Mathematical Institute

Master Thesis

Statistical Science for the Life and Behavioural Sciences

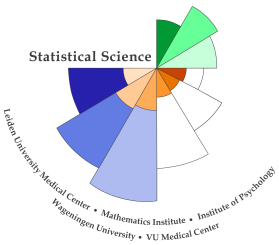
A Goodness-of-fit test for Hardy-Weinberg Equilibrium in the Presence of Covariates

Author:
Anna Morra

First Supervisor:
Dr. Stefan Böhringer
Leiden University Medical Center

Second Supervisor:
Prof.dr. Aad van der Vaart
Mathematical Institute Leiden University

March 2016



Contents

Chapter 1

Introduction

1.1 Genetic Terminology and Concepts

Genetic markers are positions (loci) in the DNA where variation occurs, *i.e.* several alternative DNA sequences can be observed. Genetic markers can therefore be viewed as random variables and realizations thereof can be measured by determining the DNA sequences at a marker present at the loci. These realizations are called alleles. For autosomal markers (markers not located on sex chromosomes) two alleles are observed per individual, each of which is inherited from one of the parents. This pair of alleles is called a genotype. Polymorphisms are markers for which at least two alleles have frequency $> 1\%$. The reasoning behind this definition is that alleles of polymorphisms should not have strongly favorable or detrimental effects on reproductive success for otherwise all but one allele would quickly be eliminated from the population, a process called fixation by selection. Approximately, polymorphisms can therefore be viewed as markers with a neutral evolutionary history, *i.e.* a history unaffected by selection. An important class of polymorphisms are single nucleotide polymorphisms (SNPs), polymorphisms with exactly two alleles (bi-allelic locus) which differ in a single letter of the DNA sequence.

1.2 Hardy-Weinberg Equilibrium

Hardy-Weinberg Equilibrium is an important concept in population genetics and genetic association studies and implies independence of alleles within genotypes as well as constant allele frequencies across generations. The more restrictive concept of Hardy-Weinberg proportions only implies independence of alleles within genotypes and is an important assumption underlying many statistical models. Hardy-Weinberg proportions are the consequence of Mendelian

inheritance and random mating.

For autosomal markers, an individual receives an allele from each parent. According to Mendel's law of segregation, both parental alleles have probability 1/2 to be transmitted to an offspring and the transmission process is independent between parents (Mendelian inheritance). In the case of a bi-allelic locus, assuming that the parental alleles combine randomly to form offspring's genotypes (random mating assumption) it follows that the genotype distribution is completely determined by the frequency ρ of one arbitrarily chosen allele. Denoting by a the allele with frequency ρ in the population and by A the allele with frequency $1-\rho$, the three possible genotypes at the locus are $\{A, A\}$, $\{A, a\}$ and $\{a, a\}$. This set notation ignores parental origin by considering unordered pairs of alleles. In practice, parental origin is unobserved most of the time and it is therefore reasonable to model the unordered case. Genotype frequencies after a single round of random mating and Mendelian inheritance are given by the following equations:

$$\begin{aligned} \pi_1 &:= P(\{a, a\}) = \rho^2 \\ \pi_2 &:= P(\{A, a\}) = 2\rho(1-\rho) \\ \pi_3 &:= P(\{A, A\}) = (1-\rho)^2. \end{aligned} \tag{1.2.1}$$

These equations define the concept of Hardy-Weinberg Equilibrium (HWE) proportions and are referred to as HWE or HWE proportions in the following. Based on HWE proportions, the frequency of allele a in the offspring generation is $\big(2\rho^2 + 2\rho(1-\rho)\big)/2 = \rho$ for an infinite population meaning that, if HWE holds, allele frequencies remain constant from generation to generation for infinite populations.

Further implicit assumptions in this argument include discrete generations, the absence of mutation, selection and genetic drift besides the infinite population size already mentioned. In practice, however, it is not necessary to assume constant allele frequencies across generations. Therefore only the random mating and Mendelian inheritance assumptions are needed to ensure that HWE proportions hold. HWE is usually assumed as it allows to simplify statistical models. This leads to efficiency gains and guarantees identifiability of models (for example haplotype inference [joint distribution of loci] based on separate genotype measurements). However, there are several reasons why HWE proportions could fail to hold, among which is the presence of population substructure. Population substructure in a population is defined as the situation where individuals are not drawn from a single distribution following HWE. The most simple form of population substructure is known as population stratification and refers to the presence of mutually exclusive strata in a population where the allele frequency is the same for all the

individuals in the same stratum but it varies between strata. It is straightforward to see that in a stratified population the number of heterozygotes is smaller compared to a population in HWE, while the variance of the variable representing the counts of one of the two allele (in a bi-allelic locus) is inflated [10]. In general, population substructure corresponds to continuous changes in allele frequencies across geographic regions (*i.e.* a north-south gradient). Among the several methods used to account for population stratification in samples is Principal Component Analysis (PCA) ([4], [5], [8], [11]). PCA is used to project data along a set of axes of variation (referred to as principal components) that are orthogonal by construction. More precisely, the first axis is chosen as to maximize the variance in the data among all possible axes of variation. Iteratively, the following axes are chosen to maximize the variation in the data among all the axes being orthogonal to all the previously selected ones. In the case of genetic data each principal component is a linear combination of single-nucleotide polymorphisms (SNPs) or other genetic variants. The coordinates of the samples along a certain number of principal components are often included as covariates in genetic association studies in order to control for the confounding due to population stratification. However, the number of principal components used is subject to an a-priori choice, a-posteriori judgments of the so-called inflation factor (a measure of alpha-level control) and (or) graphical plotting. The purpose of the present project is to develop a likelihood-based goodness-of-fit test that can evaluate the effect of covariates on correcting for HWE. For example, it can be used to select principal components that give a contribution in correcting for HWE. The underlying idea is to model HW proportions (HWP) on the individual level. HWP are based on an allele frequency that is predicted using covariates. For example, allele frequencies are predicted as a function of a linear predictor based on a number of principal components. Under the null hypothesis, this model is expected to correctly model the distribution for each individual following HWP. To model the distribution under the alternative, a parameter η is introduced, which measures the deviation of the proportion of heterozygous genotypes with respect to HWE. This parameter is shared between all individuals. This approach allows to construct a likelihood ratio statistic to assess fit based on the parameter η, where η equals zero corresponds to the null hypothesis. The parameter η allows for a quantification of deviation from the null including the construction of confidence intervals. Since no closed-form solution exists for the maximum likelihood (ML) estimates, the ML estimates of parameters are obtained by numerical optimization.

The thesis is organized as follows. In chapter 2 the likelihood framework for the new goodness-of-fit test is developed for a single SNP. In particular, both the linear predictor used to model allele frequencies on the individual level and the above mentioned parameter η are introduced and discussed. The model for individual allele frequencies is used, in combination with the

parameter η, to construct the likelihood ratio statistic for assessing fit to HWE. The standard Pearson χ^2 goodness of fit test for HWE is also introduced and a paragraph is dedicated to the ML estimation of the parameters involved in the model for individual allele frequencies. Chapter 3 focus on a special application of the likelihood ratio goodness-of-fit test introduced in Chapter 2, where a number of principal components of generated data matrices are used as covariates to model potential deviations from HWE. The results of two sets of simulations are presented. The simulations have been performed to assess the finite sample behavior of the proposed testing procedure. In particular, genetic data from stratified populations (including 2 or 3 strata) have been generated by modeling genetic drift from a hypothetical ancestral population and the new test has been applied to each SNPs in the corresponding data matrices separately. The joint goodness-of-fit has been measured based on the distribution of p-values of the individual tests. In chapter 4 are reported the results of the analysis of a real data set. In particular, two different subsets of SNPs have been analyzed based on the stratification correction derived from the new test. The last chapter is dedicated to the comparison and discussion of the results obtained for the simulations and the data analysis.

Chapter 2

Goodness of fit Test

In the present chapter we first introduce the standard Pearson goodness of fit test for HWE and then we develop the likelihood framework to construct a likelihood ratio statistic for assessing fit to HWE of a single SNP. In particular, we model HWP on the individual level and express individual allele frequencies as a function of a linear predictor based on a number of covariates. We then introduce a new parameter η, which measures the deviation of the proportion of heterozygous genotypes with respect to HWE and is shared between all individuals in a population. We construct a likelihood ratio statistic to assess fit based on the parameter η, where η equals zero corresponds to the null hypothesis that the considered SNP is in HWE. A paragraph concerning the estimation of the parameters involved in the model for individual allele frequencies via the ML method is also included.

2.1 The Pearson goodness of fit test for HWE

Suppose we have a population of N individuals measured at a single nucleotide polymorphism (SNP). The three possible genotypes at a SNP are determined by the counts of one arbitrarily chosen allele. Let G be the genotype and denote by A and a the two different alleles. We define the genotype score G^s as the number of A alleles in genotype G:

$$G^s = \begin{cases} 0 & \text{if } G = \{a, a\} \\ 1 & \text{if } G = \{A, a\} \\ 2 & \text{if } G = \{A, A\}. \end{cases}$$

The probabilities associated to the three possible genotypes are

$$\pi_1 = P(G^s = 0)$$
$$\pi_2 = P(G^s = 1)$$
$$\pi_3 = P(G^s = 2),$$

where $\sum_{j=1}^{3} \pi_j = 1$.
Let ρ be the frequency of allele a in the population. If the genotype frequencies follow HWE proportions, then $\pi_j = f_j(\rho), j = 1, 2, 3$ where

$$f_1(\rho) = \rho^2$$
$$f_2(\rho) = 2\rho(1-\rho)$$
$$f_3(\rho) = (1-\rho)^2.$$

Suppose we draw a random sample of size n from the population. Let n_j indicate the number of individual in the sample with genotype score equal to j for $j = 0, 1, 2$. The estimate for the frequency of the a allele in the population is $\hat{\rho} = (2n_0 + n_1)/2n$. In order to test for HWE, the genotype counts observed in the sample are compared to the expected counts according to the HWE proportions via the standard Pearson goodness-of-fit test ([10]). The associated test statistic is:

$$X^2 = \sum_{j=0}^{2} \frac{(n_j - n f_{j+1}(\hat{\rho}))^2}{n f_{j+1}(\hat{\rho})}$$

and, under the null hypothesis that HWE proportions hold in the population, it follows asymptotically the chi-square distribution with one degree of freedom. In the remainder we will refer to the Pearson goodness-of-fit test for HWE simply as the standard HWE test.

2.2 Likelihood framework for the new goodness-of-fit test

Suppose that our population of size N contains K homogeneous strata in which HWE holds. This leads to a discrete type of population structure, where each stratum k is characterized by a specific allele frequency ρ_k for a given SNP. Let P_i and G_i^s denote, respectively, the subpopulation membership and the genotype score of individual i. The probability that the genotype score for

individual i is equal to $j = 0, 1, 2$, is given by:

$$P(G_i^s = j) = \sum_{k=1}^{K} P(G_i^s = j, P_i = k) = \sum_{k=1}^{K} \underbrace{P(G_i^s = j | P_i = k)}_{\pi_{j+1}(\rho_k)} P(P_i = k)$$
$$= \sum_{k=1}^{K} \rho_k^{2I(j=0)} 2\rho_k(1-\rho_k)^{I(j=1)}(1-\rho_k)^{2I(j=2)} P(P_i = k).$$

In general, assuming a scenario where individuals belong to discrete populations or to linear combinations of discrete populations may not be the best way to model population history. Indeed, typical scenarios that can be observed in practice are more similar to the one shown in

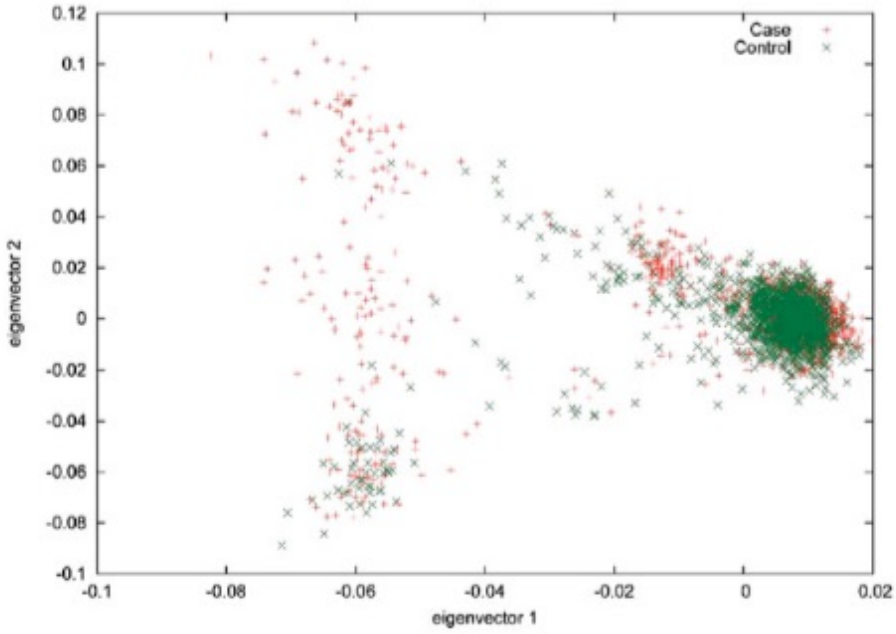


Figure 2.1: Eigenvalue decomposition of the IMAGE sample, a study on attention deficit and hyperactivity disorder in 1,000 children sample from Belgium, Israel, France, Germany and Switzerland. Source: N. M. Laird, C. Lange. The fundamentals of modern statistical genetics. *([10])*

figure 2.1, suggesting that the use of a continuous model is more appropriate. This is achieved by assuming that the number of strata equals the number of individuals in the population. As a consequence each subject is characterized by a specific allele frequency ρ_i. Since this would result in too many parameters, the idea is to express each individual allele frequency as a function of covariates. Let $\mathbf{X}$ be a $N \times (p+1)$ matrix containing the values of p covariates plus an intercept for each of the N individuals in the population. Usually, we consider $\mathbf{X}$ to be fixed covariates which may be informative on population stratification. The likelihood function associated to the

model is then

$$L((\beta_0, \beta_1, \ldots, \beta_p); \mathbf{G}, \mathbf{X}) = \prod_{i=1}^{N} \rho_i^{2I(G_i=0)} \big(2\rho_i(1-\rho_i)\big)^{I(G_i=1)} (1-\rho_i)^{2I(G_i=2)} \tag{2.2.1}$$

$$\rho_i = \text{expit}(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}),$$

where $\mathbf{G}$ is a $N \times 1$ vector of genotype scores and the inverse of the logistic function, defined as

$$\text{expit}(x) = (1 + \exp(-x))^{-1}$$

has been introduced because the linear combination $\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$ can assume values outside the interval $[0, 1]$. Equation 2.2.1 represents the likelihood, assuming each individual is drawn from the *correct* population with frequency ρ_i.

2.3 A measure of deviation from HWE proportions

A new parameter is introduced in order to measure the deviation of the proportion of heterozygous genotypes with respect to HWE proportions:

$$\eta = \text{sgn}(\pi_2 - f_2(\rho)) \Big(\sum_{j=1}^{3} \frac{(\pi_j - f_j(\rho))^2}{f_j(\rho)} \Big)^{1/2}.$$

The genotype probabilities can be expressed in terms of parameter η and the allele frequency $\rho = \pi_1 + \pi_2/2$. The pair (η, ρ) is in fact a reparametrization of the genotype distribution for a population at a SNP:

$$\begin{aligned}
\pi_1 &= \rho\big(1 - (1-\rho)(1+\eta)\big) \\
\pi_2 &= 2\rho(1-\rho)(1+\eta) \\
\pi_3 &= 1 - \rho - \rho(1-\rho)(1+\eta).
\end{aligned}$$

Therefore the Pearson goodness-of-fit test comparing observed genotype frequencies to expected genotype frequencies under HWE is equivalent to test if $\eta \neq 0$. Based on these equations, the following restrictions for the parameter η are obtained:

$$\begin{aligned}
\eta &\geq \max\left(-1, -\frac{1+\rho}{\rho}, -\frac{2-\rho}{1-\rho} \right) = -1 \\
\eta &\leq \min\left(\frac{1-\rho}{\rho}, \frac{\rho}{1-\rho}, \frac{1-2\rho(1-\rho)}{2\rho(1-\rho)} \right) = \frac{1-2\rho(1-\rho)}{2\rho(1-\rho)}.
\end{aligned}$$

Since its definition depends on the general allele frequency ρ in a population, the parameter η measures a joint deviation from HWE for all individuals in the population. However, we want to be able to insert the parameter η into the framework of likelihood 2.2.1, which involves individual allele frequencies ρ_i. Therefore we consider, for each individual i, the value

$$\eta_{max,i} = \min\left(\frac{1-\rho_i}{\rho_i}, \frac{\rho_i}{1-\rho_i}, \frac{1-2\rho_i(1-\rho_i)}{2\rho_i(1-\rho_i)}\right),$$

and multiply it for the population parameter η. The values $\eta_{max,i}$ allow to take into account the limitations imposed to η by each individual allele frequency. Their interpretation is in terms of maximum percentage to which the general deviation from HWE proportions, as measured by η, holds for the single individual in the population. The likelihood function becomes

$$\begin{aligned} L((\eta, \beta_0, \beta_1, \ldots, \beta_p); \mathbf{G}, \mathbf{X}) = \prod_{i=1}^{N} \Big\{ &\Big(\rho_i\big(1-(1-\rho_i)(1+\eta_{max,i}\eta)\big)\Big)^{I(G_i=0)} \\ &\times \big(2\rho_i(1-\rho_i)(1+\eta_{max,i}\eta)\big)\big)^{I(G_i=1)} \\ &\times \big(1-\rho_i-\rho_i(1-\rho_i)(1+\eta_{max,i}\eta)\big)\big)^{I(G_i=2)} \Big\}. \end{aligned} \tag{2.3.1}$$

2.4 Parameters estimation

The maximum-likelihood estimate for the parameter vector $\boldsymbol{\theta} = (\eta, \beta_0, \beta_1, \ldots, \beta_p)$ can be obtained by maximizing the likelihood function 2.3.1. For numerical reasons, it is more convenient to maximize the logarithm of the likelihood function, referred to as the *log-likelihood* function:

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{G}, \mathbf{X}) = \sum_{i=1}^{N} \Big\{ &I(G_i=0) \log\Big(\rho_i\big(1-(1-\rho_i)(1+\eta_{max,i}\eta)\big)\Big) \\ &+ I(G_i=1) \log\big(2\rho_i(1-\rho_i)(1+\eta_{max,i}\eta)\big)\big) \\ &+ I(G_i=2) \log\big(1-\rho_i-\rho_i(1-\rho_i)(1+\eta_{max,i}\eta)\big)\big)\Big\}. \end{aligned} \tag{2.4.1}$$

Since the logarithm is a strictly monotonically increasing function, the maximum does not change if we consider the log-likelihood instead of the likelihood function.
Let $\Theta = \{(\eta, \boldsymbol{\beta}) | \eta \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}$ be the parameter space. The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argsup}}\, l(\boldsymbol{\theta}; \mathbf{G}, \mathbf{X})$$

and it is asymptotically normally distributed

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \xrightarrow{\mathscr{D}} \mathcal{N}(\mathbf{0}, \mathcal{V}(\boldsymbol{\theta_0}))$$

where $\boldsymbol{\theta_0}$ is the true value of the parameter vector. The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is:

$$\mathcal{V}(\boldsymbol{\theta_0}) = \left\{ -\frac{1}{N} E\left[\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{G}, \mathbf{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\theta=\theta_0} \right] \right\}^{-1} = \mathcal{I}(\boldsymbol{\theta_0})^{-1}$$

where $\boldsymbol{\theta}'$ is the transpose of $\boldsymbol{\theta}$ and $\mathcal{I}(\boldsymbol{\theta})$ is the *Fisher information matrix*. Since for our model no closed-form solution to the maximization problem is available, the MLE has been obtained via numerical optimization using the *optim* function in *R*. In addiction to the MLE for the parameter vector, the function can also return an estimate of the Hessian at the solution found. The diagonal elements of the square root of the inverse of the Hessian matrix give an estimate of the standard error of the MLE. Therefore confidence intervals for the parameter estimates can be readily computed.

2.5 Likelihood Ratio Test for deviation from HWE proportions

Based on the likelihood function 2.3.1 it is possible to construct a likelihood ratio test to determine whether the parameter η differs significantly from zero. Let

$$\Theta_0 = \{(\eta, \boldsymbol{\beta}) | \eta = 0, \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}$$

and

$$\Theta_1 = \{(\eta, \boldsymbol{\beta}) | \eta \neq 0, \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}$$

be the parameter spaces corresponding to the null and alternative hypothesis, respectively. The associated test statistic is:

$$LR = -2\left(\sup_{\boldsymbol{\theta} \in \Theta_0} l(\boldsymbol{\theta}; \mathbf{G}, \mathbf{X}) - \sup_{\boldsymbol{\theta} \in \Theta_1} l(\boldsymbol{\theta}; \mathbf{G}, \mathbf{X}) \right). \tag{2.5.1}$$

Under the null hypothesis LR has an asymptotic chi-square distribution with one degrees of freedom:

$$LR \xrightarrow{\mathscr{D}} \chi^2_1.$$

The likelihood ratio test here presented is, in fact, equivalent to the standard test for HWE when no covariates are considered.

Chapter 3

Simulations

In the present section we focus on a special application of the likelihood ratio goodness-of-fit test for HWE introduced in Chapter 2. As already mentioned, one of the principal causes of deviation from HWE proportions is the presence of population substructure in samples. This refers to a situation where individuals are drawn from underlying populations for which allele frequencies differ for a given marker. In the simulations considered here, we focus on discrete substructure, *i.e.* different discrete populations each following HWE. For markers that are independent in the sub-population, any linear combinations of standardized genotypes (*i.e.* centered and scaled to unit variance) has the same variance if allele frequencies are identical in all sub-populations. Differences in allele frequencies induce covariance between markers thereby increasing variance for linear combinations having high weight on such markers. This consideration motivates the use of principal components as a means to describe and correct for population stratification. The sample covariance matrix of standardized genotypes is additionally subject to sample fluctuation, an effect, that is potentially important in this application as the number of samples is usually much smaller than the number of markers (*e.g.* number of samples $N = 10^3$, number of markers $M = 10^5$).

In the ensuing simulations, the proposed likelihood ratio test is applied to each SNP in each data set, using a number of principal components as explanatory variables to model population substructure as determined by allele frequencies for each individual. We use between one and four principal components in the analysis. Simulation results are evaluated by analyzing the distribution of p-values of these tests. Particular attention is given to the probability of making a type I error as different (subgroups of) principal components are included into the model for allele frequencies of individuals to judge the appropriateness of using principal components for correcting population substructure.

3.1 Simulation model

In our simulations we consider bi-allelic markers throughout. The genetic data consist of a $N \times M$ matrix of genotypes counts for N samples measured at M *independent* bi-allelic markers. We use the principal components of this matrix as the columns of matrix $\mathbf{X}$ introduced in Chapter 2. We remark that in case of genetic populations we generally speak about individuals rather than about samples. Therefore in the following the term *individual* is often used in place of *sample*.

The genotype distribution for a single unstratified population assumes Hardy-Weinberg proportions so that it is defined by the frequency of an arbitrarily chosen allele. In this setting the genotype frequencies, for a single individual at a single marker, are given by:

$$P(G = i) = \begin{cases} \rho^2 & \text{if } i = 0 \\ 2\rho(1-\rho) & \text{if } i = 1 \\ (1-\rho)^2 & \text{if } i = 2 \end{cases}$$

where ρ is the minor allele frequency of the bi-allelic marker and G is a random variable representing the counts of the major allele. In order to simulate a stratified population with a number S of strata, we generate S different populations for which we establish an approximate model for simulating genetic drift. First, we assume an ancestral population for which the minor allele frequency (the minimum of the two allele frequencies) of each marker is drawn from a uniform distribution on the interval $(0.05, 0.5)$. This roughly corresponds to the allele frequency distribution as used by current DNA chip technology. The reason for the lower bound of 0.05 is to avoid problems due to lack of convergence due to few observations in one of the allele categories. To simulate drift of allele frequencies, we use the following approximation to the Wright-Fisher model. For constant population size, the allele distribution of an offspring distribution is binomial. Here, all alleles are treated as being independent and the next generation is given by a draw from this binomial distribution. The population frequency in the new generation determines the binomial distribution for the next generation. This change in frequency can be approximated by perturbing the allele frequency by a normal random variable. For a given allele frequency ρ, the binomial distribution of the number of minor alleles in the present generation given the number of minor alleles in the immediately previous generation has variance $2N\rho(1-\rho)$, for a population of size N. The variance for the allele frequency in the present generation is therefore $\rho(1-\rho)/(2N)$. Assuming independent variation in each generation, the change over T generations can be roughly approximated by a random perturbation $R \sim N(0, T\rho(1-\rho)/(2N))$. The

denominator in the variance of R corresponds to the fact that larger populations drift away from each other slower than smaller ones. A simplification made here, is that drift only depends on the initial allele frequency. Here, N denotes the "effective" number of individuals in the population during the drifting period which does not have to correspond to the current population size, for example, if a population undergoes exponential growth. In this sense it represents the size of a population of constant size which exhibits the same drift as the population as hand and for an expanding population it is reasonable to assume that it is smaller than the current population size. In the following, we denote this population size as N_{drift}, *i.e.* $\mathrm{Var}(R) = T\rho(1-\rho)/(2N_{drift})$. Simulated population size is denoted as N, and the parameter N_{drift} is assumed to be fixed and it is kept smaller than N. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)$ be the vector of minor allele frequencies of M bi-allelic markers from the hypothetical (homogeneous) ancestral population and denote with $\tilde{\boldsymbol{\rho}}_{\boldsymbol{s}} = (\tilde{\rho}_{s1}, \ldots, \tilde{\rho}_{sM})$ the vector of drifted allele frequencies of stratum s, for $s = 1, \ldots, S$. In particular, $\tilde{\boldsymbol{\rho}}_{\boldsymbol{s}}$ is obtained by adding to $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)$ the random vector $\boldsymbol{r}_{\boldsymbol{s}} = (r_{s1}, \ldots, r_{sM})$, whose elements r_{sj} are realizations of the random variable R used to simulate drift of allele frequencies for stratum s. Since R is normally distributed with mean zero, random variates can assume both small positive and negative values so that some of the values $\tilde{\rho}_{sj} = \rho_j + r_{sj}$ could happen to fall outside the interval $(0.05, 0.5)$. Therefore each $\tilde{\rho}_{sj}$ has been mapped into the interval $(0, 0.5)$ through the function:

$$f(x) = \frac{\exp(x - 0.5)}{1 + \exp(x - 0.5)}. \tag{3.1.1}$$

Given the specific values of T and N_{drift} considered in the simulations, the allele frequencies $\tilde{\rho}_{sj}$ are generally mapped to the interval $(0.05, 0.5)$. Each of the S vectors of drifted allele frequencies is used to generate the matrix of genotype scores of M bi-allelic markers for a single substratum, whose distribution assumes Hardy-Weinberg proportions. Matrices for the different homogeneous substrata are combined into a single data matrix for the whole population, characterized by the following genotype frequencies:

$$P(G_{ijs} = k) = \begin{cases} [f(\tilde{\rho}_{sj})]^2 & \text{if } k = 0 \\ 2f(\tilde{\rho}_{sj})(1 - f(\tilde{\rho}_{sj})) & \text{if } k = 1 \\ (1 - f(\tilde{\rho}_{sj}))^2 & \text{if } k = 2, \end{cases}$$

where G_{ijs} is the genotype score of individual i at SNP j in stratum s and f is the function defined in (3.1.1).

3.2 Simulation scenarios

The different scenarios considered in the simulations are obtained by varying the parameters reported in Table A.1. Parameter N_{strata} represents the number of homogeneous strata included in the global population and varies from 1 to 3. For each value of N_{strata} four distinct scenarios have been simulated by considering two different values for the total number N of individuals in the population (1000 and 5000) and for the number T of generation used to model the genetic drift of the strata from an hypothetical ancestral population (500 and 1000). Furthermore, each of the 12 scenarios obtained in this way is considered three times corresponding to three values chosen for the maximum number of principal components p used to model the allele frequencies of individuals. Notice, that the parameter T is only relevant for the simulation of populations where $N_{strata} = 2$ or $N_{strata} = 3$. Therefore, in terms of population stratification we can identify five basic scenarios out of the total 36 obtained considering all the possible combinations of parameters reported in Table A.1. A graphical representation of those five scenarios in terms of the first two principal components is given in Figure A.1. The top left panel shows the picture of a homogeneous population ($N_{strata} = 1$). The top right panel represents the case of a stratified population with two different strata whose genetic drift from a hypothetical ancestral population has been simulated taking the number of generation T equal to 500. The middle left panel shows again a population with two strata where the parameter T has been set equal to 1000. Analogously the middle right and the bottom panels show the pictures of stratified populations composed of three strata for $T = 500$ and $T = 1000$, respectively. The plots show how increasing the value of T from 500 to 1000 results in a more extreme separation of the different strata along the first two principal components. It should also be noticed that for simplicity, populations have been generated with the same number of individuals in each stratum. The different strata in a population are simulated to have drifted away from a single hypothetical ancestral population starting at the same time point, *i.e.* the ancestry is star-shaped instead of tree-shaped. The scenarios described so far have been run for two different values of parameter N_{drift} as introduced in section 3.1, which, together with the parameter T, determines the level of separation among the strata present in the simulated population in terms of the principal components. The two values considered for N_{drift} were 300 and 600.
Taken together 36 scenarios were considered in the simulations.

3.3 Results

Each of the 36 scenarios summarized in Table A.1 has been simulated 100 times for $N_{drift} = 600$ and 10 times for $N_{drift} = 300$ and the results were averaged over the number of repetitions. The small numbers of repetitions considered here, compared to what is normally used in a simulation study, are due to the long running times required by the current implementation. Simulations were run on the compute cluster of the LUMC and took about 50000 core hours corresponding to approximately three weeks of computations by making parallel use of 100 cores. As simulation re-starts were necessary, more extensive simulations could not be performed in the available time.

In the present section we report the results related to the cases where a maximum of $p = 4$ principal components were used for modeling allele frequencies of individuals. In particular, the different principal components were included into the model first separately and then sequentially resulting in seven different tests, which were applied separately on each SNPs in the generated data. The standard HWE goodness-of-fit test was also performed in order to compare to the standard scenario for evaluating HWE for given SNPs. In particular, homogeneous population scenarios, *i.e.* scenarios not including drift, have been generated to assess the behavior of the testing procedure under the null model, *i.e.* a homogeneous population. Results for these analyses are reported in table 3.1, where the column *ST* stands for *Standard Test* and refers to the standard HWE goodness-of-fit test while the numbers between parenthesis in the following columns indicate the principal components included in the model for the allele frequencies of individuals when testing for deviation from HWE via the proposed likelihood-based goodness-of-fit test. To re-iterate, the corresponding null hypothesis is given by H_0: $\eta = 0$. The values reported in the tables are the proportions of p-values which are smaller or equal the level of significance specified in the second column (α) of the corresponding line. These proportions are computed across SNPs for each generated data set and then averaged over the repetitions for a specific scenario. As expected, the results for the standard HWE test keep the four different α-levels considered in all the scenarios, since they represent unstratified populations. Focusing on scenario 3, we can see that when the first principal component is used to model allele frequencies the α-level of .05 is well maintained, whereas α-level of (.01) is reasonably well kept while the observed proportions of p-values no greater than .001 and .0001 are, respectively, smaller than .001 and .0001 making the test conservative. The same happens when the second to the fourth principal components are added separately to the model as well as when the first two and three principal components are added together to the model. When all the first four principal components are used to model the allele frequencies of individuals the proportion of p-values

Scenario	α	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
3	.05	.05018	.05077	.05085	.05077	.05078	.05154	.05192	.03681
	.01	.01012	.00790	.00796	.00790	.00794	.00811	.00826	.00555
	.001	.00099	.00020	.00020	.00021	.00020	.00022	.00022	.00014
	.0001	.00009	.00000	.00000	.00000	.00000	.00000	.00000	.00000
12	.05	.04961	.04973	.04992	.04975	.04990	.04967	.04852	.03506
	.01	.00980	.01002	.00989	.00987	.00988	.01007	.01002	.00688
	.001	.00097	.00100	.00101	.00101	.00101	.00104	.00107	.00072
	.0001	.00006	.00006	.00007	.00007	.00007	.00007	.00008	.00005
21	.05	.05045	.05103	.05099	.05104	.05085	.05170	.05193	.03645
	.01	.00978	.00769	.00771	.00768	.00775	.00789	.00804	.00527
	.001	.00105	.00020	.00019	.00020	.00019	.00021	.00021	.00015
	.0001	.00009	.00000	.00000	.00000	.00000	.00000	.00000	.00000
30	.05	.05005	.05026	.05045	.05023	.05029	.05040	.04911	.03558
	.01	.00990	.01004	.01005	.00995	.01001	.01017	.01010	.00690
	.001	.00107	.00108	.00108	.00107	.00108	.00110	.00112	.00075
	.0001	.00012	.00013	.00013	.00012	.00013	.00013	.00015	.00010

Table 3.1: ***Homogeneous populations with $N_{drift} = 600$.*** *Scenarios* 3 *and* 21 *refer to populations of* $N = 1000$ *samples, while scenarios* 12 *and* 30 *refer to populations of* $N = 5000$ *samples. The number of markers* M *is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

smaller or equal to .05 and .01 also decrease. A totally analogous pattern shows up for the results related to scenario 21. As we focus on the results of scenarios 12 and 30, we see that the first six (likelihood-ratio) tests keep the different α-levels properly as well as the standard HWE test. Only when all the first four principal components are used to model the allele frequencies the observed proportions of small p-values decrease again compared to expected. These results show that as the sample size increases the control of type I error improves. However, the inclusion of four principal components in the model for the allele frequencies of individuals still results in conservative behavior.

Tables 3.2 and 3.3 show the results related to populations with two strata for $N_{drift} = 600$ and $N_{drift} = 300$, respectively. Starting with table 3.2, it is interesting to observe that the standard HWE test keeps the different α-levels pretty exactly for the first three scenarios and only in the last one we can see an increase in the observed proportions of small p-values. If we examine the graphical representation of the different scenarios in terms of projections of individuals along the first two principal components, we can see that in scenarios 6 (top right panel in figure A.1) and 15 (top right panel in figure A.2) the different strata are already clearly separated even though they are close to each other. However, an effect on the observed proportion of small p-values for the standard HWE test is only visible in scenario 33 (bottom panel in figure A.2) where the

Scenario	α	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
6	.05	.05038	.05030	.05080	.05079	.05080	.05094	.05128	.03555
	.01	.00997	.00765	.00775	.00764	.00768	.00793	.00802	.00515
	.001	.00103	.00017	.00017	.00016	.00018	.00017	.00018	.00014
	.0001	.00009	.00000	.00000	.00000	.00000	.00000	.00000	.00000
15	.05	.05062	.04943	.05058	.05066	.05069	.04932	.04811	.03398
	.01	.01037	.01007	.01044	.01041	.01036	.01021	.01007	.00684
	.001	.00097	.00088	.00095	.00096	.00094	.00090	.00092	.00062
	.0001	.00011	.00010	.00011	.00010	.00011	.00010	.00010	.00007
24	.05	.05422	.05043	.05403	.05416	.05403	.05100	.05108	.03407
	.01	.01145	.00769	.00873	.00870	.00882	.00776	.00790	.00474
	.001	.00123	.00025	.00026	.00025	.00024	.00019	.00019	.00011
	.0001	.00014	.00007	.00000	.00000	.00000	.00000	.00000	.00000
33	.05	.07024	.04967	.06924	.06932	.06911	.04943	.04736	.03263
	.01	.01996	.00998	.01944	.01945	.01952	.01000	.00983	.00628
	.001	.00436	.00093	.00417	.00420	.00418	.00091	.00090	.00054
	.0001	.00131	.00014	.00126	.00126	.00126	.00009	.00009	.00005

Table 3.2: ***Populations with two strata and $N_{drift} = 600$.*** *(i) Scenario* 6 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 15*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 24*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 33*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

two strata in the population are quite distant from each other. In principle, the standard HWE goodness-of-fit test should be sensitive to these scenarios.

Examining the results for scenario 33, we can see that as the first principal component is used to model the allele frequencies the control of type I error improves (all the four different α-levels are well maintained). On the other hand, when one of the other three principal components is included separately to the model for the allele frequencies of individuals then the results are similar to the ones obtained for the standard HWE test. This is expected because the simulated data contain a single dimension of population structure and therefore only the first principal component should be relevant for the correcting of population stratification.

When the first two or three principal components are considered together to test for deviation from HWE the α-levels are well maintained again, due to the fact that we are correcting for the first principal component, which is the one that explains the stratification. However, when we consider all the four principal components together the observed proportions of small p-values are again smaller than expected, suggesting overfitting.

The results for the remaining scenarios (6,15 and 24) are analogous to the results obtained for the homogeneous populations, since for those scenarios there seem to be no effect of population stratification on the type I error of the standard HWE test.

Scenario	α	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
6	.05	.05254	.04874	.05234	.05284	.05202	.04910	.04906	.03300
	.01	.01138	.00778	.00852	.00860	.00846	.00780	.00790	.00464
	.001	.00128	.00020	.00024	.00018	.00024	.00016	.00014	.00006
	.0001	.00014	.00004	.00000	.00000	.00000	.00000	.00000	.00000
15	.05	.07174	.05074	.07044	.07082	.07048	.05020	.04790	.03368
	.01	.01978	.01048	.01926	.01922	.01940	.01052	.01018	.00684
	.001	.00428	.00120	.00426	.00414	.00422	.00124	.00120	.00072
	.0001	.00114	.00014	.00108	.00106	.00106	.00012	.00012	.00004
24	.05	.10026	.04856	.09778	.09796	.09770	.04870	.04834	.02908
	.01	.04040	.00696	.03392	.03416	.03398	.00708	.00716	.00412
	.001	.01468	.00018	.00782	.00792	.00792	.00016	.00020	.00010
	.0001	.00674	.00000	.00214	.00214	.00212	.00000	.00000	.00000
33	.05	.19296	.05056	.19050	.19002	.19012	.04940	.04608	.03020
	.01	.11852	.01044	.11656	.11648	.11714	.01000	.00942	.00584
	.001	.07280	.00120	.07178	.07174	.07182	.00116	.00108	.00066
	.0001	.05142	.00016	.05084	.05064	.05072	.00012	.00012	.00006

Table 3.3: ***Populations with two strata and $N_{drift} = 300$.*** *(i) Scenario* 6 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 15*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 24*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 33*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

As we consider the results for the populations with two strata where $N_{drift} = 300$ the situation changes. In particular, by decreasing the value of N_{drift} from 600 to 300 we increase the variance of the random variable used to model the genetic drift from an ancestral population through generations. This results in more accentuated separation between population strata for all the scenarios. The effects of such a more extreme separation are visible on the observed proportions of small p-values for the standard HWE test for scenarios 15, 24 and 33. In particular, the results of scenarios 15 and 33 are analogous to the results of scenario 33 with $N_{drift} = 600$. We observe that results of scenarios 6 and 24 show evidence of conservative behavior for all the likelihood ratio tests performed, especially when the two smallest α-levels (10^{-3} and 10^{-4}) are considered.

Tables 3.4 and 3.5 show the results related to populations with three strata for $N_{drift} = 600$ and $N_{drift} = 300$, respectively. In this case, the populations were generated to have two dimensions of population structure. Therefore only the first two principal components are expected to be of importance in correcting for population stratification. Looking at Table 3.4, we can see that the results for the standard HWE test are analogous to the corresponding results reported in Table 3.2 for the two-strata population scenarios. In particular, the standard HWE test seems to maintain the different α-levels well for the first three scenarios and only in scenario 36 an increase in the observed proportions of small p-values is visible. As expected, the results for this

Scenario	α	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
9	.05	.04983	.04992	.04985	.05014	.05012	.05029	.05054	.03523
	.01	.01003	.00775	.00775	.00782	.00780	.00783	.00797	.00518
	.001	.00100	.00020	.00019	.00018	.00019	.00020	.00020	.00013
	.0001	.00009	.00000	.00000	.00000	.00000	.00000	.00000	.00000
18	.05	.05122	.04981	.04984	.05077	.05080	.04901	.04736	.03366
	.01	.01034	.00996	.01006	.01030	.01025	.00995	.00981	.00677
	.001	.00108	.00106	.00102	.00107	.00105	.00103	.00103	.00067
	.0001	.00013	.00013	.00012	.00013	.00012	.00013	.00013	.00009
27	.05	.05476	.05147	.05145	.05416	.05440	.05001	.05004	.03243
	.01	.01149	.00813	.00813	.00878	.00888	.00786	.00794	.00475
	.001	.00126	.00021	.00023	.00025	.00025	.00018	.00019	.00011
	.0001	.00015	.00001	.00003	.00000	.00000	.00000	.00000	.00000
36	.05	.07594	.05923	.05985	.07430	.07427	.04986	.04812	.03158
	.01	.02167	.01376	.01421	.02094	.02093	.01017	.00992	.00621
	.001	.00437	.00201	.00213	.00415	.00416	.00111	.00110	.00064
	.0001	.00104	.00037	.00038	.00101	.00099	.00011	.00012	.00006

Table 3.4: ***Populations with three strata and $N_{drift} = 600$.*** *(i) Scenario* 9 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 18*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 27*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 36*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

last scenario show that both the first two principal components contribute to improve the control of type I error. When the first two and the third principal components are included together into the model for allele frequencies of individuals all four different α-levels are well maintained. On the other hand, when either the third or the fourth principal component is included separately into the model the results are similar to the corresponding result obtained for the standard HWE test. Again, when all the four principal components are considered together the likelihood ratio test shows evidence of conservative behavior. An analogous pattern shows up for the results of scenarios 18, 27 and 36 in Table 3.5.

In order to get deeper insight into the distribution of p-values for the different tests performed we considered a tail strength measure proposed in 2005 by Taylor and Tibshirani ([3]) and defined as

$$TS(p_1, \ldots, p_M) = \frac{1}{M} \sum_{k=1}^{M} \Big(1 - p_{(k)} \frac{M+1}{k}\Big), \tag{3.3.1}$$

where p_k for $k = 1, \ldots, M$ are the p-values associated to M null hypotheses and $p_{(k)}$ for $k = 1, \ldots, M$ are the corresponding ordered p-values. If all the M null hypotheses are true (global null hypothesis) then the corresponding p-values all have a uniform distribution in $[0, 1]$, provided that they are computed from a continuous test statistic. In this case, each $p_{(i)}$ has expected value

Scenario	α	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
9	.05	.05516	.05144	.05284	.05430	.05436	.05082	.05126	.03344
	.01	.01192	.00822	.00834	.00914	.00906	.00800	.00806	.00490
	.001	.00140	.00022	.00026	.00024	.00022	.00024	.00024	.00012
	.0001	.00008	.00000	.00002	.00000	.00000	.00000	.00000	.00000
18	.05	.07498	.05746	.05916	.07330	.07288	.04850	.04680	.03056
	.01	.02122	.01360	.01416	.02048	.02068	.00984	.00954	.00532
	.001	.00436	.00202	.00210	.00422	.00424	.00090	.00086	.00050
	.0001	.00124	.00046	.00034	.00112	.00118	.00008	.00006	.00006
27	.05	.11618	.07450	.07610	.11346	.11366	.05090	.05002	.02870
	.01	.04664	.01850	.01968	.03948	.03922	.00768	.00764	.00354
	.001	.01596	.00222	.00270	.00792	.00800	.00008	.00008	.00004
	.0001	.00654	.00038	.00072	.00160	.00164	.00000	.00000	.00000
36	.05	.25454	.13730	.14182	.24992	.25032	.04722	.04364	.02674
	.01	.16136	.06812	.07380	.15856	.15910	.00904	.00878	.00472
	.001	.09852	.03514	.03800	.09676	.09650	.00102	.00100	.00058
	.0001	.06544	.02100	.02254	.06430	.06428	.00014	.00014	.00006

Table 3.5: ***Populations with three strata and $N_{drift} = 300$.*** *(i) Scenario* 9 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 18*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 27*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 36*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

$i/(M+1)$ and the expectation of TS is zero. However, when $p_{(i)} < i/(M+1)$ the corresponding term between parentheses in 3.3.1 is greater than zero. Therefore, large positive values of the tail strength indicate that there are more small p-values than what would be expected by chance, providing evidence against the global null hypothesis. Contrarily, negative values of the tail strength indicate a globally conservative behavior. The tail strength can therefore be used to make a global assessment of the p-value distribution as opposed to assessing one or a few α-levels.

Tables 3.6 and 3.7 show TS values for all the scenarios discussed so far. The results are presented in blocks according to the number of strata in the population and averaged over the number of repetitions for each scenario. Focusing on Table 3.6, we can see that the TS values of the standard HWE test for scenarios referring to homogeneous populations (3, 12, 21 and 30) are mostly positive (except for scenario 12) and close to zero, indicating no evidence of deviation of the p-value distribution from the expected uniform distribution. When the first principal components are added separately to the model for allele frequencies of individuals the TS values for the likelihood ratio test remain close to zero but become negative. When the first two and three principal components are considered together, the TS values are still negative but slightly increase in absolute value indicating that p-values for the likelihood ratio test are,

Scenario	ST	(1)	(2)	(3)	(4)	(1, 2)	(1, 2, 3)	(1, 2, 3, 4)
3	.0026	-.0061	-.0063	-.0061	-.0062	-.0089	-.0266	-.2181
12	-.0026	-.0073	-.0074	-.0074	-.0074	-.0204	-.0534	-.2634
21	.0010	-.0076	-.0078	-.0077	-.0077	-.0102	-.0285	-.2198
30	.0002	-.0046	-.0047	-.0047	-.0045	-.0179	-.0511	-.2633
6	.0005	-.0107	-.0093	-.0094	-.0093	-.0150	-.0328	-.2351
15	.0069	-.0095	.0003	.0005	.0001	-.0224	-.0588	-.2873
24	.0244	-.0084	.0119	.0115	.0117	-.0155	-.0404	-.2598
33	.0981	-.0099	.0878	.0876	.0875	-.0302	-.0726	-.3146
9	.0014	-.0098	-.0098	-.0089	-.0088	-.0153	-.0336	-.2349
18	.0072	-.0066	-.0065	-.0002	-.0002	-.0251	-.0651	-.2857
27	.0260	-.0040	-.0034	.0116	.0117	-.0225	-.0454	-.2865
36	.1200	.0361	.0401	.1077	.1077	-.0395	-.0761	-.3619

Table 3.6: ***Tail strength for simulations with $N_{drift} = 600$.*** *(i) Scenarios* 3, 12, 21 *and* 30 *refer to homogeneous populations. (ii) Scenarios* 6, 15, 24 *and* 33 *refer to populations with two strata. (iii) Scenarios* 9, 18, 27 *and* 36 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 100 *repetitions for each scenario.*

in those cases, slightly larger then we would expect under the uniform distribution. If all four principal components are included together a substantial increase in the percentage of large p-values is observable. For instance, for scenario 30 the TS value is $-.26$ indicating that there are (on average) 26% more larger p-values than expected by chance.

Examining the results of non-homogeneous populations, we can see that for scenario 36, which refer to a population with three strata, the TS value for the standard HWE test is .12, so that there are on average 12% more significant test statistics than expected by chance. The percentage decrease to 4% for the likelihood ratio test when the first two principal components are used, separately, to model allele frequencies of individuals and rise again to 10% when the third or fourth principal components are considered. An increase in the percentage of large p-values is observable when the principal components are added sequentially into the model for allele frequencies of individuals and reaches a value 36% when all the four principal components are considered together.

The results reported in Table 3.7 are analogous and consistent to the results reported in Table 3.6. The only difference is that, in the case of Table 3.7 the different strata in the non-homogeneous populations are farther apart from each other, resulting in higher absolute values for the TS of the standard HWE test and for some of the likelihood ratio tests.

3.4 Remarks

The results reported in the present chapter indicate a generally conservative behavior of the proposed likelihood-ratio test when it is used to evaluate the effect of correcting for population

Scenario	ST	(1)	(2)	(3)	(4)	(1,2)	(1,2,3)	(1,2,3,4)
3	.0026	-.0064	-.0056	-.0059	-.0053	-.0089	-.0267	-.2198
12	.0054	.0012	.0013	.0011	.0007	-.0108	-.0438	-.2489
21	-.0042	-.0121	-.0125	-.0122	-.0126	-.0142	-.0329	-.2285
30	.0028	-.0021	-.0021	-.0020	-.0029	-.0161	-.0504	-.2643
6	.0183	-.0150	.0050	.0051	.0056	-.0233	-.0477	-.2651
15	.0987	-.0100	.0886	.0881	.0880	-.0308	-.0697	-.3136
24	.1923	-.0194	.1796	.1795	.1796	-.0330	-.0707	-.3412
33	.3826	-.0145	.3723	.3722	.3716	-.0500	-.1112	-.3789
9	.0278	-.0009	.0007	.0136	.0130	-.0176	-.0394	-.2788
18	.1197	.0333	.0381	.1077	.1073	-.0450	-.0815	-.3704
27	.2442	.0997	.1077	.2281	.2287	-.0349	-.0778	-.3679
36	.4855	.2669	.2778	.4747	.4742	-.0802	-.1370	-.4348

Table 3.7: ***Tail strength for simulations with N_{drift} = 300.*** *(i) Scenarios* 3, 12, 21 *and* 30 *refer to homogeneous populations. (ii) Scenarios* 6, 15, 24 *and* 33 *refer to populations with two strata. (iii) Scenarios* 9, 18, 27 *and* 36 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 10 *repetitions for each scenario.*

stratification using principal components. This behavior is more evident in smaller populations. An interesting fact is that even in the case of populations including completely separated strata, as shown in the top right and middle left panels of Figure A.2, for which genotype frequencies deviate from HWE proportions, the standard HWE test does not have power to detect those. The results for the tail strength of the different likelihood ratio tests indicate that the number of p-values larger than expected under a uniform distribution increase remarkably when all the first four principal components are included in the model for allele frequencies of individuals, giving stronger evidence of a conservative behavior and overfitting for this particular test.

Chapter 4

Data Analysis

In the present chapter we report the results obtained by applying the proposed likelihood based test for HWE on a Wellcome Trust Case Control Consortium (WTCCC) dataset containing the genetic information of 4798 individuals at 469612 SNPs. A number of preparatory steps were necessary before proceeding with the test. First, some data quality control has been performed in order to assess genotyping quality. Quality control checks led to the exclusion of a number of individual and markers from subsequent analyses. Since SNPs in close regions of the same chromosome are highly correlated, we considered a subset of approximately independent SNPs out of the total set remained in the analysis after quality control. Principal components used to perform the test were computed based on a subset of 10000 out of the considered set of approximately independent SNPs. For computational reasons, only two subsets of markers out of the 10000 selected to compute the principal components were analyzed. Our findings suggest a substantial over correction when using more than alleged truly explanatory PCs.

4.1 Data and Quality Controls

The data set considered includes the genetic information of a total of 4798 individuals, 1916 males and 2882 females, measured at 469612 SNPs. All individuals in the sample come from Great Britain, whose population has been shown to contain evidence of stratification in modest extent ([6], [13]). The idea is then to apply the proposed likelihood ratio test for HWE to the data set and to see how it works in correcting for population stratification. Due to the technical complexity of the genotyping process, some statistical quality control steps are required after its completion in order to ensure that genotyping quality is adequate for statistical analyses ([10]). Individuals and markers failing the quality control are excluded from further analyses.

We considered the following criteria for exclusion:

1. allele frequencies of markers;
2. missing data analysis for both markers and individuals;
3. sex check for individuals;
4. estimation of inbreeding coefficient for individuals.

4.1.1 Allele frequencies

Genotypes at different SNPs are usually measured via SNP-chips based on the intensities for each allele. The chips divide the observed data points into clusters corresponding to the three possible genotypes of a SNP. When the minor allele frequencies (MAFs) are very small, it becomes difficult to determine the genotype clusters and errors happen easily because the cluster corresponding to the minor allele homozygous genotype is difficult to identify ([10]). This quality control step has been performed using the software *PLINK* (v1.07) with threshold 0.05 for minor allele frequencies. A total of 101739 SNPs with MAF smaller than 0.05 were excluded from further analyses.

4.1.2 Missing data analysis

The missing data analysis consists in computing the proportion of missing genotypes per SNP and per individual. The presence of missing genotypes is an indication that genotyping error has occurred. Therefore, both SNPs and individuals characterized by a percentage of missing genotypes larger than certain thresholds are considered as problematic and excluded from statistical analyses ([10]). The quality control step has been performed using *PLINK* specifying the thresholds 0.02 and 0.03 for individuals and markers, respectively. No individuals or markers in the dataset failed this quality control.

4.1.3 Sex check

The sex check quality control consists in comparing, for each individual, the gender reported in the data file with the one estimated based on heterozygosity rates using X chromosome data. When the reported gender differs from the gender estimated from given genomic data, the individual is seen as problematic and excluded from further analyses. The sex check has been performed using *PLINK* and led to exclusion of 22 individuals.

4.1.4 Inbreeding

The inbreeding coefficient F represents the probability that an individual in a given population inherits two copies of the same ancestral allele. Estimates of the inbreeding coefficient were computed, per individual, using the software *PLINK* and are characterized by a mean value of 0.001 with a 95% confidence interval, based on normality assumption, equal to $(-0.015, 0.016)$. Given the small value of the mean inbreeding coefficient and its narrow 95% confidence interval centered around zero, no individuals were excluded from further analyses based on this quality control.

4.1.5 Quality control summary

A total of 4776 individuals and 367873 markers remained in the analysis after quality control. In particular:

- 101739 SNPs were excluded based on the allele frequencies check with threshold 0.05;
- 22 individuals were excluded based on the results of the sex check.

No further individuals or markers were excluded based on the remaining considered criteria.

4.2 Linkage Disequilibrium and Data Pruning

As explained in chapter 3, genotype data have been simulated, either for 1000 or 5000 individuals, at 5000 *independent* SNPs. With the term *independent* we mean that the genotypes at different SNPs have been generated as if the corresponding alleles were independently and randomly sampled based on their individual allele frequencies. This is often not the case for genome-wise association datasets, where genetic variants at close loci on the same chromosome are highly correlated and easily transmitted together to the next generation. The presence of statistical associations between alleles at different loci defines the concept of *Linkage Disequilibrium* (LD).

Suppose we have two bi-allelic markers. Let A, a denote the alleles at the first marker and B, b the alleles at the second marker. Let ρ_A, ρ_a, ρ_B, ρ_b denote the allele frequencies at each locus and ρ_{AB}, ρ_{Ab}, ρ_{aB}, ρ_{ab} the frequencies of the four possible haplotypes (specific sets of alleles that offspring inherited from a single parent) at the two loci. The markers are in *Linkage Equilibrium* (LE) if the haplotype frequencies are given by the product of the corresponding

allele frequencies:

$$\begin{aligned}\rho_{AB} &= \rho_A \rho_B \\ \rho_{Ab} &= \rho_A \rho_b \\ \rho_{aB} &= \rho_a \rho_B \\ \rho_{ab} &= \rho_a \rho_b. \end{aligned} \tag{4.2.1}$$

Notice that the haplotype frequencies are the joint probabilities that the two corresponding alleles occur together in a randomly selected gamete. The right-hand sides of equations 4.2.1 are the probabilities that the two alleles *independently* occur in the same gamete, which are given by the products of the individual allele frequencies. If equations 4.2.1 fail to hold, the markers are said to be in LD.

Since many of the 367873 markers that remained in the dataset after quality control are in LD, the software *PLINK* has been used to get a pruned subset of 71872 SNPs approximately in LE with each other. The *PLINK* command for getting the pruned subset requires the specification of three parameters; namely, the window size in SNPs, the number of SNPs to shift the window at each step, the VIF threshold. The VIF is $1/(1 - R^2)$ where R^2 is the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously. The values specified for the three parameters were 100, 5 and 1.5, respectively.

4.3 PCA and Evidence of Population Substructure

Principal components used to perform the test were computed based on a subset of 10000 out of the pruned set of 71872 approximately independent SNPs.
In order to check if different subsets of 10000 (approximately independent) SNPs give a stable picture of population substructure, four different samples of size 10000 were randomly selected and the PCA was performed on each of them. Plots of the individuals along the first two principal components are shown in figure 4.1 and reveal, in all four cases, some evidence of population stratification, although the pictures differ with respect to rotation and the angular difference between two substrata. Considering all SNPs would yield a more accurate assessment of substructure. On the other hand, the use of about 10000 SNPs for computing PCs is a common practice in applied studies and it is considered as a good approximation to capture and correct for confounding due to population stratification. Therefore, a further random sample of size 10000 out of the pruned set of 71872 markers has been drawn, after filtering for those markers whose p-value for the standard Hardy-Weinberg equilibrium test was smaller than 10^{-10}, and

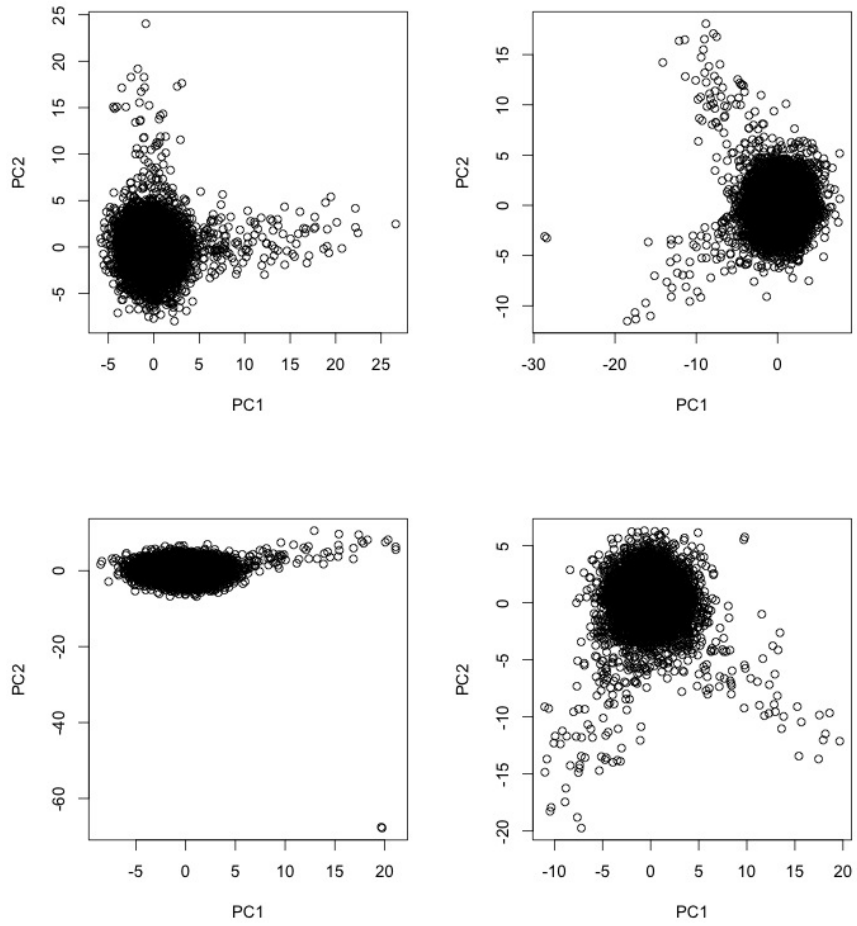


Figure 4.1: Plots of the individuals along the first two principal components computed on four different randomly selected samples of 10000 SNPs out of the 71872 approximately in LE.

used to compute the principal components to be included in the model for individual allele frequencies. The reason for the filtering is to exclude markers presenting very low p-values which are more likely the consequence of genotyping errors rather than effects of population stratification. Figure 4.2 shows some evidence of population substructure as captured by the selected subset of 10000 markers. It can be seen that most of the samples are located in the square $[-5, 5] \times [-5, 5]$ and that the first two principal components do not seem to differ much in terms of variability present in the data, as is also confirmed by the similar magnitudes of the correspondent standard deviations[1].

[1]The square roots of the eigenvalues of the covariance (or correlation) matrix of the genetic data matrix

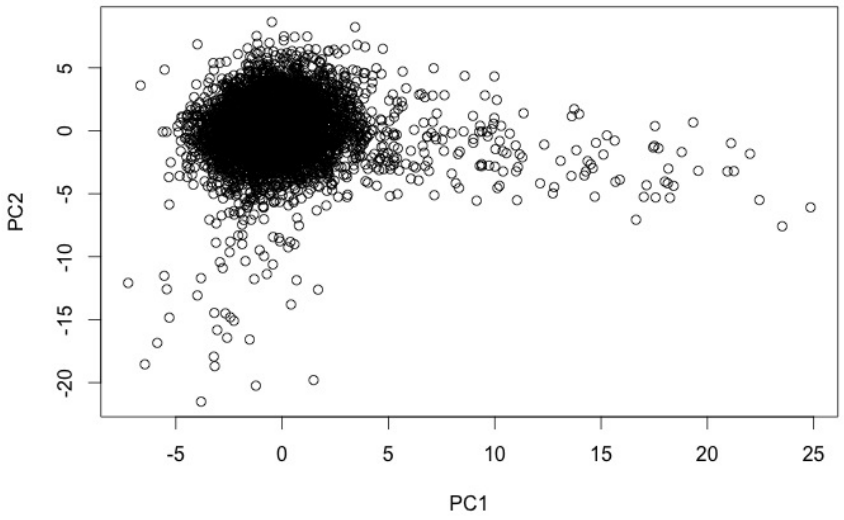


Figure 4.2: Plots of individuals along the first two principal components computed on a specific sample of size 10000, randomly selected from the pruned set of 71872 markers approximately in linkage equilibrium after filtering for the markers with a p-value for the standard HWE test smaller than 10^{-10}.

4.4 Results and Remarks

For computational reasons, two subsets were analyzed. The new likelihood ratio test has been performed on two different subsets of markers out of the 10000 selected to compute the principal components, namely:

- a subset of markers whose p-value for the standard HWE test fell in the interval $(10^{-10}, 10^{-6})$;
- a random subset of size 1000.

The first of the two subsets mentioned above has been considered in order to see to which extent the inclusion of certain principal components improves the fit with respect to Hardy-Weinberg equilibrium for those markers presenting the smallest p-values for the standard test. The second subset has been considered to get the distribution of the p-values under the null hypothesis that Hardy-Weinberg equilibrium holds in the population. The maximum number of principal components considered was four. The first four principal components were added to the model both individually and sequentially. The correspondent standard deviations were 2.66, 2.59, 2.51 and 2.48, respectively. In figure 4.3 the qq-plots of the p-values for eight different tests are shown. In particular, the top left panel shows the qq-plot for the standard HWE test while the following panels show the plots obtained when applying the likelihood ratio test to the specific subset of markers, first including the different principal components separately and then sequentially. It

can be seen that in all cases, most of the p-values are a lot higher than their expected values under the null hypothesis (HWE). A slight improvement in terms of goodness-of-fit can be observed when including one of the principal components. The improvement looks the same no matter which of the separate principal components is included in the model. This is probably due to the fact that the eigenvalues of the principal components considered are quite similar to each other. The improvement in the fit becomes slightly more evident when adding the principal components sequentially. However, when all the first four principal components are added together to the model, an increase in high p-values (left bottom p-values in figure 4.3) appears, indicating some overfitting issues in the procedure: the included principal components are not adjusting for possible population stratification only but also for random noise present in the data. Figure 4.4 shows the plots of the p-values for the different tests on a sample of 1000 markers randomly selected from the subset of the 10000 approximately in linkage equilibrium. Since the markers are here selected randomly (and, in particular, independently from their p-values), the different plots should give, for each test, an idea of the distribution of the corresponding p-values under the null hypothesis. Even in this case it can be seen that many p-values are far bigger than expected. The top six plots look almost the same, indicating that adding the first four principal components one at the time does not improve the fit with respect to HWE. The same holds for the case in which the first two principal components are added together. Some more evident improvement is obtained when adding the first three principal components together, but then, again, when including the four principal components together quite strong evidence of overfitting appears.

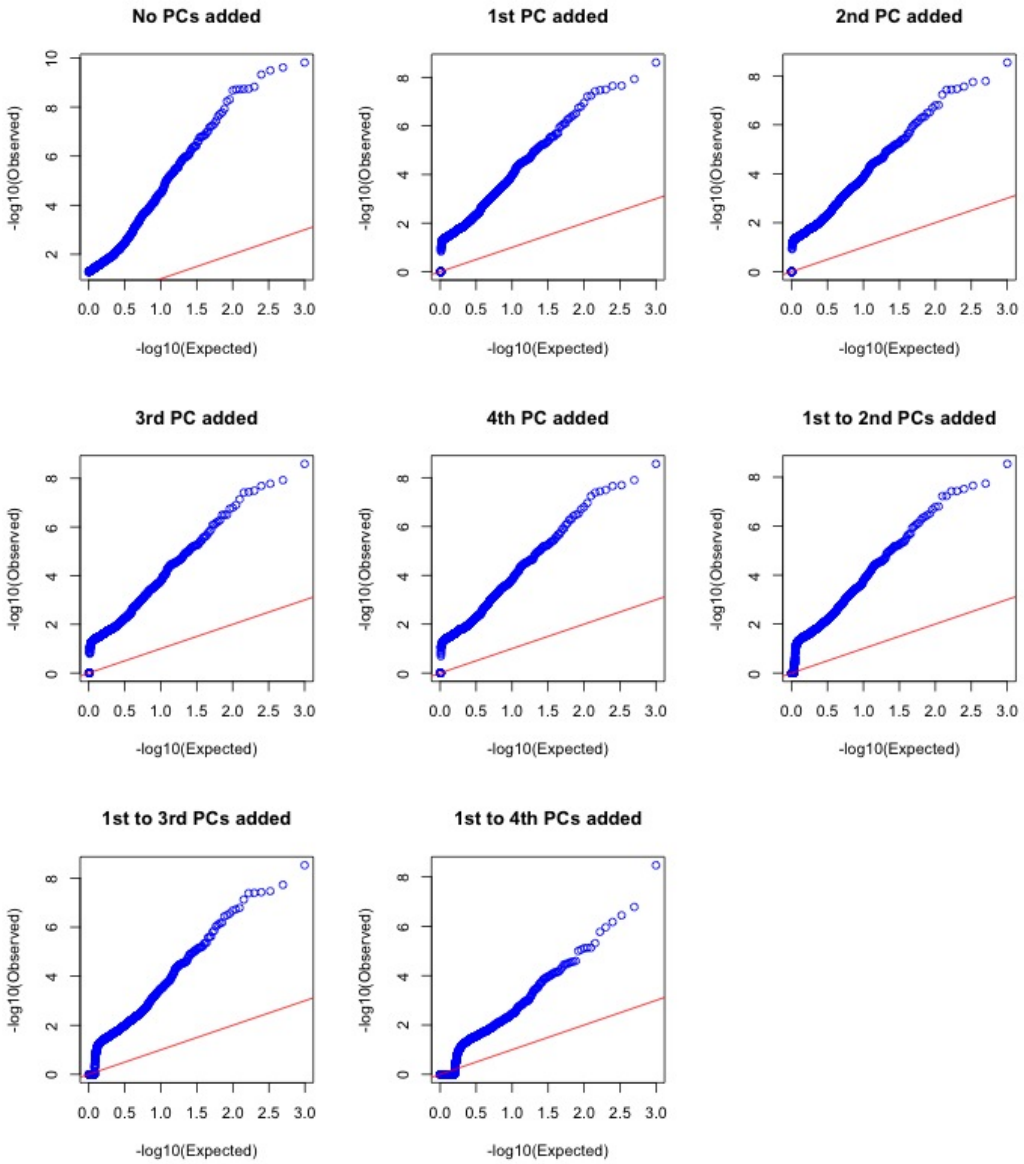


Figure 4.3: Uniform QQ-plots of the p-values obtained by applying the proposed likelihood ratio test to a subset of SNPs (out of the 10000 selected to compute the principal components) whose p-value for the standard HWE test are in the interval $(10^{-10}, 10^{-6})$. *The top left panel shows the qq-plot for the standard HWE. The following panels show, from left to right, the plots obtained when applying the likelihood ratio test to the specific subset of markers, first including the different principal components separately and then sequentially.*

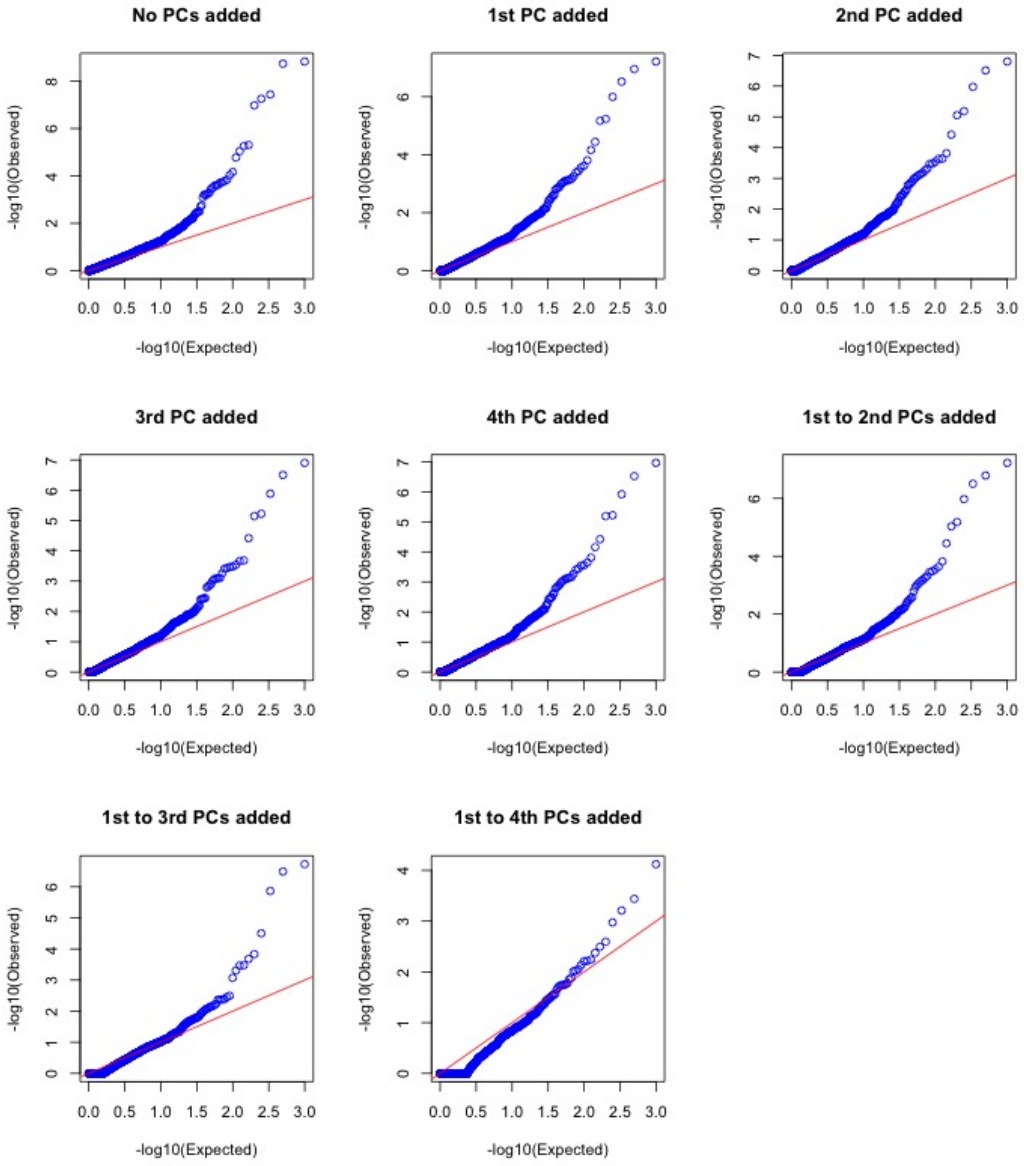


Figure 4.4: Uniform QQ-plots of the p-values obtained by applying the proposed likelihood ratio test to a sample of 1000 markers randomly selected from the subset of the 10000 used to compute the principal components. The top left panel shows the qq-plot for the standard HWE. The following panels show, from left to right, the plots obtained when applying the likelihood ratio test to the specific subset of markers, first including the different principal components separately and then sequentially.

Chapter 5

Discussion

We developed a likelihood-based goodness-of-fit test to evaluate the effect of covariates to model potential deviations from HWE. In particular, the main interest was to investigate confounding due to population stratification. As most often principal components of standardized genotypes are used to model population stratification, these were chosen as covariates. For example, in genome wide association studies correction with principal components is common practice. In a wider context, principal components are used to correct for so-called *batch effects* ([9], [12]). Batch effects generally refer to technical sources of variation that have been added to samples during handling and measurement. In many studies, samples cannot be handled and measured all at once due to the number of samples. The samples are then split into subsets and each subset is measured on a different day. Technical variation is inevitably added to the data every time samples are measured and it is reasonable to assume that the technical variation comes from the same distribution for samples measured on the same day but it differs across different days. We can refer to this phenomenon as *batch effect* and the goal is to correct for this technical variation and to avoid that it confounds with biological variation in the samples. The common way to do that is to choose a plotting tool, like PCA, assign different colors to samples labeled in different days and see how the colors distribute in the plot (for example along the first two axis of maximal variation). This is exactly what we do to correct for population stratification. In particular, we generate population composed by discrete strata where HWE proportions hold, we assign a different color to different strata and we plot the individual along the first two principal components computed based on the genetic data matrix. In this respect, population stratification correspond to a batch effect. Principal components are linear combinations of different SNPs and can be thought as containing a (biological) signal plus some noise. In this context the signal represents *true* deviation from HWE once stratification is accounted for, while

stratification is interpreted as noise in the data. The goal is then to try to test for HWE while accounting for population stratification. However, it is important to keep in mind that every time we try to correct for noise in the data we take the risk of correcting not only for the noise, but for the actual, biological, signal as well.

The results of the simulations we performed indicate a general conservative behavior of the proposed likelihood ratio test when principal components are used to model population stratification. In particular, this is more evident for scenarios related to simulations with fewer individuals (1000 instead of 5000). This is likely due to the fact that the distribution of the likelihood ratio test statistic for our proposed test is not exact but asymptotic. Therefore the approximation improves as the number of individuals in the population increases. Focusing on scenarios related to homogeneous populations with 5000 individuals, the behavior of the likelihood ratio test, as assessed on the basis of the four α-levels considered, looks reasonably good when a maximum of three principal components are used to model individual allele frequencies. However, the inclusion of four principal components in the model for the individual allele frequencies results in conservative behavior. A possible explanation for that is that by including four principal components an overfit is induced as we are actually correcting not only for population stratification (for which a maximum of two are informative) but for arbitrary variation in the data. This overfit leads to a higher number of high p-values and a smaller number of low p-values, as shown by the TS values for the corresponding likelihood ratio test. This result implies that the number of principal components used to correct for population stratification should be carefully judged to avoid overcorrection.

We remarked that, even in the case of stratified populations characterized by completely separated strata (i.e. top right and middle left panels of Figure A.2), the results for the standard HWE test do not show an increase in the observed proportions of low p-values as judged by a number of α-levels. In principle, every stratified population violates HWE proportions since population stratification generally results in a decrease in the frequency of heterozygotes compared to a population in HWE ([10]). A possible explanation for this fact is that based on the simulation model, the population stratification may be driven by a small number of SNPs whose allele frequencies substantially differ between substrata while for the majority of SNPs the differences in allele frequency between strata is smaller, so that most of the SNPs do not suffer from the expected decrease in the frequency of heterozygotes.

The results of the data analysis show a pattern which seems to be consistent with the results of the simulations. However, it is important to consider the limitations related to all the performed analyses. As already mentioned, the two different set of simulations have been run for 10 and 100 iterations for each scenario, respectively. The small numbers of replications per scenario are due

to the long computing time required by the testing procedure and have an impact, for example, on the possibility to get accurate estimates for the standard error of the tail strength. Regarding the data analysis several steps were undertaken in order to obtain the data set on which the likelihood-ratio test was applied. In particular, after the pruning applied to the complete data matrix in order to get a subset of approximately independent SNPs, a random set of 10000 SNPs was selected to compute the principal components used to model allele frequencies of individuals. Studies have shown that a sufficient amount of SNPs to get a complete picture of the population substructure is in the order of tens of thousand ([11]). The use of a subset of 10000 SNPs is in line to what is commonly done in practice and can be considered as a reasonably good approximation. However, computing the principal components based on the 71872 obtained after the pruning would have led to a more accurate assessment of population substructure.
Furthermore, for computational reasons only two subsets of SNPs out of the 10000 used to compute the principal components were finally analyzed. It is also important to notice that the behavior of the likelihood ratio test on the two selected samples were assessed graphically by means of QQ-plots (against a uniform distribution). This was not possible for the results of the simulations since it would have been problematic to store all the p-values for all the tests performed. Indeed the simulation function was programmed in a way to extract and save for each test only the proportions of p-values smaller that the few considered α-levels and the values of the tail strength. Therefore, no QQ-plots were produced relatively to the results of the simulations and a direct comparison with the results of the data analysis was then not possible. However, the conclusions that can be drawn based on the data analysis results are in accordance to the conclusions based on the simulation results and provide evidence of an increasing overfitting issue as more principal components are considered in the likelihood ratio test and suggesting that by including four principal components we are actually correcting for more than simple population stratification.

In conclusion, we have established a goodness-of-fit test for HWE in the presence of covariates and evaluated the performance of the test by means of simulations and a data analysis. Results suggest that the common practice of including a fixed number of principal components in regression models to correct for population stratification leads to conservative testing procedures and can be optimized by a more precise choice of principal components.

Appendix A

Tables and Figures

Scenario	T	M	N	N_{strata}	p
1	500	5000	1000	1	2
2	500	5000	1000	1	3
3	500	5000	1000	1	4
4	500	5000	1000	2	2
5	500	5000	1000	2	3
6	500	5000	1000	2	4
7	500	5000	1000	3	2
8	500	5000	1000	3	3
9	500	5000	1000	3	4
10	500	5000	5000	1	2
11	500	5000	5000	1	3
12	500	5000	5000	1	4
13	500	5000	5000	2	2
14	500	5000	5000	2	3
15	500	5000	5000	2	4
16	500	5000	5000	3	2
17	500	5000	5000	3	3
18	500	5000	5000	3	4
19	1000	5000	1000	1	2
20	1000	5000	1000	1	3
21	1000	5000	1000	1	4
22	1000	5000	1000	2	2

23	1000	5000	1000	2	3
24	1000	5000	1000	2	4
25	1000	5000	1000	3	2
26	1000	5000	1000	3	3
27	1000	5000	1000	3	4
28	1000	5000	5000	1	2
29	1000	5000	5000	1	3
30	1000	5000	5000	1	4
31	1000	5000	5000	2	2
32	1000	5000	5000	2	3
33	1000	5000	5000	2	4
34	1000	5000	5000	3	2
35	1000	5000	5000	3	3
36	1000	5000	5000	3	4

Table A.1: ***Simulation parameters****. T represents the number of generation based on which the allele frequency changes are simulated. M is the number of simulated markers, N the total number of samples in the simulated population and p the maximum number of principal components included in the model for individual allele frequencies.*

Scenario	α	ST	(1)	(2)	(1, 2)
1	.05	.05039	.05083	.05089	.05158
	.01	.01003	.00786	.00787	.00805
	.001	.00097	.00019	.00019	.00020
	.0001	.00008	.00000	.00000	.00000
10	.05	.05037	.05081	.05067	.05073
	.01	.01024	.01028	.01031	.01037
	.001	.00107	.00108	.00107	.00108
	.0001	.00011	.00011	.00011	.00011
19	.05	.05010	.05059	.05058	.05129
	.01	.01014	.00791	.00787	.00809
	.001	.00105	.00021	.00021	.00023
	.0001	.00011	.00000	.00000	.00000
28	.05	.04996	.05013	.05009	.05006
	.01	.00998	.00996	.01004	.01009
	.001	.00096	.00097	.00095	.00097
	.0001	.00011	.00011	.00011	.00012

Table A.2: ***Homogeneous populations with*** $N_{drift} = 600$***.*** *Scenarios* 1 *and* 19 *refer to populations of* $N = 1000$ *samples, while scenarios* 10 *and* 28 *refer to populations of* $N = 5000$ *samples. The number of markers* M *is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(1, 2)
4	.05	.04988	.05015	.05021	.05057
	.01	.01005	.00773	.00783	.00792
	.001	.00100	.00021	.00022	.00022
	.0001	.00013	.00000	.00000	.00000
13	.05	.05116	.05005	.05122	.04991
	.01	.01036	.00992	.01038	.01001
	.001	.00107	.00099	.00106	.00100
	.0001	.00011	.00009	.00010	.00009
22	.05	.05441	.05054	.05427	.05109
	.01	.01146	.00775	.00882	.00787
	.001	.00134	.00026	.00029	.00020
	.0001	.00019	.00006	.00001	.00000
31	.05	.07103	.05013	.06995	.04971
	.01	.02069	.01020	.02020	.01018
	.001	.00457	.00111	.00440	.00106
	.0001	.00143	.00016	.00136	.00011

Table A.3: ***Populations with two strata and*** $N_{drift} = 600$***.*** *(i) Scenario* 4 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 13*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 22*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 31*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(1, 2)
7	.05	.05100	.05083	.05080	.05092
	.01	.01016	.00787	.00781	.00792
	.001	.00112	.00024	.00022	.00023
	.0001	.00011	.00001	.00001	.00001
16	.05	.05159	.05057	.05066	.04995
	.01	.01042	.00998	.01004	.00997
	.001	.00113	.00107	.00108	.00107
	.0001	.00012	.00011	.00010	.00011
25	.05	.05565	.05235	.05252	.05105
	.01	.01201	.00833	.00849	.00799
	.001	.00140	.00026	.00025	.00021
	.0001	.00017	.00002	.00002	.00000
34	.05	.07602	.05889	.05960	.04945
	.01	.02161	.01366	.01411	.01003
	.001	.00428	.00189	.00210	.00099
	.0001	.00100	.00032	.00038	.00009

Table A.4: ***Populations with three strata and $N_{drift} = 600$.*** *(i) Scenario* 7 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 16*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 25*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 34*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(1, 2)
4	.05	.05476	.05048	.05452	.05090
	.01	.01120	.00732	.00850	.00744
	.001	.00122	.00028	.00030	.00024
	.0001	.00018	.00006	.00000	.00000
13	.05	.07000	.04888	.06914	.04828
	.01	.01994	.01006	.01944	.01020
	.001	.00428	.00080	.00412	.00076
	.0001	.00138	.00010	.00130	.00004
22	.05	.10054	.04958	009882	.05040
	.01	.04142	.00778	.03542	.00776
	.001	.01532	.00028	.00862	.00028
	.0001	.00762	.00000	.00280	.00000
31	.05	.19268	.04926	.18986	.04748
	.01	.11698	.00992	.11502	.00986
	.001	.07320	.00110	.07214	.00110
	.0001	.05192	.00012	.05126	.00010

Table A.5: ***Populations with two strata and $N_{drift} = 300$.*** *(i) Scenario* 4 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 13*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 22*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 31*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(1, 2)
7	.05	.05542	.05238	.05216	.05086
	.01	.01202	.00794	.00836	.00788
	.001	.00112	.00014	.00020	.00020
	.0001	.00014	.00000	.00000	.00000
16	.05	.07430	.05788	.05798	.04920
	.01	.02052	.01278	.01286	.00934
	.001	.00380	.00168	.00198	.00104
	.0001	.00100	.00028	.00042	.00014
25	.05	.11752	.07566	.07730	.05210
	.01	.04706	.01814	.01940	.00746
	.001	.01610	.00214	.00242	.00010
	.0001	.00640	.00034	.00034	.00000
34	.05	.25510	.13618	.14098	.04678
	.01	.15938	.06696	.07104	.00980
	.001	.09540	.03330	.03666	.00094
	.0001	.06296	.01910	.02292	.00010

Table A.6: ***Populations with three strata and $N_{drift} = 300$.*** *(i) Scenario* 7 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 16*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 25*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 34*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
2	.05	.0502	.0509	.0508	.0508	.0515	.0518
	.01	.0100	.0078	.0078	.0078	.0080	.0082
	.001	.0010	.0002	.0002	.0002	.0002	.0003
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
11	.05	.0500	.0503	.0503	.0502	.0503	.0492
	.01	.0100	.0100	.0101	.0101	.0101	.0101
	.001	.0009	.0010	.0010	.0010	.0010	.0010
	.0001	.0001	.0001	.0001	.0001	.0001	.0001
20	.05	.0500	.0505	.0505	.0505	.0512	.0515
	.01	.0103	.0080	.0080	.0081	.0082	.0084
	.001	.0010	.0002	.0002	.0002	.0002	.0002
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
29	.05	.0500	.0501	.0502	.0503	.0501	.0489
	.01	.0100	.0102	.0101	.0102	.0103	.0102
	.001	.0010	.0010	.0010	.0010	.0010	.0010
	.0001	.0001	.0001	.0001	.0001	.0001	.0001

Table A.7: ***Homogeneous populations with $N_{drift} = 600$.*** *Scenarios* 2 *and* 20 *refer to populations of* $N = 1000$ *samples, while scenarios* 11 *and* 29 *refer to populations of* $N = 5000$ *samples. The number of markers* M *is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
5	.05	.0501	.0505	.0505	.0505	.0511	.0513
	.01	.0101	.0077	.0079	.0078	.0079	.0081
	.001	.0010	.0002	.0002	.0002	.0002	.0003
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
14	.05	.0515	.0502	.0513	.0513	.0500	.0485
	.01	.0105	.0101	.0105	.0105	.0101	.0100
	.001	.0010	.0010	.0011	.0011	.0010	.0010
	.0001	.0001	.0001	.0001	.0001	.0001	.0001
23	.05	.0540	.0506	.0539	.0538	.0511	.0511
	.01	.0116	.0078	.0088	.0088	.0079	.0081
	.001	.0013	.0003	.0003	.0003	.0002	.0002
	.0001	.0002	.0001	.0000	.0000	.0000	.0000
32	.05	.0718	.0503	.0706	.0706	.0499	.0478
	.01	.0205	.0102	.0200	.0199	.0102	.0099
	.001	.0045	.0011	.0043	.0044	.0011	.0011
	.0001	.0014	.0001	.0013	.0013	.0001	.0001

Table A.8: ***Populations with two strata and $N_{drift} = 600$.*** *(i) Scenario* 5 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 14*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 23*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 32*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
8	.05	.0504	.0505	.0505	.0507	.0508	.0510
	.01	.0102	.0078	.0079	.0079	.0079	.0080
	.001	.0010	.0002	.0002	.0002	.0002	.0002
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
17	.05	.0519	.0507	.0507	.0517	.0499	.0483
	.01	.0106	.0103	.0103	.0106	.0101	.0100
	.001	.0011	.0011	.0011	.0011	.0011	.0011
	.0001	.0001	.0001	.0001	.0001	.0001	.0001
26	.05	.0550	.0519	.0520	.0546	.0506	.0507
	.01	.0116	.0081	.0081	.0088	.0078	.0078
	.001	.0013	.0003	.0003	.0003	.0002	.0002
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
35	.05	.0756	.0584	.0594	.0740	.0490	.0471
	.01	.0213	.0138	.0140	.0207	.0100	.0097
	.001	.0043	.0019	.0021	.0041	.0010	.0010
	.0001	.0010	.0003	.0004	.0010	.0001	.0001

Table A.9: ***Populations with three strata and $N_{drift} = 600$.*** *(i) Scenario* 8 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 17*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 26*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 35*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 100 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
5	.05	.0544	.0499	.0542	.0541	.0503	.0502
	.01	.0118	.0077	.0091	.0090	.0076	.0077
	.001	.0012	.0002	.0002	.0002	.0001	.0002
	.0001	.0001	.0000	.0000	.0000	.0000	.0000
14	.05	.0690	.0484	.0679	.0680	.0480	.0464
	.01	.0193	.0096	.0188	.0189	.0095	.0093
	.001	.0043	.0008	.0040	.0041	.0008	.0008
	.0001	.0012	.0000	.0012	.0012	.0000	.0000
23	.05	.1018	.0501	.1002	.1002	.0503	.0497
	.01	.0415	.0073	.0357	.0356	.0074	.0075
	.001	.0156	.0002	.0090	.0091	.0002	.0001
	.0001	.0078	.0001	.0027	.0027	.0000	.0000
32	.05	.1936	.0504	.1904	.1902	.0492	.0460
	.01	.1198	.0102	.1177	.1178	.0103	.0094
	.001	.0726	.0011	.0714	.0715	.0010	.0009
	.0001	.0510	.0002	.0503	.0503	.0002	.0002

Table A.10: ***Populations with two strata and $N_{drift} = 300$.*** *(i) Scenario* 5 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 14*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 23*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 32*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	α	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
8	.05	.0546	.0515	.0518	.0541	.0501	.0506
	.01	.0118	.0084	.0085	.0090	.0081	.0083
	.001	.0013	.0003	.0003	.0003	.0002	.0002
	.0001	.0002	.0000	.0000	.0000	.0000	.0000
17	.05	.0763	.0587	.0602	.0742	.0503	.0479
	.01	.0208	.0127	.0135	.0201	.0095	.0091
	.001	.0039	.0017	.0019	.0037	.0010	.0009
	.0001	.0009	.0002	.0004	.0009	.0001	.0001
26	.05	.1182	.0751	.0775	.1149	.0512	.0508
	.01	.0475	.0186	.0208	.0409	.0082	.0083
	.001	.0164	.0024	.0027	.0077	.0003	.0003
	.0001	.0063	.0004	.0005	.0014	.0000	.0000
35	.05	.2566	.1378	.1433	.2528	.0494	.0458
	.01	.1620	.0690	.0725	.1596	.0099	.0093
	.001	.0968	.0345	.0372	.0949	.0008	.0009
	.0001	.0646	.0212	.0236	.0636	.0001	.0001

Table A.11: ***Populations with three strata and $N_{drift} = 300$.*** *(i) Scenario* 8 *refers to a population of* $N = 1000$ *samples where the allele frequency changes are simulated based on* $T = 500$ *generations. (ii) Scenario* 17*:* $N = 5000$ *and* $T = 500$*. (iii) Scenario* 26*:* $N = 1000$ *and* $T = 1000$*. (iv) Scenario* 35*:* $N = 5000$ *and* $T = 1000$*. The number* M *of SNPs considered is* 5000 *for all the scenarios. The results are averaged over the* 10 *repetitions for each scenario and represent the proportions of p-values smaller or equal the* α*-level specified in the corresponding line. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	ST	(1)	(2)	(1,2)
1	.0014	-.0076	-.0073	-.0099
10	.0028	-.0015	-.0019	-.0149
19	.0007	-.0077	-.0079	-.0103
28	-.0010	-.0057	-.0057	-.0185
4	.0010	-.0098	-.0090	-.0144
13	.0088	-.0076	.0021	-.0214
22	.0238	-.0095	.0112	-.0167
31	.0993	-.0101	.0889	-.0304
7	.0037	-.0074	-.0076	-.0134
16	.0103	-.0036	-.0029	-.0217
25	.0294	-.0005	.0007	-.0183
34	.1207	.0364	.0406	-.0399

Table A.12: ***Tail strength for simulations with $N_{drift} = 600$.*** *(i) Scenarios* 1, 10, 19 *and* 28 *refer to homogeneous populations. (ii) Scenarios* 4, 13, 22 *and* 31 *refer to populations with two strata. (iii) Scenarios* 7, 16, 25 *and* 34 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 100 *repetitions for each scenario and represent the tail strength values of the different tests performed. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
2	.0012	-.0072	-.0073	-.0075	-.0097	-.0273
11	.0010	-.0037	-.0038	-.0038	-.0167	-.0495
20	.0018	-.0068	-.0069	-.0071	-.0093	-.0274
29	.0000	-.0046	-.0046	-.0047	-.0178	-.0504
5	-.0000	-.0106	-.0097	-.0097	-.0153	-.0323
14	.0094	-.0073	.0026	.0025	-.0206	-.0567
23	.0235	-.0096	.0107	.0108	-.0177	-.0422
32	.1031	-.0067	.0926	.0926	-.0261	-.0696
8	.0027	-.0087	-.0086	-.0076	-.0143	-.0332
17	.0093	-.0046	-.0045	.0016	-.0234	-.0638
26	.0284	-.0013	-.0003	.0141	-.0190	-.0418
35	.1205	.0362	.0396	.1081	-.0403	-.0776

Table A.13: ***Tail strength for simulations with $N_{drift} = 600$.*** *(i) Scenarios* 2, 11, 20 *and* 29 *refer to homogeneous populations. (ii) Scenarios* 5, 14, 23 *and* 32 *refer to populations with two strata. (iii) Scenarios* 8, 17, 26 *and* 35 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 100 *repetitions for each scenario and represent the tail strength values of the different tests performed. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	ST	(1)	(2)	(1,2)
1	-.0012	-.0101	-.0105	-.0129
10	-.0018	-.0067	-.0071	-.0195
19	-.0091	-.0180	-.0180	-.0203
28	.0014	-.0034	-.0038	-.0174
4	.0230	-.0105	.0101	-.0187
13	.0973	-.0164	.0866	-.0378
22	.1934	-.0139	.1809	-.0270
31	.3824	-.0159	.3716	-.0504
7	.0284	-.0021	-.0024	-.0204
16	.1146	.0314	.0330	-.0447
25	.2492	.1081	.1157	-.0246
34	.4861	.2694	.2807	-.0750

Table A.14: ***Tail strength for simulations with $N_{drift} = 300$.*** *(i) Scenarios* 1, 10, 19 *and* 28 *refer to homogeneous populations. (ii) Scenarios* 4, 13, 22 *and* 31 *refer to populations with two strata. (iii) Scenarios* 7, 16, 25 *and* 34 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 10 *repetitions for each scenario and represent the tail strength values of the different tests performed. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

Scenario	ST	(1)	(2)	(3)	(1,2)	(1,2,3)
2	-.0030	-.0114	-.0107	-.0115	-.0129	-.0304
11	-.0053	-.0095	-.0097	-.0101	-.0225	-.0578
20	.0029	-.0059	-.0051	-.0058	-.0085	-.0269
29	.0066	.0010	.0012	.0009	-.0137	-.0477
5	.0264	-.0085	.0133	.0133	-.0163	-.0407
14	.0935	-.0164	.0834	.0825	-.0360	-.0793
23	.1984	-.0131	.1855	.1857	-.0273	-.0659
32	.3857	-.0163	.3759	.3755	-.0514	-.1120
8	.0256	-.0035	-.0020	.0114	-.0206	-.0426
17	.1212	.0355	.0401	.1080	-.0401	-.0806
26	.2479	.1040	.1123	.2321	-.0314	-.0770
35	.4886	.2772	.2829	.4778	-.0594	-.1190

Table A.15: ***Tail strength for simulations with $N_{drift} = 300$.*** *(i) Scenarios* 2, 11, 20 *and* 29 *refer to homogeneous populations. (ii) Scenarios* 5, 14, 23 *and* 32 *refer to populations with two strata. (iii) Scenarios* 8, 17, 26 *and* 35 *refer to populations with three strata. The values of the parameters defining the scenarios are reported in* Table *A.1. The results are averaged over the* 10 *repetitions for each scenario and represent the tail strength values of the different tests performed. The column* ST *refers to the standard HWE test. The numbers between parentheses in the following columns indicate the principal component(s) considered when testing for HWE via the proposed likelihood ratio test.*

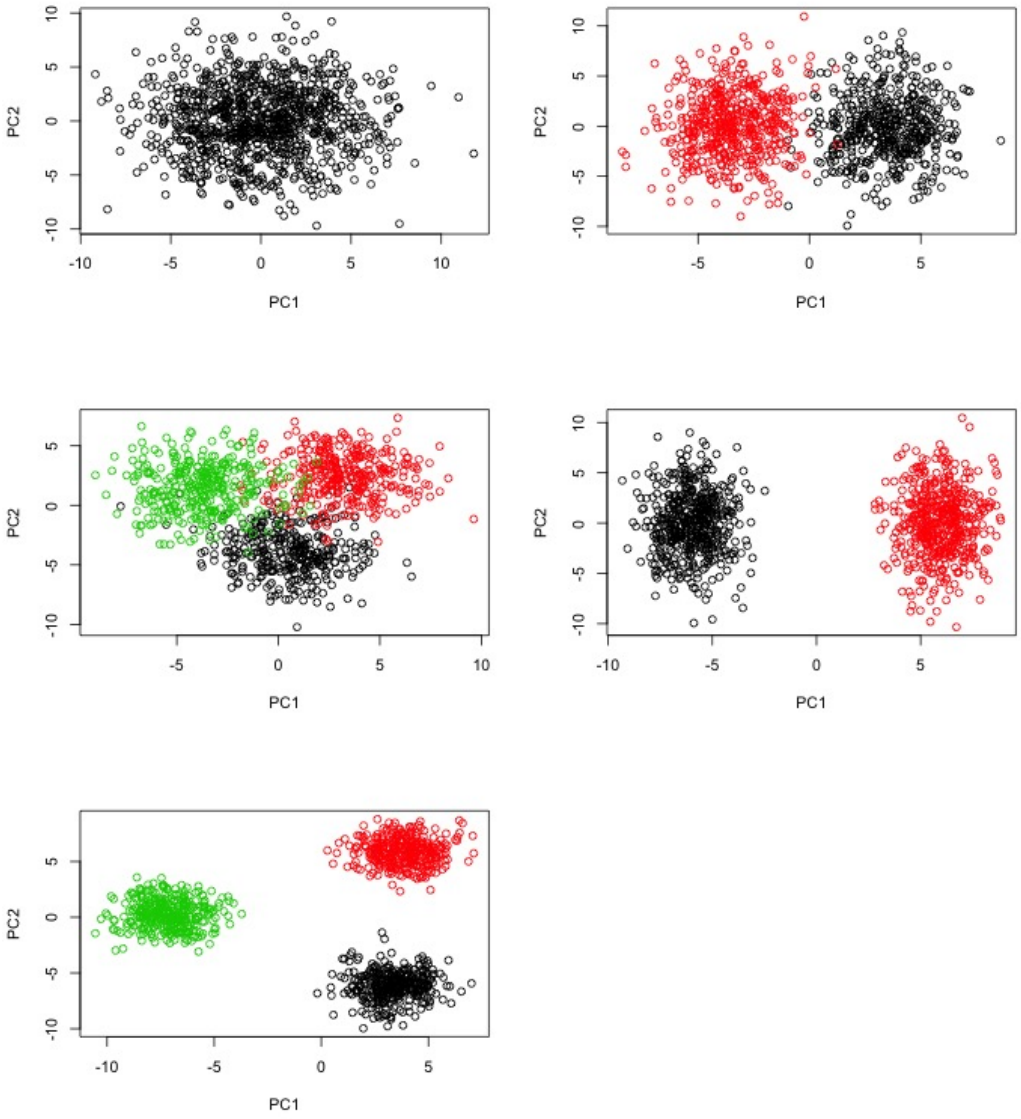


Figure A.1: ***Simulation scenarios with*** $N = 1000$ ***and*** $N_{drift} = 600$. *Starting from the top left panel are represented, in order, a single homogeneous population, two stratified populations with two less or more distant strata and two population with three less or more distant strata.*

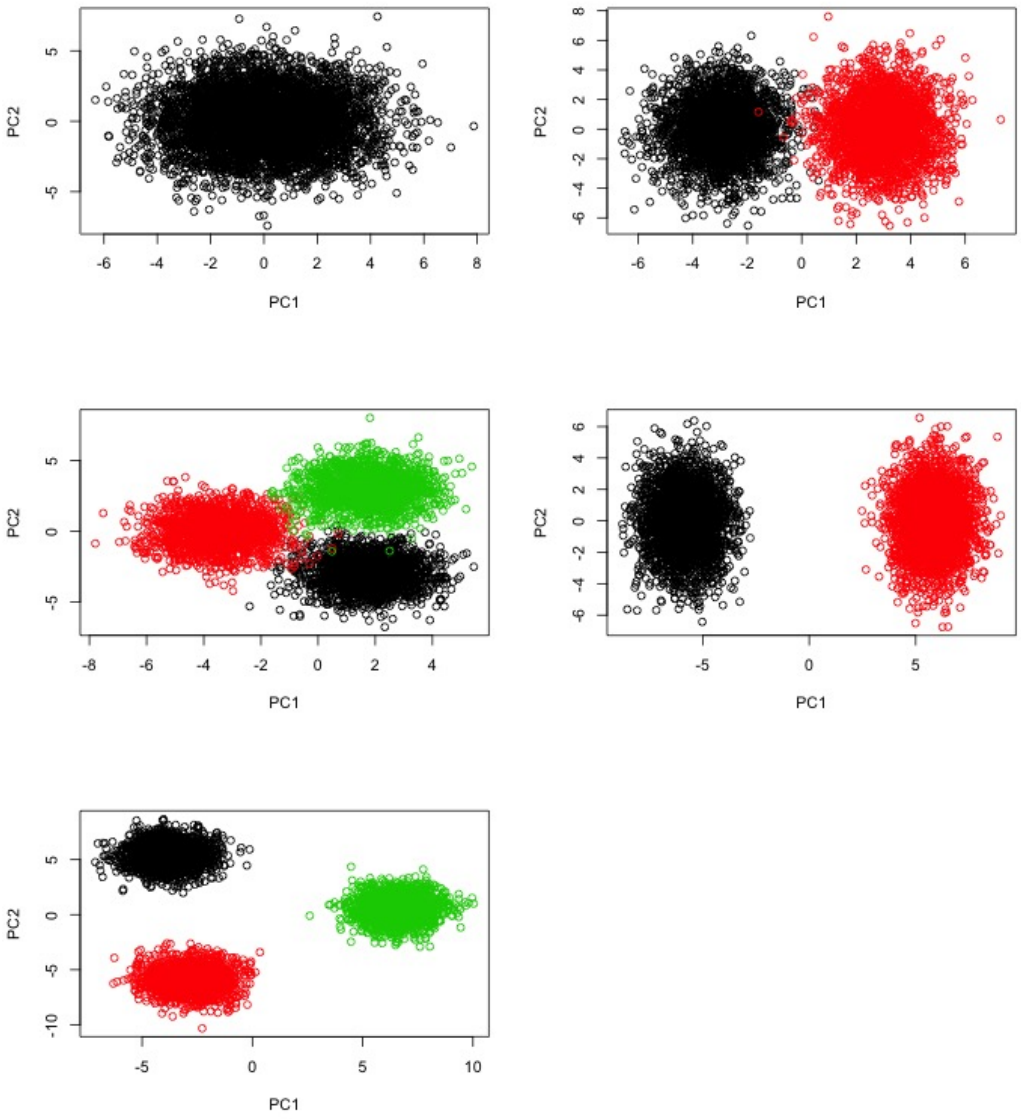


Figure A.2: ***Simulation scenarios with*** $N = 5000$ ***and*** $N_{drift} = 600$. *Starting from the top left panel are represented, in order, a single homogeneous population, two stratified populations with two less or more distant strata and two population with three less or more distant strata.*

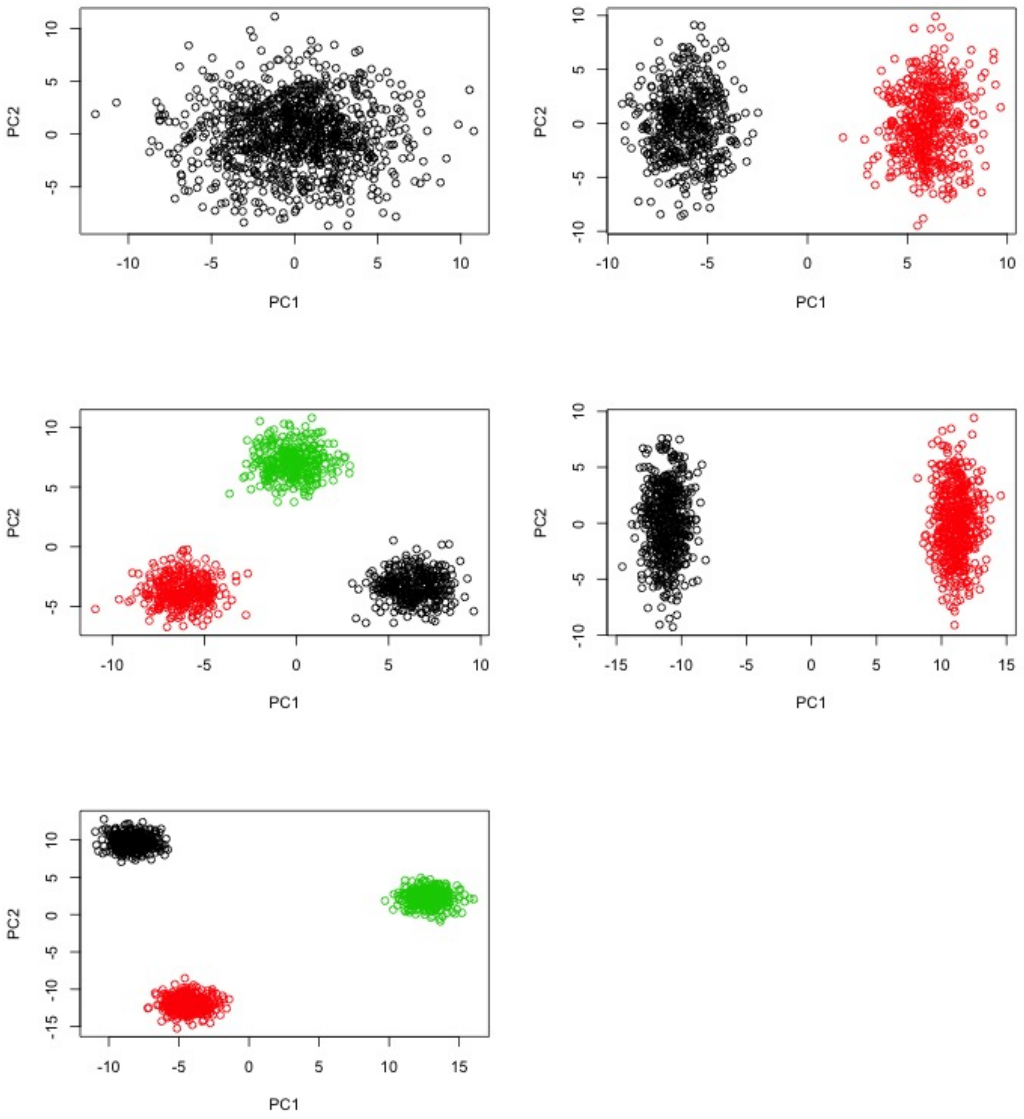


Figure A.3: ***Simulation scenarios with*** $N = 1000$ ***and*** $N_{drift} = 300$. *Starting from the top left panel are represented, in order, a single homogeneous population, two stratified populations with two less or more distant strata and two population with three less or more distant strata.*

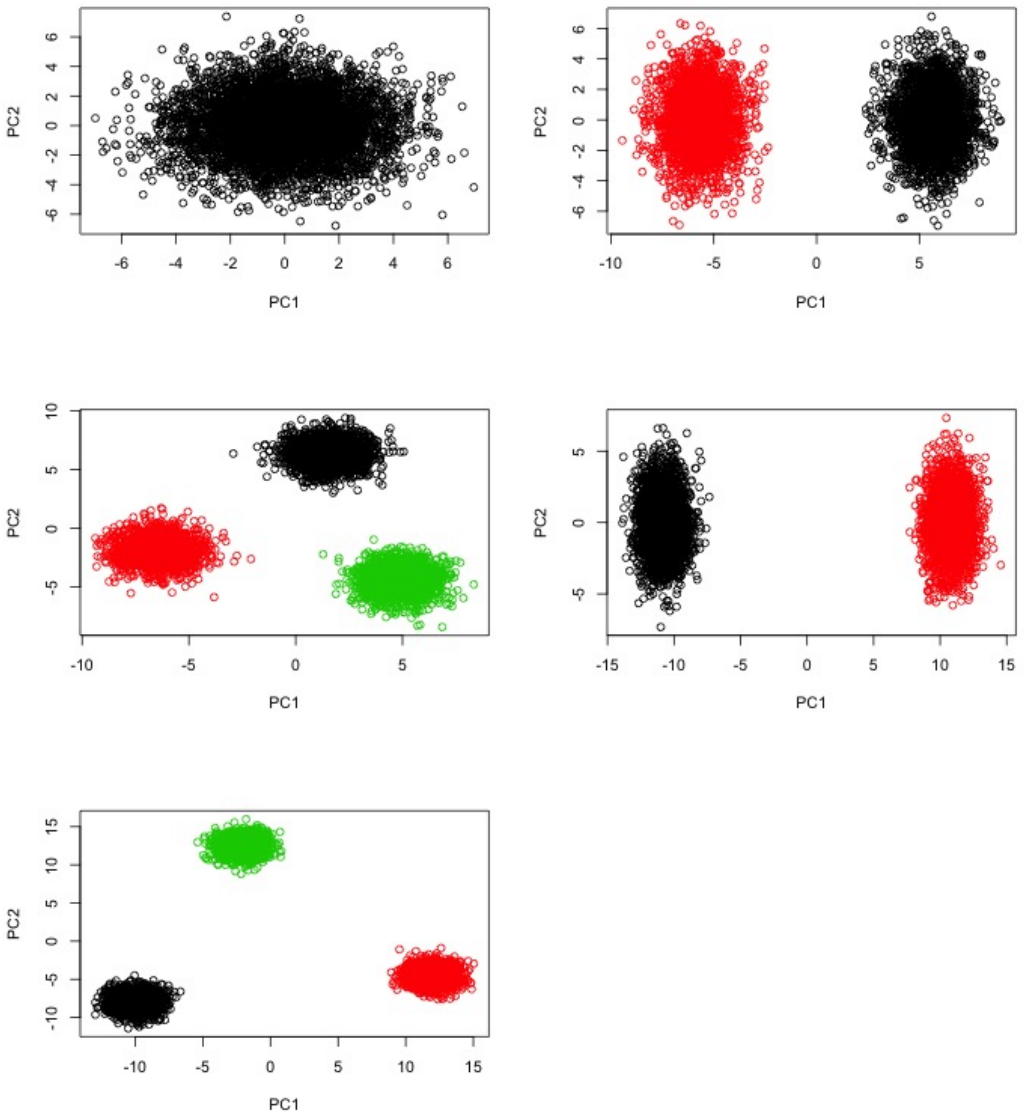


Figure A.4: ***Simulation scenarios with*** $N = 5000$ ***and*** $N_{drift} = 300$. *Starting from the top left panel are represented, in order, a single homogeneous population, two stratified populations with two less or more distant strata and two population with three less or more distant strata.*

Appendix B

R code

B.1 Source file

```
# Goodness-of-fit test to test given genotypes for fit to HWE

af2hwe<-function(q){
  c((1-q)^2,2*q*(1-q),q^2)
}

hwe.test = function(gts) {
  n = sum(gts)
  af = ((2 * gts[1] + gts[2])/(2 * n))
  hwf = af2hwe(1 - af)
  E = n * hwf
  t = sum((gts - E)^2/E)
  pchisq(t, 1, lower.tail = F)
}

hwe.test2 = function(gts) {
  n = sum(gts)
  af = ((2 * gts[1] + gts[2])/(2 * n))
  hwf = af2hwe(1 - af)
  E = n * hwf
  t = sum((gts - E)^2/E)
  list(pval=pchisq(t, 1, lower.tail = F),af=af)
}


# Simulation of genotypes

simGtsLoc <- function(rho, N){
  0:2%*%rmultinom(N,1,c(rho^2,2*rho*(1-rho),(1-rho)^2))
}
simGts <- function(rhos, N){
  sapply(rhos, function(rho) simGtsLoc(rho, N))
```

```
}


# Create different allele frequancies for stratified populations

newFreqs2 = function(rhos,N,M) sapply(rhos, function(rho){
  tem<-rho +rnorm(1, 0, M * rho * (1 - rho)/(2 * N));
  exp(tem-0.5)/(1+exp(tem-0.5))
  })


# Expit functions

expit = function(x,c=0) (1/(1 + exp(-(x-c))))
expit2 = function(x,min=0,max=1,c=0) ((max-min)*expit(x,c)+min)


## Likelihood single SNp

LikSingleSNP<- function(beta,G, PC, eta2=0){
  eta<-expit2(eta2,-1,1)
  #print(c(beta,eta))
  PC<-cbind(1,PC)
  rho <- PC %*% matrix(beta,ncol=1);
  #if(any(rho<0|rho>1)) return(-Inf)
  rho<-expit(rho)
  eta_max<-pmin(rho/(1-rho),(1-rho)/rho,
                (1-2*rho*(1-rho))/2*rho*(1-rho))
  eta<-eta*eta_max
  l <- ifelse(G==0,log(rho*(1-(1-rho)*(1+eta))),
              ifelse(G==1,log(2*rho*(1-rho)*(1+eta)),
                     log(1-rho-rho*(1-rho)*(1+eta)))
              )
  if(any(is.nan(l))) return(-Inf)
  else return(sum(l))
  #sum(l)
}

AuxLikSingleSNP <- function(par, G, PC){
  LikSingleSNP(beta=par[-1], G=G, PC=PC, eta2=par[1])
} # make it possible to use the optim function for LikSingleSNP
# when eta!=0


## Test deviation from HWE

HWETestLoc <- function(gen, pcs, ind, met){
  alf<-(sum(gen==1)+2*sum(gen==0))/(2*length(gen))
  Alt<-optim(c(0,alf,rep(0,length(ind))),AuxLikSingleSNP,
             G=gen,PC=pcs[,ind], method=met,
             control=list(fnscale=-1, maxit=1000)
             )
```

```
  Null <- optim(c(alf,rep(0,length(ind))),LikSingleSNP,
                G=gen,PC=pcs[,ind], method=met,
                control=list(fnscale=-1, maxit=1000)
                )
  LR <- -2*(Null$value-Alt$value)
  c('pval'=pchisq(LR,df=1,lower.tail=F),'eta'=Alt$par[1],
    'alt'=Alt$par[-1],'null'= Null$par)
}

HWETest <- function(Gts, pcs, ind, met){
  apply(Gts, MARGIN=2, HWETestLoc, pcs=pcs, ind=ind, met=met)
}


## To generate stratified populations

Strata <-function(N,rhos=runif(N,0.05,.5),M,Nind, Ndrift=300){
  l <- lapply(1:length(Nind),function(i){
    simGts(newFreqs2(rhos=rhos,N=Ndrift,M=M[i]),N=Nind[i])
    })
  L<-NULL
  for(i in 1:length(l)){
    L<-rbind(L,l[[i]])
  }
  L
}

# Auxiliary functions

Freq <- function(N,rhos=runif(N,0.05,.5),M,Nind, Ndrift=600){
  res <- sapply(1:length(Nind), function(i){
    newFreqs2(rhos=rhos,N=Ndrift,M=M[i])
    })
  v <- apply(res, 1, var)
  return(res[which.max(v),])
}

Table012 <- function(x){
  c('0'=sum(x==0),"1"=sum(x==1),"2"=sum(x==2))
}


## Likelihood ratio test function

TestFun <- function(G, p, met){

  PCs <- prcomp(x=G, center = T, scale. = T)$x
  Mgts <- t(apply(G,2,Table012))
  TL<-vector('list',length=2*p)
  TL[[1]]<-apply(Mgts,1,hwe.test)
  for(i in 1:p){
    TL[[i+1]] <- HWETest(Gts=G, pcs=PCs, ind=i, met=met)[1,]
```

```
    if((i+p+1)<2*p+1){
      TL[[i+p+1]] <- HWETest(Gts=G,pcs=PCs,ind=1:(i+1),met=met)[1,]
    }
  }
  TL
}

Testf<-function(G,p,met,PC){

  Mgts <- t(apply(G,2,Table012))
  TL<-vector('list',length=2*p)
  TL[[1]]<-apply(Mgts,1,hwe.test)
  for(i in 1:p){
    TL[[i+1]] <- HWETest(Gts=G, pcs=PC, ind=i, met=met)[1,]
    if((i+p+1)<2*p+1){
      TL[[i+p+1]] <- HWETest(Gts=G,pcs=PC,ind=1:(i+1),met=met)[1,]
    }
  }
  TL
}


## Simulation function

SimulationSingle2 <- function(N, M, Nind, p, met){
  GMat <- Strata(N=N, M=M, Nind=Nind)
  TestFun(G=GMat, p=p, met=met)
}

SigLevLoc <- function(alpha, x) mean(x<=alpha)
# function that retains a small number of alpha-lev per simulation
SigLev <- function(alphas, x){
  sapply(alphas, function(alpha) SigLevLoc(alpha, x))
}

TailStr <- function(x){
  # function that compute the Tail strength
  TS <- (1/length(x))*sum((1-sort(x)*((length(x)+1)/(1:length(x)))))
  TS
}

SimulationFull2 <- function(N, M, Nind, p, met, alphas, Niter){
  res <- Lapply(1:Niter, function(i,N,M,Nind,p,met,alphas,Niter){
        if (1) {
          temp <- SimulationSingle2(N=N,M=M,Nind=Nind,p=p,met=met)
    r<-rbind(sapply(temp,SigLev,alphas=alphas),sapply(temp,TailStr))
    rownames(r)<-c(as.character(alphas),'TS')
        }
         #r0 <- list(r = r, N = N, M = M, Nind = Nind, p = p, met = met);
  #r0 <- list(N = N, M = M, Nind = Nind, p = p, met = met);
        r0 = r;
  r0
```

```
}, N=N, M=M, Nind=Nind, p=p, met=met, alphas=alphas, Niter=Niter)
  res
}

SimulationFunction=function(Niteration,Nsnp,N,Npop,m,p,Ngen){

  SimulationFull2(
    N=Nsnp,
    M=rep(Ngen,Npop),
    Nind=round(vector.std(m^(1:Npop))*N),
    p=p,
    met='Nelder-Mead',
    alphas=c(0.05,0.01,1e-3,1e-4),
    Niter=Niteration
    )
}

runSimulation <- function(modelList,
                          output ='startification-simulations.RData'){
  result = iterateModels(modelList, SimulationFunction, parallel = T);
  if (!is.null(output)) save(result, file = output);
  result
}
```

B.2 Simulation script

```
#
#  simulation.R
#

source('SourceFile.R');
source('RgenericAll.R', chdir = T);
Source('https://git.lumc.nl/s.boehringer/configuration-shark/
␣␣␣␣␣␣␣␣raw/master/RcomputeResources.R')
library('parallelize.dynamic')

parallelize_setEnable(FALSE)
parallelize_declare(source = c('RgenericAll.R', 'SourceFile.R'));
parallelize.dynamic:::Log.setLevel(7);

if (1) {
  modelList2 = list(

    #global = list(list(Niteration = 1e2, m=1)),
    global = list(list(Niteration = 100, m=1)),
    Ngen=c(500,1000),
    Nsnp = 5e3,
    N = c(1, 5) * 1e3,
    Npop = c(1, 2, 3),
    p = c(2,3,4)
  );
```

```
}

# real simulation
if (1) {
  parallelize_initialize(Parallelize_config__,
                         backend = 'ogs-shark-all',
                         sourceFiles = c('RgenericAll.R','SourceFile.R'),
                         force_rerun = F, parallel_count = 200);
  rtest <- runSimulation(modelList2, output = NULL)
}
```

B.3 Data analysis

B.3.1 Data Quality Control

```
### Missing rates for individuals and markers

# The PLINK command:
# ./plink --bfile datamerged --missing
# creates..

missInd = read.table('/plink-1.07-mac-intel/plink.imiss',
                     header = T, stringsAsFactors = F);
missInd[missInd$F_MISS>0.02|is.na(missInd$F_MISS),2]
# extract individuals with proportion of missing genotypes >0.02
# OR individuals with NA/NaN value for missingness prop F_MIS

missMar = read.table('/plink-1.07-mac-intel/plink.lmiss',
                     header = T, stringsAsFactors = F);
length(missMar[missMar$F_MISS>0.03|is.na(missMar$F_MISS),2])
# n of markers with proportion of sample missing >0.03
# OR markers with NA value for missingness prop F_MISS


### Filter for Minor Allele frequencies

# ./plink --bfile datamerged --out dmerged --freq
# The command creates the file dmerged.frq

frq<-read.table('/plink-1.07-mac-intel/dmerged.frq', header = T,
                stringsAsFactors = F)

length(frq[frq$MAF<0.05|is.na(frq$MAF),2])
# n of markers with MAF<0.05 OR with NA value for the MAF
# 101739 markers with MAF<0.05


### Sex check

# The PLINK command:
# ./plink --bfile datamerged --out dmerged --maf 0.05 --check-sex
```

```
# creates the dmerged.sexcheck file

est_sex<-read.table('/plink-1.07-mac-intel/dmerged.sexcheck',
                    header = T, stringsAsFactors = F)
sum(est_sex$STATUS=='PROBLEM') # 22 individuals fails the qc


### Inbreeding coefficient estimates

# ./plink --bfile datamerged --out dmerged --het
# The command creates the file dmerged.het

het.est<-read.table('/plink-1.07-mac-intel/dmerged.het',
                    header = T, stringsAsFactors = F)
summary(het.est[,6])
#meanF<-as.vector(summary(het.est[,6])['Mean'])
#sdv<-sd(na.exclude(het.est[,6]))
#L<-meanF-qnorm(.975)*sdv
#U<-meanF+qnorm(.975)*sdv
#L;U (-0.01512911,0.01689671)


### New Lists of individuals and markers for the analysis

# list of individual to keep:
ind_to_keep<-est_sex[-which(est_sex$STATUS=='PROBLEM'),1:2]
# list of markers to keep:
Mar_to_keep<-frq[frq$MAF>=0.05 & !is.na(frq$MAF),2]

# Create txt files for plink
write(t(ind_to_keep),ncolumns=2,file='individuals.txt')
write(Mar_to_keep,file='markers.txt')

# PLINK command:
# ./plink --bfile datamerged --out dmclean --keep /individuals.txt
# --extract /markers.txt --make-bed
# creates a new bfile with only desired markers and individuals:
# dmclean.bed, dmclean.bim, dmclean.fam
```

B.3.2 Principal Components Analysis and Likelihood Ratio Test for HWE

```
### Pruned subset of snps approximately in LE (plink)

# The PLINK command
# ./plink --bfile dmclean --indep 100 5 1.5 --out dmpruned
# creates files dmpruned.prune.in and dmpruned.prune.out

causal_snps<-read.table('/plink-1.07-mac-intel/dmpruned.prune.in',
                        header = F,
                        stringsAsFactors = F)
```

```
# The file dmpruned.prune.in is a list of SNPs in approximate LE

# The PLINK command
# ./plink --bfile dmclean --out dmprune --extract dmpruned.prune.in
# --make-bed
# creates a new bfile with only the desired markers and individuals:
# dmprune.bed, dmprune.bim, dmprune.fam

# Reading plink files and get a genotypes matrix
data<-read.plink('/plink-1.07-mac-intel/dmprune.bed',
                 '/plink-1.07-mac-intel/dmprune.bim',
                 '/plink-1.07-mac-intel/dmprune.fam')
dataMat<-data$genotypes
write.SnpMatrix(dataMat,file='dataMat2')
Data<-read.table('dataMat2')

### Preparatory step 2: PCA

# Selection of 4 different random subset of the pruned set of snps.
# This is to check if different subsets of pruned snps gave the same
# picture of the expected population structure since it was unfeasible
# to perform the PCA on the entire pruned set

snplist<-causal_snps$V1

# Plot

par(mfrow=c(2,2))

for(i in 1:4){
sam<-sample(snplist,size=10000)
sam<-gsub('-','.',sam)
dat<-Data[,sam]
PCA<-prcomp(dat, center = T, scale. = T)
plot(PCA$x)
}

par(mfrow=c(1,1))

# Check on the matrix of the 71872 SNPs approximatly in LE
Tab<-apply(Data, 2,Table012)
Tab<-t(Tab)
Hwes<-apply(Tab2,1,hwe.test)
Hwes_sort<-sort(Hwes2)

# Select from the 71872 snps the ones with pval>1e-10
snplist<-names(Hwes[which(Hwes>1e-10)])
rsub<-sample(snplist,10000)

# PCA based on rsub
PCA<-prcomp(Data[,rsub], center = T, scale. = T)
plot(PCA$x) # plot
```

```
title('Population␣substructure␣captured␣by␣a␣subsample␣of␣10000
␣␣␣␣␣␣markers')

### Analysis on 2 different subsets of snps in rsub3

## Extract subsample of snps to test with smallest pval in rsub3
stt<-sort(Hwes[rsub])
max(sort(Hwes[rsub])[1:1000])
stt<-names(stt[1:1000])

## Extract random subsample of snps to test in rsub3
set.seed(2)
stt2<-names(sample(Hwes[rsub], size=1000))

## Test on the two selected subsets of snps
for(i in 1:2){
  if(i==1) res<-Testf(G=Data[,stt],p=4,met='Nelder-Mead',PC=PCA$x)
  else resI<-Testf(G=Data[,stt2],p=4,met='Nelder-Mead',PC=PCA$x)
}
```

Bibliography

[1] Jonathan K. Pritchard and Noah A. Rosenberg. *Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies.* American Journal of Human Genetics, 65: 220-228, 1999.

[2] Janis E. Wigginton, David J. Cutler and Gonçalo R. Abecasis. *A Note on Exact Tests of Hardy-Weinberg Equilibrium.* American Journal of Human Genetics, 76: 887-893, 2005.

[3] Jonathan Taylor and Robert Tibshirani. *A tail strength measure for assessing the overall univariate significance in a dataset.* Biostatistics, 7(2): 167-181, 2006.

[4] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E Weinblatt, Nancy A Shadick and David Reich. *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 38(8): 904-909, 2006.

[5] Nick Patterson, Alkes L. Price, David Reich. *Population Structure and Eigenanalysis.* PLoS Genetics, `www.plosgenetics.org`, 2(12): 2074-2093, 2006.

[6] The Wellcome Trust Case Control Consortium *Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls.* Nature, 447: 661-678, 2007.

[7] Simon C. Heath, Ivo G. Gut, Paul Brennan, James D. McKay, Vladimir Bencko, Eleonora Fabianova, Lenka Foretova, Michel Georges, Vladimir Janout, Michael Kabesch, Hans E. Krokan, Maiken B. Elvestad, Jolanta Lissowska, Dana Mates, Peter Rudnai, Frank Skorpen, Stefan Schreiber, José M. Soria, Ann-Christine Syvänen, Pierre Meneton, Serge Herçberg, Pilar Galan, Neonilia Szeszenia-Dabrowska, David Zaridze, Emmanuel Génin, Lon R. Cardon and Mark Lathrop. *Population Structure and Eigenanalysis.* European Journal of Human Genetics, 16: 1413-1429, 2008.

[8] Gil McVean. *A Genealogical Interpretation of Principal Components Analysis.* PLoS Genetics, `www.plosgenetics.org`, 5(10): 1-10, 2009.

[9] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zhao, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods and J. Zhang. *A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data*. The Pharmacogenomics Journal, 10: 278-291, 2010.

[10] N. M. Laird, C. Lange. *The fundamentals of modern statistical genetics*. Springer, 2010.

[11] Matthieu Bouaziz, Christophe Ambroise, Mickael Guedj. *Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies*. PLoS Genetics, `www.plosgenetics.org`, 6(12): 1-13, 2011.

[12] Johann A. Gagnon-Bartsch, Terence P. Speed. *Using control genes to correct for unwanted variation in microarray data*. Biostatistics, 13(3): 539-552, 2012.

[13] Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C. Royrvik, Barry Cunliffe, Wellcome Trust Case Control Consortium, International Multiple Sclerosis Genetics Consortium, Daniel J. Lawson, Daniel Falush, Colin Freeman, Matti Pirinen, Simon Myers, Mark Robinson, Peter Donnelly and Walter Bodmer. *The fine-scale genetic structure of the British population*. Nature, 519: 309-314, 2015.