of $\mathbf{E}_k$, as shown (**??**).

After technical correction, correspondingly

$$\tilde{\chi}^2_1 = \chi^2_1/\lambda_{GC_k} \tag{1.7}$$

a correction of the test statistic. Compared with its original form, the correction has several implications

- Statistically, as (**??**) has its response variable from $\mathbf{X}$, the correction removes its overfitting.
- Genetically, it corrects for genetic drift such as soaked in $\mathbf{E}_1$. Here the quantity of the genetic drift is measured by the median of the $m$ $\chi^2_1$ values observed.