UNIVERSITY OF COPENHAGEN
DEPARTMENT OF MATHEMATICAL SCIENCES

Mathematical tools for population genetics based on genotype data

A PhD thesis by Song Li

January 2023

THIS THESIS HAS BEEN SUBMITTED TO THE PHD SCHOOL OF
THE FACULTY OF SCIENCE, UNIVERSITY OF COPENHAGEN

SONG LI
song.li@math.ku.dk

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
2100 COPENHAGEN, DENMARK

Supervisor: Carsten Wiuf
Department of Mathematical Sciences
University of Copenhagen, Denmark

Assessment committee: Bo Markussen (Chair)
Department of Mathematical Sciences
University of Copenhagen, Denmark

Lars Nørvang Andersen
Department of Mathematics
Aarhus University, Denmark

Jotun Hein
Department of Statistics
University of Oxford, UK

Date of submission: January 15, 2023

Date of defence: March 20, 2023

ISBN: 978-87-7125-068-8

Mathematical tools for population genetics based on genotype data

Song Li

A thesis presented for the degree of
Doctor of Philosophy

January 2023

Abstract

This thesis covers work in mathematical and statistical aspects of population genetics.

In the first part of the thesis, we present the behaviour of two F-statistics, which are defined as the difference in the allele frequency at a given time point in one population and the difference in allele frequency between two populations, respectively. In order to calculate the first two moments of the allele frequency present in the mathematical expression of the F-statistics, mutation and migration as linear evolutionary forces are incorporated into Wright-Fisher model. We give some parameter conditions that cause the behaviour of the F-statistics to be non-monotonic over time, that is, to increase and then decrease over time.

In the second part of the thesis, we propose a new method to evaluate the statistical fit of principal component analysis (PCA) in admixture models and population structure inference. Statistical tools such as residual and correlation coefficients are utilized to detect violations of model assumptions. We give the mathematical expressions of two correlation coefficient matrices of the residuals and some theorems about their properties. The method is demonstrated in both simulated and real data through matrix visualization.

Finally, I introduce a mathematical definition and estimate of kinship coefficient based on pedigree and genotype data. I propose a procedure for dividing the sample containing related individuals who have alleles copied from a common ancestor into some data sets. Each data set corresponds to the individual under study and is used in PCA method to estimate the allele frequency of the studied individual present in the kinship coefficient. In the presence of population structure, the proposed estimate performs well for the case of full-siblings.

Resumé

Denne afhandling dækker arbejde i matematiske og statistiske aspekter af populationsgenetik.

I specialets første del præsenterer vi adfærden for to F-statistikker, der er defineret som forskellen i allelfrekvensen på et givet tidspunkt i henholdsvis én population og forskellen i allelfrekvensen mellem to populationer. For at beregne de første to momenter af allelfrekvensen, der er til stede i det matematiske udtryk for F-statistikken, er mutation og migration som lineære evolutionære kræfter inkorporeret i Wright-Fisher-modellen. Vi giver nogle parameterbetingelser, der får F-statistikkens opførsel til at være ikke-monotonisk over tid, det vil sige, at den øges og derefter falder over tid.

I anden del af afhandlingen foreslår vi en ny metode til at evaluere den statistiske tilpasning af principal komponentanalyse (PCA) i blandingsmodeller og befolkningsstrukturinferens. Statistiske værktøjer såsom residualer og korrelationskoefficienter bruges til at opdage brud på modelantagelser. Vi giver de matematiske udtryk for to korrelationskoefficientmatricer af residualerne og nogle sætninger om deres egenskaber. Metoden demonstreres i både simulerede og reelle data gennem matrixvisualisering.

Til sidst introducerer jeg en matematisk definition og estimat af slægtskabskoefficient baseret på stamtavle og genotypedata. Jeg foreslår en procedure til at opdele prøven, der indeholder relaterede individer, som har alleler kopieret fra en fælles forfader i nogle datasæt. Hvert datasæt svarer til individet under undersøgelse og bruges i PCA-metoden til at estimere allelfrekvensen for det undersøgte individ, der er til stede i slægtskabskoefficienten. I nærværelse af befolkningsstruktur fungerer det foreslåede skøn godt for helsøskende.

Contents

Chapter 1

Introduction

The main content of this thesis covers the three manuscripts. The first manuscript that makes up chapter 2 is available on *bioRxiv*, while the second and third manuscripts are not public yet.

The first manuscript focuses on the application of F-statistics to population-specific allele frequencies over time within and between populations. The Wright-Fisher model with mutation and migration provides the first two moments expressions of allele frequency, allowing F-statistics to be analyzed for the behaviour.

The second manuscript develops a method for evaluating inferred population structure using the properties of the residuals, that is, the difference between the predicted and the observed genotypes under the admixture model. We consider principal component analysis or clustering algorithm to obtain the residuals. The visualization of residual correlation matrix based on simulated and real data can be utilized to investigate cases that violate the assumptions of the model.

In the preliminary work of the third manuscript, the kinship coefficient is introduced to address the issue of related individuals in structured populations and a formula of estimating kinship coefficient based on genotype data is proposed. I consider the method of principal component analysis to estimate individual-specific allele frequency parameters contained in the formula. Using simulated scenarios with some population structure contexts, the implementation of the estimation method in full-siblings case performs better than some existing methods, which can be motivated to expand to other cases in the future.

The introduction is meant to provide the necessary background of the topics covered, thereby deepening the familiarity with concepts and terms that appear in manuscripts. In section 1.2 of this chapter, I present the main contributions of the work done in my manuscripts, as well as perspectives for the future research directions noted. The bibliography part is the list of references cited in the introduction, and the references of each manuscript are in their own chapters.

1.1 Background

When it comes to tracing the ancestral history of a specific allele, one can consider the relationship between individuals and populations. Set some assumptions on the relationship, follow some laws of heredity, simulate the evolution process and extract some information we are interested in, so as to explore the genetic variation. As depicted in Figure 1.1, the source is defined as a population

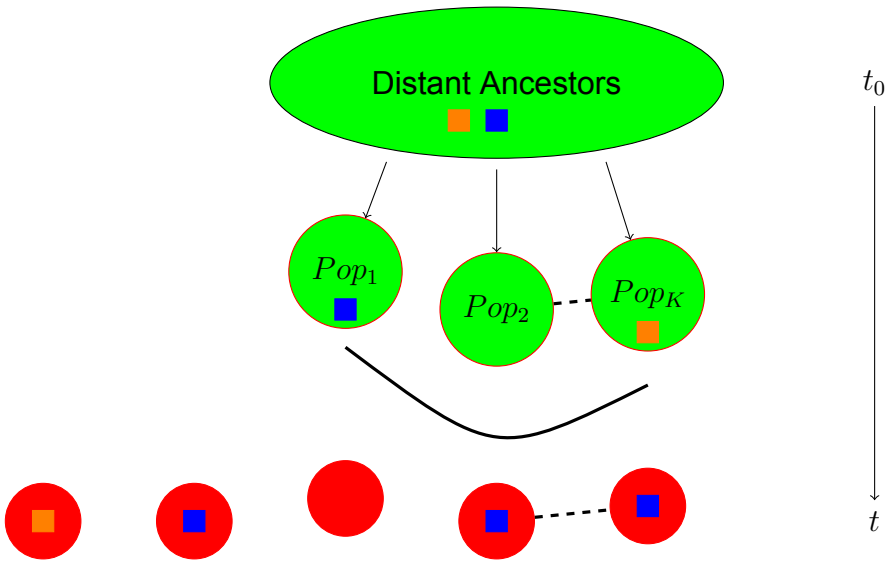


Figure 1.1: Illustration of individuals in generation t tracing their ancestry to generation t_0. The K populations share the common distant ancestors and give rise to many current individuals. A particular allele in an individual can be traced back to a population and further down to a distant ancestor. Green circles represent populations, red circles represent individuals and rectangles represent alleles.

of distant ancestors. As a result of evolutionary forces and mating modes, the descendants of ancestors continue to diverge into some distinct populations. Mating within and between populations leads to different individuals. This process is accompanied by the transmission of genetic information. Extract such genetic information as data, which is used for analyzing many subjects including population structure and individual genetic differences. Defining variables according to different research priorities is the first step in using data efficiently, followed by the establishment of theoretical models to describe these variables based on appropriate assumptions. The specific statistics selected under the model can be used as indicators to explore the predicted trend of changes in the corresponding variables. The issue of estimating the parameters involved in the model and evaluating model fit leaves a lot of room for methodology. Meanwhile, the correlation reflected in the data reveals the understanding of the relationship between individuals to which data belongs. For the above mentioned I expand in the following sections.

1.1.1 Variation and data

A DNA or RNA molecule is made of deoxynucleotides. Deoxynucleotides carry the important functional group called the base and there are four kinds of bases in DNA: Adenine (A), Cytosine (C), Guamine (G) and Thymine (T). A series of polynucleotides are joined together according to the rules of base pairing and are arranged in sequence to form the helical structure of DNA. Proteins are added as protective shells for DNA sequences, giving rise to the chromosomes of eukaryotes, such as mammals. So cells appear, and then algae, multicellularity, plants, humans. Humans are diploid creatures, carrying two copies of 22 autosomes and a pair of sex chromosomes. The fact that individuals have such a set of chromosomes means that they share the genetic information contained in the majority of DNA sequence. But the process by which chromosomes divide to make copies is full of randomness. The variation that follows randomness is the evidence that a population of individuals has evolved over time. The most common type of human genetic variation, single nucleotide polymorphism (SNP), represents a difference at a single nucleotide, such as base A being replaced by base G.

On average, SNPs occur once per 1,000 nucleotides. For example, a comparison of two parts of the genome at the same location reveals a nucleotide substitution, then a SNP and two alleles are said to be present. The two alleles are named the major and the minor allele based on the frequency of occurrence in a population at that locus. In population genetics, the frequency of an allele occurring at a locus in a population is a fraction or percentage to describe the amount of variation, defined as allele frequency (Gillespie, 2004). If the major allele is marked "B" and the minor allele is marked "b" at a locus, the term genotype is used to describe three different combinations: Bb, BB and bb. An individual is homozygous at a locus if the genotype is BB or bb, while is heterozygous at a locus if the genotype is Bb. The strand of DNA responsible for carrying translatable information along the 5′-to-3′ direction is called the coding strand, where 5′ and 3′ are positions defined by chemical concepts. A pair of homologous chromosomes has two coding strands, each of which is a sequence of different bases. At a given locus, the combination of bases taken from the two coding strands also represents the genotype. For example, in Figure 1.2, a diploid individual has two copies of each chromosome consists of DNA, which is composed of the four bases A, T, C and G, then the genotype of SNP 1 is AT. Homozygotes and heterozygotes are distinguished by genotype, for example, TT, CC and GG indicate that the individual is homozygous for SNP 4,6,7,9 and 10, while AT, TG, GC and TC indicate that the individual is heterozygous for SNP 1,2,3,5 and 8. For a population of individuals, most sites have only two types of bases, but some sites could have three or all four. The pair could be any of six possible AG, AC, AT, GC, TC and TG. Millions of SNPs have been reported in humans. By setting one of the two simplest alleles at a locus as the reference allele and the other as the alternate allele, the number of the reference allele is encoded as the genotype value (see Table 1.1). If the genotype value is 0 or 2 then the individual is said to be homozygous at that locus, otherwise the individual

SNP	1	2	3	4	5	6	7	8	9	10	⋯
$5'\cdots$	A	T	G	T	A	T	C	T	T	G	$\cdots 3'$
$5'\cdots$	T	G	C	T	T	T	C	C	T	G	$\cdots 3'$

Figure 1.2: Two coding strands of the homologous chromosomes.

is heterozygous. The genotype value is an important entry point for linking data from variables. In Table 1.1, the reference and alternate allele are randomly assigned. Note that the reference allele is in practice determined by the reference genome, which is not the same as the major allele depending on the frequency of occurrence in a particular population.

Table 1.1: The genotype value of two alleles combined in Figure 1.2

SNP	ref	alt	value
1	A	T	1
2	T	G	1
3	G	C	1
4	T	C	2
5	T	A	1
6	T	C	2
7	A	C	0
8	T	C	1
9	T	G	2
10	C	G	0

In addition to single base substitutions, alleles arise from insertions and deletions of multiple base pairs, which are classified as DNA mutations. Mutations, which occur during replication or recombination or both, change the sequence of bases that represent a segment of DNA. The base sequence shown in Figure 1.3 represents an allele that is altered by each of the three possible mutations: substitution is replacing base C with base A, insertion is adding bases G and A to the sequence, and deletion is removing bases G, C, T from the sequence.

Allele:	A	C	T	A	G	C	T	A	G	C	T
Substitution:	A	C	T	A	G	A	T	A	G	C	T
Insertion:	A	C	T	A	G	G	C	A	T	A	G	C	T
Deletion:	A	C	T	A					A	G	C	T

Figure 1.3: Three types of DNA mutations.

1.1.2 Model and theory

The Wright Fisher model

The simplest tool to describe the variation in the number of individuals carrying a reference allele in a population of discrete, nonoverlapping generations and random mating is the Wright–Fisher model (Fisher, 1930; Wright, 1931). Let Z_t represent the number of haploid individuals with a reference allele "A" in a population of constant size N at generation t and denote allele frequency $X_t = Z_t/N$. The Wright–Fisher model can be defined as follows in the form of a binomial distribution

$$Z_{t+1}|X_t = x_t \sim \mathcal{B}in(N, x_t). \tag{1.1}$$

The Wright-Fisher model focuses on the effects of genetic drift, assuming that other evolutionary forces such as selection, mutation, and migration are not taken into account. Natural selection plays an important role in population genetics, but the null model ignores the existence of natural selection in the first place. The null model is based on the neutral theory proposed by Motoo Kimura in 1968. The Wright-Fisher model can be considered neutral in the sense that each offspring individual randomly selects a parent from the previous generation. Under the framework of the Wright-Fisher model, the first two moments that are computed using the fact that the expectation and variance of a binomial distribution are

$$\begin{aligned} \mathbb{E}\left[X_{t+1} \mid X_t = x_t\right] &= x_t, \\ \mathbb{V}\left[X_{t+1} \mid X_t = x_t\right] &= \frac{x_t(1-x_t)}{N}, \end{aligned} \tag{1.2}$$

where $\mathbb{E}$ and $\mathbb{V}$ are the expectation and variance operators, respectively. By the law of iterated expectation, it can be seen that the expected frequency of each generation is the same and equal to the given allele frequency of the first generation. Z_t and Z_{t+1} are rewritten as i and j to denote the number of individuals carrying the allele in generation t and $t+1$, respectively. The probability of the allele being transferred from i to j is given

$$P(Z_{t+1} = j \mid Z_t = i) := P_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}, \quad 0 \le i, j \le N. \tag{1.3}$$

From the above definition, the Wright–Fisher model is a discrete time Markov process in which the state space of allele frequency is $x_t = i/N$. The absorbing states of this Markov chain are the state 0 and N, which allows one to observe whether the final population loses or fixes an allele (see Figure 1.4). Let a_i be the probability that the final population contains only one allele type, and the number of individuals carrying this allele at the starting is i. The calculation of a_i is divided

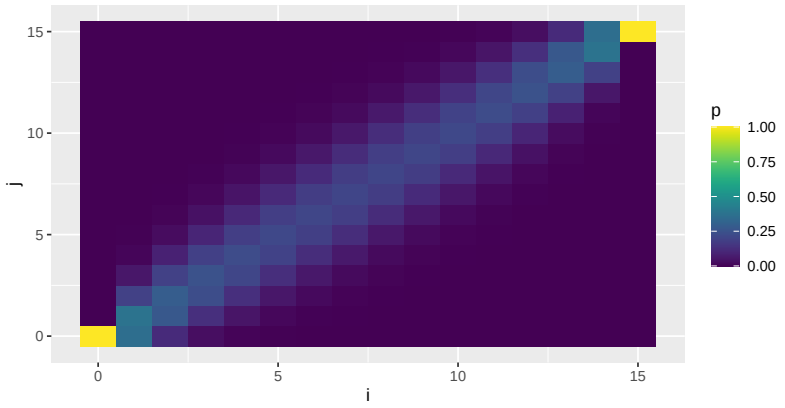


Figure 1.4: Shown is the probability transition matrix assuming $N = 15$.

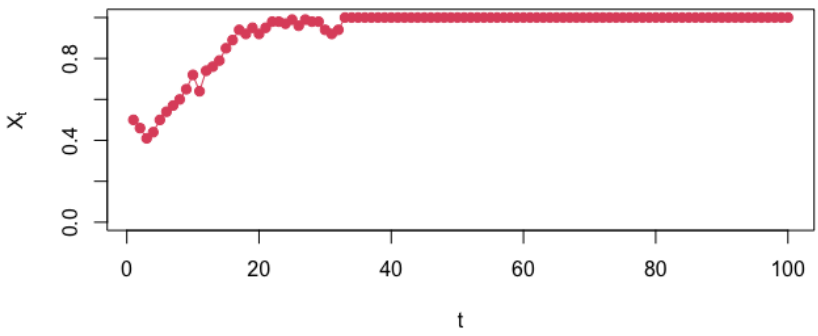
into two parts, one is the probability of the transition from the initial state $Z_0 = i$ to the state $Z_1 = j$ at the next time, and the other is the probability of reaching the absorbing state N from the state j. The combination of the two probabilities is the following

$$a_i = \sum_{j=0}^{N} P_{ij} a_j, \quad i \in \{1, 2, \cdots, N\}. \tag{1.4}$$

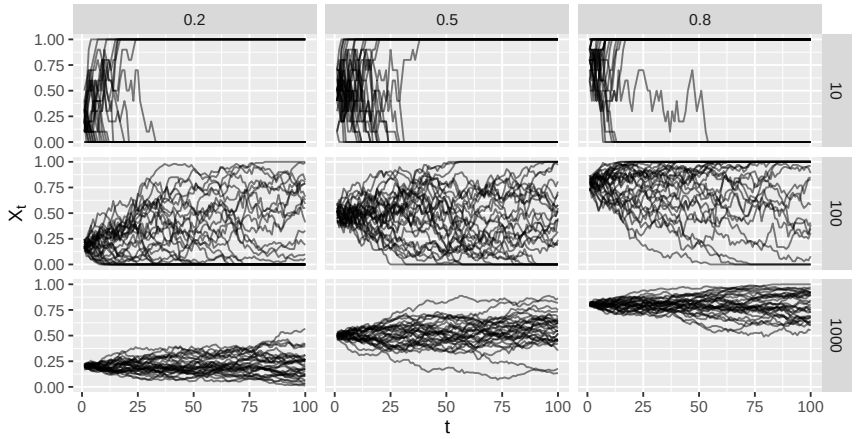
With the transition probability (1.3), $\mathbb{E}[Z_1 \mid Z_0 = i] = i$, i.e. $\sum_{j=0}^{N} P_{ij} j = i$. Then $a_i = i/N$, that means the probability that the allele will eventually be fixed is just its initial frequency (Tavaré, 2004). Figure 1.5(a) shows an example of individuals carrying only one allele in a final population using the Wright–Fisher model and Figure 1.5(b) illustrates the claim about a_i. As can be observed from Figure 1.5(b), the allele frequency trend of large population is often accompanied by the lengthening of corresponding time scale. When the population size N is set large enough, the allele frequency X_t converges to a specific constant, but with increasing time. Such limit processes depend on population size and time. Unifying measures of time and population size on a scale is a common way to obtain limits in population genetics. Note that time as a discrete variable is the number of generations. Scale generations in units of N, i.e. $t = \lfloor Nu \rfloor$ and the allele frequency becomes

$$Y_N(u) = N^{-1} Z_{\lfloor Nu \rfloor}, \quad u \geq 0, \tag{1.5}$$

where $\lfloor \cdot \rfloor$ is the rounding operation. As $N \to \infty$, $Y_N(u) \xrightarrow{d} Y_u$, where Y_u is a diffusion process in $[0, 1]$. Diffusion theory (Crow and Kimura, 1970; Karlin and Taylor, 1980; Neuhauser, 2001) provides some settings for Y_u, including the definition of the first two moments. Based on the moments, further explorations of Y_u are in the first manuscript. Under random mating, all alleles are inherited independently, then a population consisting of N diploid individuals can be represented by a haploid model of size $2N$ when applying the Wright-Fisher model. The Wright-Fisher model can be extended to non-constant population sizes. In math, one can denote that the population



(a)



(b)

Figure 1.5: (a) Shown is the allele frequency in the Wright-Fisher model varies with generation under $N = 100$ and $x_1 = 0.5$. (b) The initial allele frequencies are 0.2, 0.5 and 0.8, and population sizes are 10, 100 and 1000. The number of replicates and generations per simulation is 30 and 100, respectively. As the population size increases, the noise decreases. Furthermore, when the initial frequency is smaller, it corresponds to loss, and when the initial frequency is large, fixation is more likely.

sizes from generation 0 to t are $N_0, \ldots, N_t$, respectively. Compared with the constant population size in the ideal model, the population size in the non-ideal model that fluctuates by generations is close to the reality but difficult to use for some calculations. However, if the similarities between the two models are captured, and the latter can be described by the former through transformation. In this way, the non-ideal model (non-constant population sizes) can be converted into a new ideal model (constant population size) for analysis. For example, in the Wright-Fisher model, the final state of one allele is the absorbing state 0 or N, mean that the heterozygosity in the population is lost. The loss rate of heterozygosity, if it is the same in both models, then the constant population size defined in the ideal Wright-Fisher model is called the effective population size and denoted as N_e (Wright, 1931). Let b_0 be the probability that any two alleles in generation 0 are different, similarly $b_1, \ldots, b_t$. The probability that two alleles in generation 1 come from different parents

in generation 0 is $1 - 1/N_0$, similarly $1 - 1/N_1, \ldots, 1 - 1/N_{t-1}$. Thus,

$$b_t = (1 - \frac{1}{N_0}) \cdots (1 - \frac{1}{N_{t-1}}) b_0.$$

From the above equation, it can be seen that heterozygosity decreases gradually. For the constant population size N,

$$b_t = (1 - \frac{1}{N})^t b_0.$$

By the definition, the effective size for loss of heterozygosity is

$$N_e = \frac{t}{\frac{1}{N_0} + \ldots + \frac{1}{N_{t-1}}}.$$

The harmonic average of a set of numbers is often closer to the smallest of the numbers. This makes the above effective population size value closer to the smallest size of the population, such as a bottleneck. A population bottleneck is a drastic reduction in population size. Effective population size is a widely used concept and can be similarly defined for variance, inbreeding and coalescence versions.

Binomial distribution and Markov chain mathematize the effect of genetic drift on allele frequencies to form the simplest Wright-Fisher model. Mutation, migration, and natural selection are also expected to cause changes in allele frequency, depending on the form of mathematical expression. Mutation and migration, as linear evolutionary forces like genetic drift, can be considered in a general and unified manner, while selection can be considered as nonlinear forms (Crow and Kimura, 1970). Different forms need multiple versions of the Wright-Fisher model, which is discussed in the first manuscript.

F-statistics

The level of heterozygosity in a population is reflected by the change of allele frequency. When the frequency value of a specific allele at a certain locus in a population rises or falls significantly, it means that the heterozygosity decreases sharply. A useful mathematical tool used to describe the expected value of the difference in frequency between a population at different times, or between different populations at any times, is the F-statistics. The term F-statistics, proposed by Reich et al. (2009), are considered to measure the genetic difference between sets of populations when only genetic drift occurs. In the F-statistics theory, there are three statistics labeled F_2, F_3 and F_4. Since the latter two can be represented by the F_2, it is fundamental to consider the definition of the F_2. For two different populations P_1 and P_2, assume that the reference allele has frequency $X_{t_1,1}$ in P_1 with population size $N_{t_1,1}$ at time t_1, and $X_{t_2,2}$ in P_2 with population size $N_{t_2,2}$ at time

t_2, respectively. For P_1 at time t_1 and t_1', F_2 is defined as

$$F_2\left(P_{t_1,1}, P_{t_1',1}\right) = \mathbb{E}\left[\left(X_{t_1,1} - X_{t_1',1}\right)^2\right]. \tag{1.6}$$

For P_1 and P_2 at any given times, F_2 is defined as

$$F_2\left(P_{t_1,1}, P_{t_2,2}\right) = \mathbb{E}\left[\left(X_{t_1,1} - X_{t_2,2}\right)^2\right]. \tag{1.7}$$

The above F_2 directly shows the variation of allele frequency influenced by genetic drift. Applications based on F_2 are rapidly and widely used in population genetic studies, e.g., the testing for tree-like structure and admixture history of related populations (Reich et al., 2009), the number of founding populations for the European region (Lazaridis et al., 2014) and estimates of admixture proportions (Patterson et al., 2012).

Equation (1.7) quantifies the genetic differences between two populations, and if such differences are affected by genetic structure, the Wright's version of the F-statistics is another commonly used mathematical tool called F_{st} or fixation index (Wright, 1951). Unfortunately, there is no fixed definition of F_{st}. Wright originally defined F_{st} as a ratio of variances, while a more common definition is based on probabilities (Durrett, 2008). For example, the F_{st} of P_1 and P_2 can be defined as

$$F_{st}\left(P_1, P_2\right) = \frac{p - q}{1 - q}, \tag{1.8}$$

where p represents the probability that two alleles in a population are the same and q represents the probability that two alleles from different populations are the same. The interpretation of F_{st} is that a higher value implies greater divergence between populations, while $F_{st} = 0$ indicates that populations are randomly mating. It is worth noting that the statistics here are parameters which are part of the model.

1.1.3 Two methods for inferring population structure

Population structure

The variation in population-specific allele frequencies focuses on quantifying the effects of evolutionary forces on populations, namely splitting into some populations from distant ancestors (see Figure 1.1). In population and individual parts, each individual can trace the line of descent from the corresponding population. Specifically, alleles in the gene pool of the ancestral population have a probability of being passed on to the offspring individuals. The nonrandom mating of individuals divided into different populations causes differences in allele frequencies between populations. Population structure is the concept proposed to describe such differences. Geographic

reproductive isolation is a common factor in shaping population structure, such as the different ethnic groups produced by the five continents of the Earth, which often become a known set of ancestral populations. Other factors, such as migration and interpopulation mixing, contribute to the proportion of individuals that have inherited alleles from each ancestral population, which is the evidence of population structure. An example of SNPs on the homologous chromosomes of an individual are shown in Figure 1.6, and the proportion of the corresponding ancestral population occupying this segment is defined the admixture proportion of the individual.

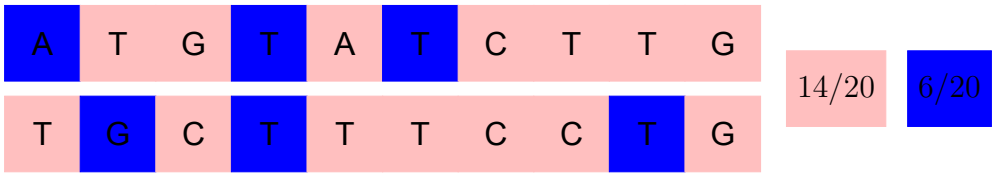


Figure 1.6: The SNPs array on some segments of the homologous chromosomes of an individual produced by two ancestral populations with admixture proportions $14/20$ and $6/20$. The pink and blue segments are inherited from population 1 and 2, respectively.

Clustering

Clustering and principal component analysis (PCA) are methods widely used in the inference of population structure, which focus on admixed individuals with different ancestry backgrounds. In terms of genetic clustering algorithms, for example, Pritchard et al. (2000) developed the software STRUCTURE based on Markov chain Monte Carlo theory, and Alexander et al. (2009) created the software ADMIXTURE based on maximum likelihood. Compared with the former, the latter improves the computing speed. The mathematical definition of the likelihood is as follows. Given K ancestral populations, M biallelic SNPs and N individuals, then g_i^s is denoted as the genotype value of individual i at SNP s. The individual-specific allele frequency π_i^s is defined as follows (Thornton et al., 2012),

$$\pi_i^s = \sum_{k=1}^{K} f_k^s q_{ki}, \tag{1.9}$$

where $q_{ki}, f_k^s \in [0,1]$ are the admixture proportion of individual i and the the reference allele frequency of the ancestral population k at SNP s, respectively, and $\sum_{k=1}^{K} q_{ki} = 1$. In admixture models, (1.9) is written as matrix notation $\mathbf{\Pi} = \boldsymbol{FQ}$, where $\mathbf{\Pi} = (\pi_i^s)_{M\times N}, \boldsymbol{F} = (f_k^s)_{M\times K}, \boldsymbol{Q} = (q_{ki})_{K\times N}$. In ADMIXTURE, the probabilities are proportional to a specific form, i.e.,

$$P(g_i^s \mid \pi_i^s) \propto (\pi_i^s)^{g_i^s} (1-\pi_i^s)^{2-g_i^s}, \quad g_i^s = 0, 1, 2.$$

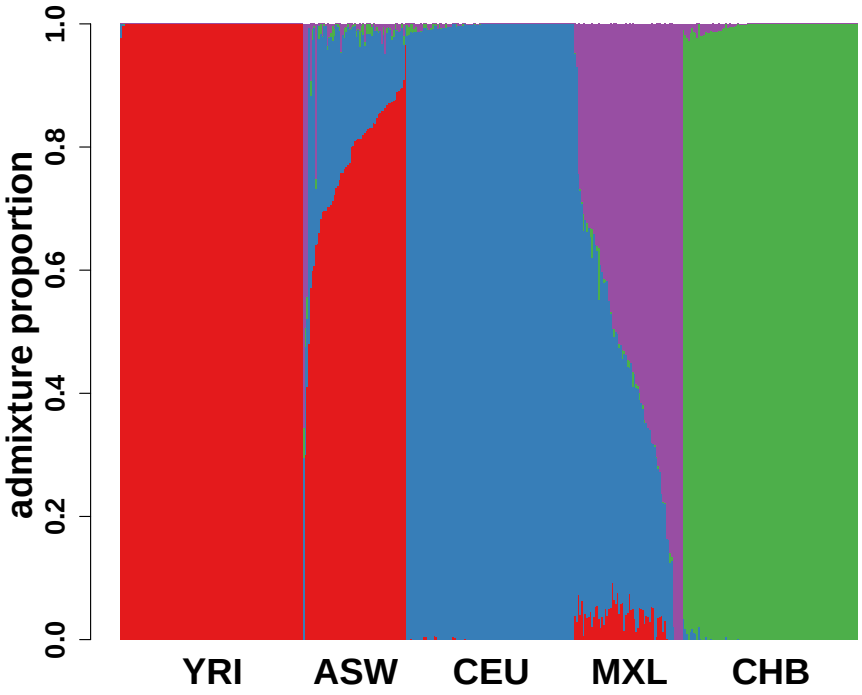


Figure 1.7: Shown is an example of admixture proportions obtained by applying real human data from the 1000 Genomes project to the software ADMIXTURE and setting $K = 4$ (Auton et al., 2015; Garcia-Erill and Albrechtsen, 2020). The data includes 434 samples of five groups, which are 108 samples of the Yoruba group from Nigeria (YRI), 61 samples of the group from Southwest US with African ancestry (ASW), 99 samples of the Utah group with Northern and Western European ancestry (CEU), 63 samples of the group from Los Angeles, California with Mexican ancestry (MXL) and 103 samples of the Han Chinese group from Beijing, China (CHB).

That produces the logarithmic likelihood

$$L = \sum_{i=1}^{N} \sum_{s=1}^{M} \left\{ g_i^s \ln \left[\sum_{k=1}^{K} f_k^s q_{ki} \right] + (2 - g_i^s) \ln \left[\sum_{k=1}^{K} (1 - f_k^s)\, q_{ki} \right] \right\}.$$

The above likelihood holds under the condition that individuals are independent of each other and linkage disequilibrium is ignored. Linkage disequilibrium (LD) is the nonrandom association of alleles of different locus. Recombination interacts in a complex way with selection, mutation and genetic drift to determine levels of LD (Slatkin, 2008). Ignoring LD means that any two alleles are said to be statistically independent. Some iterative optimization algorithms are used to estimate the parameter matrices $\boldsymbol{Q}$ and $\boldsymbol{F}$ in the process of likelihood L maximization. The admixture proportions estimated can be shown as a bar chart (see Figure 1.7).

PCA

PCA is introduced by Menozzi et al. (1978) into population genetics to condense allelic information in Europeans. Then PCA is applied to genetic data to study population structure on a solid statistical basis (Patterson et al., 2006). PCA is an optimization method that projects the high-dimensional data matrix to the first few uncorrelated principal components formed by the linear combination of the considered variables, so as to capture the variance in the data as much as possible, and discards the components corresponding to the smallest eigenvalues in the original data matrix to reduce the dimension and at the same time minimize the loss of information. In PCA method, the genotype data matrix $\boldsymbol{G} = (g_i^s)_{M\times N}$ is usually standardized first. Patterson et al. (2006) set each entry in the standardized matrix $\widetilde{\boldsymbol{G}}$ is

$$\widetilde{g}_i^s = \frac{g_i^s - 2p^s}{\sqrt{p^s(1-p^s)}}, \tag{1.10}$$

where $p^s = \sum_{i=1}^N g_i^s/(2N)$. $p^s(1-p^s)$ is the variance form often used in genetic drift and can also be replaced by empirical variance of the data,

$$\widetilde{g}_i^s = \frac{g_i^s - 2p^s}{\sqrt{\frac{1}{N-1}\sum_{i=1}^N \left(g_i^s - 2p^s\right)^2}}. \tag{1.11}$$

And the more straightforward way is

$$\widetilde{g}_i^s = g_i^s - 2p^s. \tag{1.12}$$

(1.10), (1.11) and (1.12) can be used to deal with mean centering. However, when g_i^s is some constant value at SNP s in the selected N individuals, the data of SNP s should be removed before proceeding with (1.10) or (1.11). Denote the matrix $\boldsymbol{T} = \widetilde{\boldsymbol{G}}\boldsymbol{V}$ as the new matrix obtained by the transformation, where $\boldsymbol{V} = (v_{ij})_{N\times N}$ and $\sum_{i=1}^N v_{ij}^2 = 1$ such that the transformation does not change the scale of the data. Take the first column vector $\boldsymbol{t}_1$ of $\boldsymbol{T}$, then

$$\boldsymbol{t}_1^T\boldsymbol{t}_1 = \boldsymbol{v}_1\widetilde{\boldsymbol{G}}^T\widetilde{\boldsymbol{G}}\boldsymbol{v}_1,$$

where $\boldsymbol{v}_1 = (v_{11}, \ldots, v_{N1})^T$. When $\boldsymbol{v}_1$ is the eigenvector corresponding to the largest eigenvalue of $\widetilde{\boldsymbol{G}}^T\widetilde{\boldsymbol{G}}$, $\boldsymbol{t}_1^T\boldsymbol{t}_1$ takes the maximum value, which means that the first principal component captures the maximum possible variance of the variable in the original data. Following this principle, the other principal components capture the rest. Therefore, the resulting $\boldsymbol{V}$ is a matrix consisting of the eigenvectors corresponding to the eigenvalues arranged from largest to smallest. Then $\boldsymbol{T}\boldsymbol{V}^T = \widetilde{\boldsymbol{G}}$, it can be seen that, in the coordinate system with each principal component as the new coordinate axis, the elements in each row vector of $\boldsymbol{V}$ define the position of variables of the original data.

In PCA method, $\boldsymbol{V}$ is the matrix of the right singular vectors obtained by directly performing

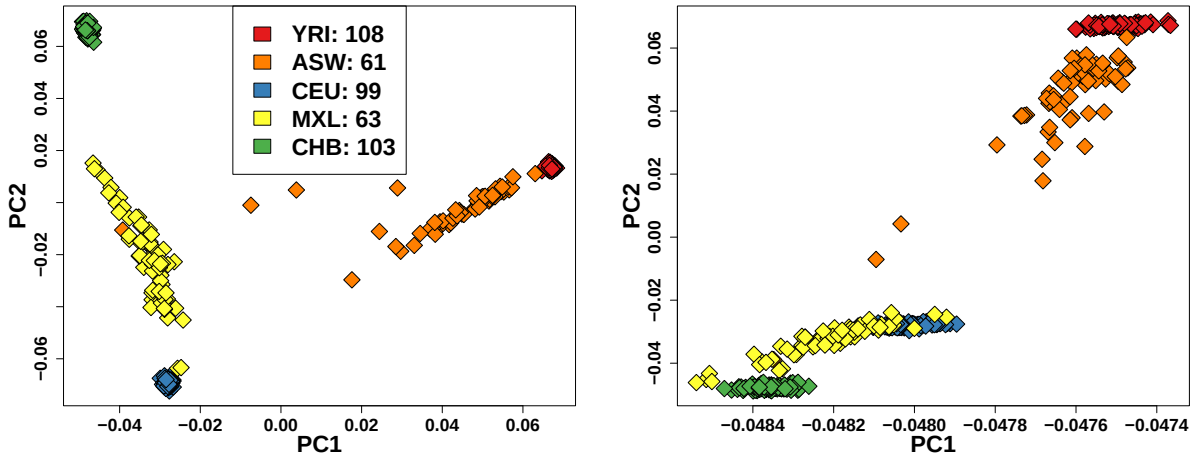
singular value decomposition (SVD) of $\widetilde{\boldsymbol{G}}$ or the matrix of the eigenvectors of $\widetilde{\boldsymbol{G}}^T\widetilde{\boldsymbol{G}}$, i.e.

$$\widetilde{\boldsymbol{G}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

or

$$\widetilde{\boldsymbol{G}}^T\widetilde{\boldsymbol{G}} = \boldsymbol{V}\boldsymbol{\Sigma}^*\boldsymbol{V}^T,$$

where $\boldsymbol{U}$ is an $M \times M$ matrix, $\boldsymbol{\Sigma}$ is an $M \times N$ rectangular diagonal matrix and $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$ is an $N \times N$ square diagonal matrix. In order to reduce the dimension of the new data matrix $\boldsymbol{T}$, only the first few principal components (PCs) are selected, which means that the eigenvectors corresponding to smaller eigenvalues are discarded. The top ranked PCs reflect the largest sample variation through the corresponding eigenvectors. The software EIGENSOFT (Price et al., 2006) is developed to explicitly model ancestry differences between cases and controls. The eigenvectors provide insights into variability among individuals and the axes of the eigenvectors with the largest eigenvalues are usually important in describing genetic variations (Byun et al., 2017). In practice, the first and second principal components are often selected to establish a two-dimensional coordinate system to display the sample and the inferred population structure. Taking the example from Figure 1.7, Figure 1.8 is the visualization of real human data using PCA method, which distinguishes various sample populations by PC1 and PC2. The comparison of between Figure 1.8(a) and Figure 1.8(b) shows the influence of mean centering on the PCA, including the change of the value range of two PC axes. In admixture models, estimating allele frequencies is often



(a) PCA on $\widetilde{\boldsymbol{G}}$ (b) PCA on $\boldsymbol{G}$

Figure 1.8: An example for PCA plotting the inferred population structure of real human data from the 1000 Genomes project (Auton et al., 2015; Garcia-Erill and Albrechtsen, 2020). (a) The PC plotting obtained after mean centralized data processing. (b) The PC plotting obtained without mean centralized data processing.

achieved using PCA. Assume that two alleles at the same locus of individual i are treated as identically distributed, and different locus are treated as independent of each other (Balding and Nichols, 1995), the genotypes follow a binomial distribution, i.e.

$$g_i^s \mid \pi_i^s \sim \mathcal{B}in\left(2, \pi_i^s\right).$$

Hao et al. (2016) propose that $\mathbf{\Pi}$ is estimated by forming the projection of $\boldsymbol{G}/2$ onto the top PCs of $\boldsymbol{G}$ with an explicit intercept. Set $K < N < M$, then the rank of $\mathbf{\Pi}$ is $K-1$ mean that the top $K-1$ PCs are selected by running SVD. The ALStructure algorithm (Cabreros and Storey, 2019) further proposes the estimation of $\boldsymbol{F}$ and $\boldsymbol{Q}$ by considering the boundary conditions, which is a unification of PCA-based and likelihood-based approaches. PCA can also be extended to infer population structure in the presence of related individuals who have alleles originate from a common ancestor. Conomos et al. (2015) perform PCA on those unrelated individuals separated from the sample as a subset of identified ancestral representatives, and then predicts the component of variation for all remaining related individuals based on genetic similarity.

1.1.4 Correlation of individuals

Related individuals issue

As described in the previous section, some parameters of unrelated individuals in the presence of population structure can be estimated using the clustering and PCA method. In term of the clustering methods, the admixture proportion estimated by running software such as ADMIXTURE can be inaccurate in the presence of related individuals, even if the number of clusters is given correctly in the population structure setting. Some close relatives may be considered clusters in these methods, mistaken for ancestral populations. In PCA methods, the top PCs fail to reflect the inferred population structure when related individuals are present in the data, but succeed when these individuals are removed (Price et al., 2010). Genetic studies often involve related individuals, which motivates the consideration of the parameters used to distinguish individual relationships.

Pedigree-based kinship

The relationships between individuals are established between the red circles in Figure 1.1. Two individuals are related if they have a recent common ancestor. Kinship or relatedness is a fundamental concept to describe the relationships between individuals, but it is difficult to define in one way (Speed and Balding, 2015). The common definition is based on a pedigree, in which the most recent common ancestor can also be defined. Non-random mating not only shapes population structure at the ancestral level, but also establishes relationships among individuals in the

pedigree. The pedigree is used to describe different members of a family and their relationships, such as parents and siblings (see Figure 1.9). In a pedigree, it is customary to mark individuals lacking parental lines as founders and others as non-founders.

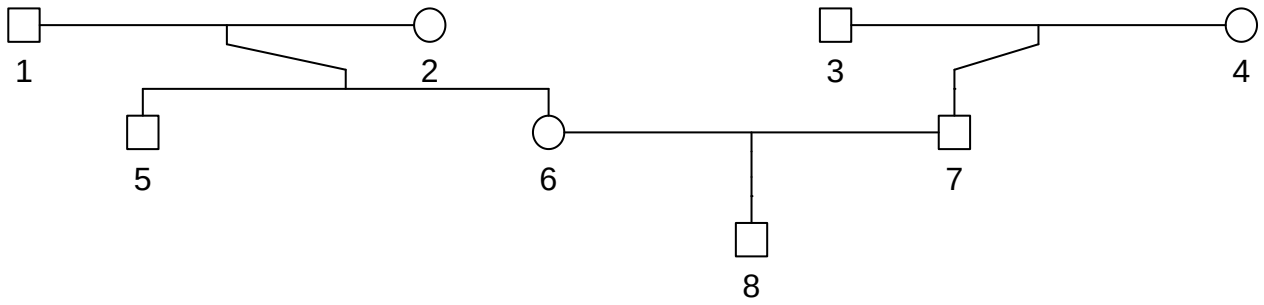


Figure 1.9: A pedigree map of 8 members marked with numbers. In this pedigree, the sexes are distinguished by squares and circles. Peers are connected by a horizontal line, parents and children are connected by vertical or diagonal lines. Individuals 1,2,3 and 4 are founders, while individuals 5,6,7 and 8 are non-founders.

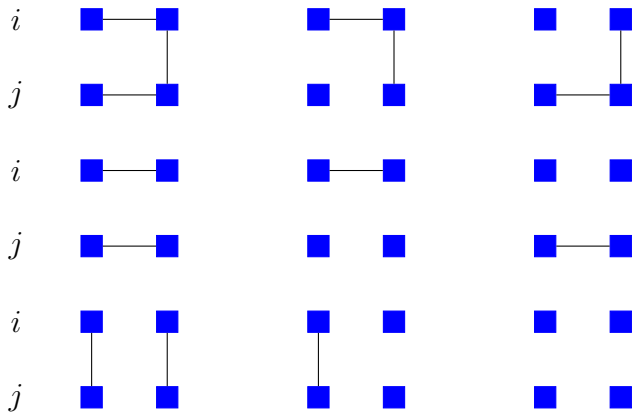


Figure 1.10: The blue square is the allele and line segment indicate that the connected pair of alleles are IBD. The three IBD states in the last row all exclude inbreeding: the number of alleles with IBD status from left to right is 2,1 and 0, respectively.

Two alleles that have originated from the replication of one single allele in a previous generation individual are said to be identical by descent (IBD) from that individual (Falconer, 1996; Thompson, 2013). IBD provides a basis for measuring inter- and intra-individual correlations. Figure 1.10 shows 9 IBD states of 4 alleles at a locus in two individuals. For a diploid individual i, the probability that a pair of alleles from i are IBD is defined as ψ_i, also known as the inbreeding

coefficient; and for two individuals i and j, the probability that a randomly selected allele from i and a randomly selected allele from j are IBD is defined as φ_{ij}, also known as the coancestry or the kinship coefficient of two individuals . In addition, the coefficient of self-kinship is defined as $\varphi_{ii} := (1 + \psi_i)/2$. Two individuals are said to be outbred if the four alleles only have the three possible cases in the last row of Figure 1.10, then $\varphi_{ij,2}, \varphi_{ij,1}$ and $\varphi_{ij,0}$ are defined as the probability that i and j share 2, 1 or 0 alleles of IBD at a locus, respectively, corresponding to three cases from left to right. When $\psi_i = \psi_j = 0$, the kinship coefficient can be expressed as follows

$$\varphi_{ij} = \frac{1}{2}\varphi_{ij,2} + \frac{1}{4}\varphi_{ij,1}.$$

With the above concepts, take Figure 1.9 as an pedigree example and consider the kinship for some pairs of individuals. Individual 8 has founders 1,2,3 and 4 as ancestors, individual 5 has founders 1 and 2 as ancestors, so the most recent common ancestors of both are founders 1 and 2. In the calculation of pedigree-based kinship, it is usually assumed that the founders are unrelated to each other, then $\psi_5 = \psi_6 = \psi_7 = \psi_8 = 0$ and

$$\varphi_{58} = \frac{1}{2}\frac{2}{4 \times 2} = \frac{1}{8}.$$

In Figure 1.9, it is observed that individuals 5 and 7 had no recent common ancestor in the pedigree even though both have ancestors, implying that $\varphi_{57} = 0$. Such calculation of the kinship coefficient depends on the structure of the pedigree and there is no complete pedigree in nature. As more ancestors are added to the pedigree, previously unrelated individuals may share a common ancestor and thus become related. In addition, the kinship coefficients may be the same for different relationships, for example,

$$\varphi_{56} = \frac{1}{2}\frac{2}{2 \times 2} = \frac{1}{4}$$

and

$$\varphi_{68} = \frac{1}{2}\frac{1}{1 \times 2} = \frac{1}{4}$$

define siblings and parent-offspring, respectively. Further considering the index $(\varphi_{ij,0}, \varphi_{ij,1}, \varphi_{ij,2})$ above to define the relation, then

$$\text{siblings:} \quad (\varphi_{56,0}, \varphi_{56,1}, \varphi_{56,2}) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$$

and

$$\text{parent-offspring:} \quad (\varphi_{68,0}, \varphi_{68,1}, \varphi_{68,2}) = (0, 1, 0).$$

Genotype-based correlation

If the structure of pedigree is not observed, correlation coefficients based on genotype data can be used to define the inbreeding and kinship coefficient in mathematical expression. Let $g_i^s = g_{i,1}^s + g_{i,2}^s$, where $g_{i,1}^s, g_{i,2}^s \in \{0, 1\}$ indicate whether the first or second allele at SNP s in individual i is A. Assume that at a locus s in a homogenous population of size N the frequency of allele A is p_s and $P(g_{i,1}^s = 1) = P(g_{i,2}^s = 1) = p_s$. Using pedigree-based inbreeding coefficient, the following expression

$$P(g_i^s = 2) = \psi_i p_s + (1 - \psi_i) p_s^2 \tag{1.13}$$

can be interpreted that if two alleles from individual i at SNP s are IBD with probability ψ_i, then only one allele is considered; otherwise, the two alleles are independent. In addition, $\psi_i = 0$ means that the individual is outbred. The estimated genotype frequencies for homozygotes at SNP s in the population can be defined as follows,

$$\widehat{f}(AA) = \frac{\sum_{i=1}^{N} I_{\{g_i^s = 2\}}}{N}$$

and

$$\widehat{f}(aa) = \frac{\sum_{i=1}^{N} I_{\{g_i^s = 0\}}}{N}.$$

where the function $I_{\{X=x\}}$ means if $X = x$ then $I = 1$, otherwise $I = 0$. Then using (1.13) the expected genotype frequencies in the presence of the inbreeding influence are

$$\mathbb{E}[\widehat{f}(AA)] = p_s^2 + p_s(1 - p_s)\bar{\psi}$$

and

$$\mathbb{E}[\widehat{f}(aa)] = (1 - p_s)^2 + p_s(1 - p_s)\bar{\psi},$$

where $\bar{\psi} = \sum_{i=1}^{N} \psi_i / N$. Compared the relationship between genotype and allele frequency described by the Hardy-Weinberg equilibrium (Hardy, 1908), it can be seen that the expected genotype frequencies for homozygotes increase by $p_s(1 - p_s)\bar{\psi}$ because of the inbreeding influence. Clearly,

$$\begin{aligned}
&\mathbb{E}(g_{i,1}^s) = \mathbb{E}(g_{i,2}^s) = p_s \\
&\mathbb{V}(g_{i,1}^s) = \mathbb{V}(g_{i,2}^s) = p_s(1 - p_s) \\
&\mathbb{E}(g_{i,1}^s g_{i,2}^s) = P(g_i^s = 2) = \psi_i p_s + (1 - \psi_i) p_s^2.
\end{aligned}$$

The inbreeding coefficient is then mathematically expressed as the following correlation coefficient

$$\psi_i = \frac{\mathbb{E}\{[g_{i,1}^s - \mathbb{E}(g_{i,1}^s)][g_{i,2}^s - \mathbb{E}(g_{i,2}^s)]\}}{\sqrt{\mathbb{V}(g_{i,1}^s)\mathbb{V}(g_{i,2}^s)}} = \rho(g_{i,1}^s, g_{i,2}^s),$$

where ρ is the correlation coefficient operator. Since ψ_i is constant across M SNPs, if p_s is known and $g^s_{i,1}, g^s_{i,2}$ are observed for all SNP s, an estimate of ψ_i can be presented as follows,

$$\widehat{\psi}_i = \frac{1}{M}\sum_{s=1}^{M}\frac{(g^s_{i,1}-p_s)(g^s_{i,2}-p_s)}{p_s(1-p_s)}.$$

And inspired by that, similarly,

$$\begin{aligned}\mathbb{E}(g^s_i) &= 2p_s\\ \mathbb{V}(g^s_i) &= 2p_s(1-p_s)(1+\psi_i).\end{aligned}$$

Using pedigree-based kinship coefficient, the following expression

$$P(g^s_{i,a}=1, g^s_{j,b}=1) = \varphi_{ij}p_s + (1-\varphi_{ij})p_s^2 \quad (a=1,2; b=1,2) \tag{1.14}$$

can be interpreted that if any two alleles from individual i and j at SNP s are IBD with probability φ_{ij}, then only one allele is considered; otherwise, the two alleles are independent. In addition, $\varphi_{ij}=0$ means that individual i and j are unrelated. Clearly,

$$\mathbb{E}(g^s_i g^s_j) = \mathbb{E}[(g^s_{i,1}+g^s_{i,2})(g^s_{j,1}+g^s_{j,2})].$$

The correlation between g^s_i and g^s_j is then mathematically expressed as follows,

$$\rho\left(g^s_i, g^s_j\right) = \frac{\mathbb{E}\{[g^s_i-\mathbb{E}(g^s_i)][g^s_j-\mathbb{E}(g^s_j)]\}}{\sqrt{\mathbb{V}(g^s_i)\mathbb{V}(g^s_j)}} = \frac{2\varphi_{ij}}{\sqrt{(1+\psi_i)(1+\psi_j)}}.$$

For two individuals that are not inbred, i.e. $\psi_i=\psi_j=0$, the above expression can be simplified to $\rho\left(g^s_i, g^s_j\right) = 2\varphi_{ij}$, where $0 \le \varphi_{ij} \le 0.5$. Let $\rho_{ij} = \sum_{s=1}^{M}\rho\left(g^s_i, g^s_j\right)/M$ denote the correlation between individual i and j, combined with the fact that φ_{ij} is constant over M SNPs, then $\varphi_{ij} = \rho_{ij}/2$ indicates that the kinship coefficient measures the degree of genetic similarity between individuals. When p_s is known and g^s_i, g^s_j are observed for all SNP s, an estimate of φ_{ij} can be presented as follows,

$$\widehat{\varphi}_{ij} = \frac{1}{4M}\sum_{s=1}^{M}\frac{(g^s_i-2p_s)(g^s_j-2p_s)}{p_s(1-p_s)}. \tag{1.15}$$

The above genotype-based correlation calculations require the assumption that alleles originate from a homogeneous population, where p_s is an known parameter. However, in practice, it makes more sense to have the presence of population structure, where p_s is replaced with the individual-specific allele frequency and unknown. Then a new estimate $\widehat{\varphi}'_{ij}$ is converted from (1.15) into the

following form

$$\widehat{\varphi}'_{ij} = \frac{1}{4M}\sum_{s=1}^{M}\frac{(g_i^s - 2\widehat{\pi}_i^s)(g_j^s - 2\widehat{\pi}_j^s)}{\sqrt{\widehat{\pi}_i^s(1-\widehat{\pi}_i^s)\widehat{\pi}_j^s(1-\widehat{\pi}_j^s)}}, \tag{1.16}$$

where $\widehat{\pi}_i^s$ and $\widehat{\pi}_j^s$ are estimates of the individual-specific allele frequencies π_i^s and π_j^s, respectively. Conomos et al. (2016) propose a complex approach called PC-Relate. This approach first uses PCs combined with linear regression to estimate allele frequencies for all individuals including related individuals, and then considers the following form

$$\widehat{\varphi}_{ij}^A = \frac{\sum_{s=1}^{M}(g_i^s - 2\widehat{\pi}_i^s)(g_j^s - 2\widehat{\pi}_j^s)}{4\sum_{s=1}^{M}\sqrt{\widehat{\pi}_i^s(1-\widehat{\pi}_i^s)\widehat{\pi}_j^s(1-\widehat{\pi}_j^s)}}, \tag{1.17}$$

as an estimate of the kinship coefficient. The difference between (1.16) and (1.17) is that the former is the average of the ratios across SNPs, while the latter is the ratio of the two averages across SNPs. For the convergence of the estimator, the former must satisfy more restrictive conditions than the latter (Ochoa and Storey, 2021). The latter has a asymptotic bias when individuals are related and have admixed ancestry. Instead of taking allele frequencies into account, Manichaikul et al. (2010) propose KING-robust kinship coefficient estimator

$$\widehat{\kappa}_{ij} = \frac{\sum_{s=1}^{M}\left[g_i^s\left(1-g_i^s\right) + g_j^s\left(1-g_j^s\right) + g_i^s g_j^s\right]}{\sum_{s=1}^{M}\left[g_i^s\left(2-g_i^s\right) + g_j^s\left(2-g_j^s\right)\right]} \tag{1.18}$$

based solely on genotype data. (1.18) is also used in PC-Relate method for preliminary screening of related and unrelated individuals, which is a key step.

1.2 Contributions and perspectives

1.2.1 Manuscript 1

The main contribution of this paper is to explore the behavior of the variation in the allele frequency over time by using the F-statistics, F_2 and F_{st}. The evolutionary force factors affecting allele frequency are set to function $g : [0,1] \to [0,1]$, and the Wright-Fisher model is modified to describe the changes in the number of individuals carrying the reference allele in the following form

$$Z_{t+1}|X_t = x_t \sim \mathcal{B}in(N_{t+1}, g(x_t)), \tag{1.19}$$

where Z_{t+1} is the number of individuals carrying the allele with population size N_{t+1} at generation $t+1$ and X_t is the allele frequency at previous generation t. We consider pure drift, mutation and

migration as linear evolutionary forces to give the form of g,

$$g(x) = (1-a)x + b, \quad 0 \leq b \leq a \leq 1, \tag{1.20}$$

where a and b are the two evolutionary force parameters that we use for analysis. In theory, we only use the first two moments of model (1.19) to calculate the explicit expression results of F_2 and F_{st}, which are proposed by Reich et al. (2009) and Wright (1951), respectively. On this basis, we give the conditions for determining the increase and decrease of the genetics difference over time and the specific parameter conditions satisfying the existence of the inflection point. We find that migration may cause a non-monotonic behavior of genetic difference. Since most of the real world population is affected by migration, then this leads to the conclusion that in most cases the behavior of the F-statistic of the real world population will be non-monotonic (Li and Wiuf, 2022).

To explain the method more clearly in this paper, some proposed expressions are limited to the case where the population size does not change over time, but are equally applicable to cases where population size varies. Based on the definition of effective population size described in the previous section, we can attempt to extend a similar analysis by giving such a value. In addition, the setting of g is limited to linear evolutionary forces, and nonlinear cases such as natural selection should also be included. For calculation methods involving only the first two moments, an explicit expression for F-statistics does not necessarily exist. The application of Taylor's expansion may reveal some of these results. Meanwhile, the diffusion approximation is an effective method to deal with the parameters, which simplifies the analysis and can be further extended to the case of infinite population size.

1.2.2 Manuscript 2

The main contribution of the second paper is to evaluate the admixture model fit by visualizing the correlation matrix of residuals defined as the differences between the observed and predicted values. Using the notation in **Manuscript 2**, we consider the genotype value G_{si} of individual $i \in \{1, \ldots, n\}$ at SNP $s \in \{1, \ldots, m\}$ as described by the following model

$$G_{si} \sim \mathcal{B}in(2, \Pi_{si}), \tag{1.21}$$

where Π_{si} is individual-specific allele frequency defined in the previous section and depends on the number of ancestral populations, their admixture proportions and the ancestral population allele frequencies. Using the matrix notation, (1.21) is

$$G \sim \mathcal{B}in(2, \Pi_k), \quad \Pi_k = F_k Q_k,$$

where matrix Q_k has dimension $k \times n$, matrix F_k has dimension $m \times k$, and k is the unknown number of ancestral populations. The process of finding a predicted value is essentially taking an estimate $\widehat{\Pi}$ of Π and plugging $\widehat{\Pi}$ into the following model

$$G' \sim \mathcal{B}in(2, \widehat{\Pi})$$

to obtain data points. Thus, the expected value of the resulting data point, i.e. $\mathbb{E}(G') = 2\widehat{\Pi}$, is taken as the predictive value. Let $\widehat{P}_{k'}$ be an orthogonal projection onto a k'-dimensional subspace, where k' is a number of ancestral populations that we suggest. Denoting $\widehat{\Pi}_{k'} = G\widehat{P}_{k'}$, the $m \times n$ matrix of residuals is defined as

$$\widehat{R}_{k'} = G - 2\widehat{\Pi}_{k'} = G(I - \widehat{P}_{k'}).$$

According to different methods, we can give the corresponding expression of $\widehat{P}_{k'}$ based on the data.

In **Manuscript 2**, we propose two correlation coefficient matrices of residuals. Define the $n \times n$ empirical covariance matrix $\widehat{B}$ with entries

$$\widehat{B}_{ij} = \frac{1}{m-1} \sum_{s=1}^{m} (\widehat{R}_{k',si} - \bar{R}_{k',i})(\widehat{R}_{k',sj} - \bar{R}_{k',j}),$$

where

$$\bar{R}_{k'i} = \frac{1}{m} \sum_{s=1}^{m} \widehat{R}_{k',si}$$

and $\widehat{R}_{k',si}$ are entries of $\widehat{R}_{k'}$. The corresponding empirical correlation matrix is denoted as $\widehat{b}$ with entries

$$\widehat{b}_{ij} = \frac{\widehat{B}_{ij}}{\sqrt{\widehat{B}_{ii}\widehat{B}_{jj}}}, \quad i, j = 1, \ldots, n.$$

The estimated covariance matrix is defined as

$$\widehat{C} = (I - \widehat{P}_{k'})\widehat{D}(I - \widehat{P}_{k'}),$$

where, $\widehat{D}$ is the $n \times n$ diagonal matrix containing the average heterozygosities of each individual, i.e.

$$\widehat{D}_{ii} = \frac{1}{m} \sum_{s=1}^{m} G_{si}(2 - G_{si}), \quad i = 1, \ldots, n,$$

and the corresponding estimated correlation matrix is denoted as $\widehat{c}$ with entries

$$\widehat{c}_{ij} = \frac{\widehat{C}_{ij}}{\sqrt{\widehat{C}_{ii}\widehat{C}_{jj}}}.$$

The above two kinds of correlation matrices are simple to compute and save the computational power. When the model is correct and the number of SNPs is large enough, the two correlation matrices are consistent. Therefore, the difference between two correlation matrices is expected to be close to zero without violating the admixture model. If the difference is significantly greater than zero, the model is misfit. In order to coordinate with the correlation matrix visualization, we give several theorems.

By using different PCA methods and the software ADMIXTURE for simulated and real data, the corresponding $\widehat{P}_{k'}$ is obtained. The validity of the model fitting determination is investigated, and it is in line with the expectation.

Correlation matrix visualization requires a given k', the number of ancient populations. The model fitting is affected by different k', although the correctness of k' does not mean the data fitting model. When PCA method is used to infer population structure, unadmixed samples can be distinguished through population number setting, which is reflected in the selection of the top principal components. In **Manuscript 2**, we consider three PCA methods: one PCA (named PCA 1 by us) proposed by Chen and Storey (2015) and the other two PCA methods proposed by Patterson et al. (2006). In the application of PCA 1, due to uncentralized data, we omit the principal component corresponding to the largest eigenvalue, which is different from the other two PCA methods. In the follow-up work, the unification of PCA methods can be improved, which is crucial for selecting k'. For data analysis, our results still need to be verified in many scenarios, such as whether the violation of several assumptions of the admixture model can be reflected when more possible F_{st} parameter values are taken into account. In addition, the genetic relationships resulting from recent hybridization can shape the existence of related individuals and affect the effectiveness of PCA for population structure inference. In such scenarios, the difference between the two correlation coefficient matrices in our approach is not significant. It is worth noting that, in principle, our approach requires a sufficient number of SNPs. The setting of the number of SNPs and sample size should reflect the validity of the proposed method in different scenarios.

1.2.3 Manuscript 3

The third manuscript is the preliminary work on kinship that I am currently doing independently. Let C be the number of founders shared by individuals i and j in a pedigree. Given $s \in \{1, 2, \ldots, M\}$ and the fact that the kinship coefficient is constant over SNPs, I present an estimate of the kinship coefficient based on pedigree and genotype,

$$\widehat{\varphi}_{ij} = \frac{1}{4} \cdot \frac{\sum_{s=1}^{M} \left(y_i^s - 2\widehat{\pi}_i^s\right)\left(y_j^s - 2\widehat{\pi}_j^s\right)}{\sum_{s=1}^{M} \frac{1}{C} \sum_{c=1}^{C} \widehat{\pi}_c^{(i,j)s}\left(1 - \widehat{\pi}_c^{(i,j)s}\right)}, \tag{1.22}$$

where $\widehat{\pi}_i^s, \widehat{\pi}_j^s$ are the corresponding individual-specific allele frequencies estimators, and $\widehat{\pi}_c^{(i,j)s}$ is the allele frequency estimator for the shared founder c, $c = 1, \ldots, C$. The difficulty of kinship estimation method shown in equation (1.22) lies in the estimation of allele frequency, especially for the part of shared founders. In the preliminary work of **Manuscript 3**, the full-siblings case is considered and a new estimation formula based on equation (1.22) is proposed,

$$\widehat{\varphi}_{ij}^F = \frac{\sum_{s=1}^M \left(y_i^s - 2\widehat{\pi}_i^s\right)\left(y_j^s - 2\widehat{\pi}_j^s\right)}{\sum_{s=1}^M \left[\left(y_i^s - 2\widehat{\pi}_i^s\right)^2 + \left(y_j^s - 2\widehat{\pi}_j^s\right)^2\right]}. \tag{1.23}$$

The estimation method of equation (1.23) only focuses on how to estimate individual-specific allele frequency. I consider the PCA 1 method in the second manuscript to obtain the allele frequency estimator. Before applying the PCA 1 method to the data set with the presence of related individuals, I follow the KING-robust estimator formula to separate unrelated individuals in the sample. In the simulation analysis, the validity of the estimation is verified.

The estimation formula and method proposed above have inspired me to extend to other situations. The present work deals with the Full-siblings case, and other more general pedigree cases have yet to be verified. For individuals with different admixture histories and population structure backgrounds, one issue is whether the estimates provided by equations (1.22) and (1.23) are better than those provided by other methods, such as PC-Relate and KING-robust estimator.

Bibliography

John H Gillespie. *Population genetics : a concise guide (2. ed.)*. Baltimore, Md.: The Johns Hopkins University Press, 2004.

R.A. Fisher. *The genetical theory of natural selection*. Oxford: Clarendon, 1930.

S Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

Simon Tavaré. *Part I: Ancestral Inference in Population Genetics*, pages 1–188. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-39874-5. doi: 10.1007/978-3-540-39874-5_1.

J.F. Crow and M. Kimura. *An introduction to population genetics theory*. New York, Evanston and London: Harper and Row, 1970.

S. Karlin and H. M. Taylor. *A Course in Stochastic Processes*. Wiley, New York, NY, 1980.

C Neuhauser. Mathematical models in population genetics. *Handbook of statistical genetics*, pages 153–178, 2001.

D. Reich, K. Thangaraj, N. Patterson, A. LPrice, and L. Singh. Reconstructing indian population history. *Nature*, 461:489–494, 2009.

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Karola Kirsanow, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I. Bos, Susanne Nordenfelt, Cesare de Filippo, Kay Prüfer, Susanna Sawyer, Cosimo Posth, Wolfgang Haak, Fredrik Hallgren, Elin Fornander, George Ayodo, Hamza A. Babiker, Elena Balanovska, Oleg Balanovsky, Haim Ben-Ami, Judit Bene, Fouad Berrada, Francesca Brisighelli, George Busby, Francesco Cali, Mikhail Churnosov, David E. C. Cole, Larissa Damba, Dominique Delsate, George van Driem, Stanislav Dryomov, Sardana A. Fedorova, Michael Francken, Irene Gallego Romero, Marina Gubina, Jean-Michel Guinet, Michael Hammer, Brenna Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Rick Kittles, Elza Khusnutdinova, Toomas Kivisild, Vaidutis Kučinskas, Rita Khusainova, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Robert W. Mahley, Béla Melegh, Ene Metspalu, Joanna Mountain, Thomas Nyambo, Ludmila Osipova, Jüri Parik, Fedor Platanov, Olga Posukh, Valentino Romano, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, Mikhail Voevoda, Joachim Wahl, Pierre Zalloua, Levon Yepiskoposyan, Tatijana Zemunik, Alan Cooper, Cristian Capelli, Mark G. Thomas, Sarah A. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Matthias Meyer, Evan E. Eichler, Joachim Burger, Montgomery Slatkin, Svante

Pääbo, Janet Kelso, David Reich, and Johannes Krause. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513:409–413, 2014. doi: 10.1038/nature13673.

N. J. Patterson, P. Moorjani, Y. Luo, S. Mallick, and N. Rohland et al. Ancient admixture in human history. *Genetics*, 192:1065–1093, 2012.

S Wright. The genetical structure of populations. *Annals of eugenics*, 15(4):323–354, 1951.

Richard Durrett. *Probability Models for DNA Sequence Evolution*. Springer New York, NY, 2008. doi: 10.1007/978-0-387-78168-6.

A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, and 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393.

G. Garcia-Erill and A. Albrechtsen. Evaluation of model fit of inferred admixture proportions. *Molecular Ecology Resources*, 20(4):936–949, 2020. doi: https://doi.org/10.1111/1755-0998.13171.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 06 2000. doi: 10.1093/genetics/155.2.945.

D. H. Alexander, J. Novembre, and K Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655—1664, 2009. doi: 10.1101/gr.094052.109.

Timothy Thornton, Hua Tang, Thomas J. Hoffmann, Heather M. Ochs-Balcom, Bette J. Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2012.05.024.

Montgomery Slatkin. Linkage disequilibrium —understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008. doi: 10.1038/nrg2361.

P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):786–792, 1978. doi: 10.1126/science.356262.

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLOS Genetics*, 2(12):1–20, 2006. doi: 10.1371/journal.pgen.0020190.

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. doi: 10.1038/ng1847.

Jinyoung Byun, Younghun Han, Ivan P. Gorlov, Jonathan A. Busam, Michael F. Seldin, and Christopher I. Amos. Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. *BMC Genomics*, 18(1):789, 2017. doi: 10.1186/s12864-017-4166-8.

David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1):3–12, 1995.

W Hao, M Song, and JD. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, 2016. doi: 10.1093/bioinformatics/btv641.

Irineo Cabreros and John D Storey. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. *Genetics*, 212(4):1009–1029, 2019. doi: 10.1534/genetics.119.302159.

M. P. Conomos, M. B. Miller, and T. A. Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic epidemiology*, 39(4):276–293, 2015. doi: 10.1002/gepi.21896.

Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010. doi: 10.1038/nrg2813.

Doug Speed and David J. Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015. doi: 10.1038/nrg3821.

D.S. Falconer. *Introduction to Quantitative Genetics, Ed. 3*. Longmans Green/John Wiley & Sons, Harlow, Essex, UK/New York, 1996.

Elizabeth A Thompson. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics*, 194(2):301–326, 06 2013. doi: 10.1534/genetics.112.148825.

G.H. Hardy. Mendelian Proportions in a Mixed Population. *Science*, 28(706):49–50, 1908. doi: 10.1126/science.28.706.49.

Matthew P. Conomos, Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1): 127–148, 2016. ISSN 0002-9297. doi: 10.1016/j.ajhg.2015.11.022.

Alejandro Ochoa and John D. Storey. Estimating FST and kinship for arbitrary population structures. *PLOS Genetics*, 17(1):e1009241, jan 2021. doi: 10.1371/journal.pgen.1009241.

Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26 (22):2867–2873, 2010. doi: 10.1093/bioinformatics/btq559.

Song Li and Carsten Wiuf. The behaviour of f-statistics over time. *bioRxiv*, 2022. doi: 10.1101/2022.08.25.505252. URL https://www.biorxiv.org/content/early/2022/08/26/2022.08.25.505252.

X. Chen and J.D. Storey. Consistent estimation of low-dimensional latent structure in high-dimensional data, 2015.

Chapter 2

Manuscript 1

The Behaviour of F-statistics over Time

Song Li[1] and Carsten Wiuf[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark.

Publication details: On *bioRxiv*.

The Behaviour of F-statistics over Time

Song Li† and Carsten Wiuf
Department of Mathematical Sciences, University of Copenhagen, Denmark

Abstract

We study the behaviour of the F_2-statistic and $F_{\rm st}$-statistic, respectively, over time in a Wright-Fisher model with mutation and migration. We give precise conditions for when the F_2-statistic is non-monotonic, that is, increases over time until a certain point and then starts decreasing. We show that even for small population sizes, the two statistics are well approximated by population size scaled expressions.

Keywords: Allele frequency; F-statistics; Wright–Fisher model; Linear evolutionary force; Finite population; Population genetics.

1 Introduction

The frequency of an allele in a population depends on various evolutionary forces. The Wright–Fisher model (Fisher, 1930; Wright, 1931) in its simplest form describes the evolution of the allele frequency at a single diploid site in a finite population, without overlapping generations, in the absence of any other evolutionary forces such as mutation and selection. For large populations sizes, using that the Wright-Fisher model can be regarded as a discrete time Markov process, continuous time diffusion approximations can be derived by scaling time and parameters in the population size (Crow and Kimura, 1970; Ewens, 2004; Karlin and Taylor, 1980; Neuhauser, 2001). The diffusion process is essentially determined by its first two moments, which greatly simplifies the exploration of the behavior of the allele frequency over time in large populations. Recently, other approximations have been proposed to study the distribution of the allele frequency over finitely many generations, based on the Wright-Fisher model, e.g., Balding and Nichols (1995); Nicholson et al. (2002); Foll and Gaggiotti. (2008); Coop et al. (2010); Gautier (2015); Haasl and Payseur (2016).

The main purpose of this paper is to explore the behavior of the two F-statistics, F_2 and F_{st}, that reflect frequency changes and population differentiation. These statistics depend

†Corresponding author. E-mail address: song.li@math.ku.dk

on the first two moments of the allele frequency distribution. For multiple populations, F_2 and F_{st} are related (Reich et al., 2009; Peter, 2016). The F_{st}, or the fixation index, is a measurement of population differences in alelle frequency and can be defined in two ways (Holsinger and Weir, 2009; Durrett, 2008). The F_{st}-statistic provides important insights into frequency processes within and between populations. Pure drift considered in the Wright-Fisher process is the simplest evolutionary force. However, evolution is complex and random, involving other factors such as mutation, migration, and natural selection (Cavalli-Sforza and Edwards, 1967). In general, evolutionary forces are divided into linear and nonlinear forms (Crow and Kimura, 1970). Linear forces are typically mutation and migration, which we also consider here. In this paper, we adoopt the definition of F_2 proposed by Reich et al. (2009), and the definition of F_{st} proposed by (Wright, 1951). The F_2 is defined as the square of the difference in allele frequency between two populations and has a range of mathematical properties (such as additivity) used in admixture inference (Patterson et al., 2012; Peter, 2016; Soraggi and Wiuf, 2019). Here, we study how F_2 and F_{st} vary over time. In particular, we show that migration might give rise to non-monotome behavior and analyse when this happens. We give precise conditions for when an inflection point occurs, that is, a time point after which the statisitc starts decreasing. Under pure drift, both statistics increase over time.

The paper is structured as follows. We describe the Wright-Fisher model in Section 2.1, allowing for mutation and migration. In Section 2.2, we give the definition of the F_2-statistic and the F_{st}-statistic, and find expressions for how they vary over time. We end with a discussion in Section 3. Proofs and mathematical details are collected in the Appendix.

2 Methods and Results

We consider a population of haploid individuals over generations. The frequency of the reference allele is set to $X_t \in [0, 1]$ in generation $t \geq 0$, and we are interested in the evolution of X_t over time. On this basis, we impose certain constraints on X_t, $t \geq 0$, to establish a model for its change.

Specifically, we consider a Wright-Fisher-like model with population size N_t in generation $t \geq 0$. The random number of individuals carrying the reference allele is denoted Z_t, hence

$X_t = Z_t/N_t$. In line with the Wirght-Fisher model, we assume the allele frequency of the next generation is determined by the previous generation and potentially external factors. The external influence acting on the allele can be defined as a function $g : [0, 1] \to [0, 1]$, and the model can be set up using a binomial distribution,

$$Z_{t+1}|X_t = x_t \sim \mathcal{B}in(N_{t+1}, g(x_t)). \tag{1}$$

Here, $\mathcal{B}in(n, p)$ is the binomial distribution with sample size n and probability p. The mathematical form of the effect g will be discussed below. The Wright-Fisher model can be extended to diploids by changing N to $2N$, or to non-constant population sizes by taking floating samples from each generation.

2.1 *Linear evolutionary pressure model*

We are mainly concerned with the influence of pure drift, mutation and migration on the evolution of the allele frequency (Cavalli-Sforza and Edwards, 1967; Cavalli-Sforza, 1973). The common characteristic is a linear constraint on how the frequency change (Siren, 2012). Specifically, we address the following cases.

In the case of mutation, a_1 is assumed to be the probability of mutation from the allele 'A' to the allele 'a', and b_1 is the probability from 'a' to 'A'. Therefore, if the frequency of 'A' carried by the parent is X_t, then we define

$$g(X_t) = X_t(1 - a_1) + (1 - X_t)b_1 = (1 - a_1 - b_1)X_t + b_1.$$

In the simplest case of migration (Tataru et al., 2015, 2016), individuals have the freedom to migrate in and out of the population. Assume that the probability of migration between populations is m and an infinitely large population with a constant allele frequency $X^* \in [0, 1]$. Then, we get the calculation

$$g(X_t) = (1 - m)X_t + mX^*. \tag{2}$$

We incorporate both the above mutation and migration into the model,

$$\begin{aligned} g(X_t) &= [(1 - a_1 - b_1)X_t + b_1](1 - m) + mX^* \\ &= \{1 - [m + (1 - m)(a_1 + b_1)]\}X_t + (1 - m)b_1 + mX^* \\ &:= (1 - a_3)X_t + b_3, \end{aligned} \tag{3}$$

where, $a_3 = m + (1-m)(a_1 + b_1)$, $b_3 = (1-m)b_1 + mX^*$. Obviously, by taking X_t to be 1, then $0 \leq 1 - a_3 + b_3 \leq 1$, so we need $0 \leq b_3 \leq a_3 \leq 1$ as a constrain.

We describe the above results directly in terms of the function g,

$$g(x) = (1-a)x + b, \quad 0 \leq b \leq a \leq 1, \quad 0 \leq x \leq 1. \tag{4}$$

In particular, when $a = b = 0$, the above model degenerates into the familiar pure drift case.

2.2 *Definitions of F-statistics*

In this section, we study the F_2 (Reich et al., 2009) and the F_{st} (Wright, 1951). The F_2 can be used to measure the difference in allele frequencies in a single population at different time points, and F_{st} describes this difference in multiple populations. The following parts are our results.

2.2.1 *The F_2*

In a single population, we consider two time points 0 and t, and express their allele frequencies as X_0 and X_t, respectively. Then the F_2 is defined as,

$$F_2(t) = F_2\left(X_0, X_t\right) = \mathbb{E}\left[\left(X_t - X_0\right)^2\right]. \tag{5}$$

Since we are interested in the variation of the allele frequency from its starting point, we assume throughout that $X_0 = x_0$, with $0 \leq x_0 \leq 1$, is fixed. Then $F_2(t)$ becomes

$$F_2(t) = \mathbb{E}[(X_t - x_0)^2] = Var(X_t) + \left[\mathbb{E}(X_t) - x_0\right]^2. \tag{6}$$

Under the linear evolutionary pressure model, let

$$\alpha := N_0 a, \qquad \gamma := \frac{b}{a},$$

with the convention that $0/0 := 1$, then $0 \leq \gamma \leq 1$. The variance can be given recursively in terms of the expectation, similarly to Tataru et al. (2015, 2016) and Siren (2012),

$$\begin{aligned} Var(X_t) &= \frac{1}{N_t}\mathbb{E}(X_t)[1 - \mathbb{E}(X_t)] + \left(1 - \frac{1}{N_t}\right)\left(1 - \frac{\alpha}{N_0}\right)^2 Var(X_{t-1}) \\ &= \sum_{i=1}^{t} \frac{1}{N_i}\left(1 - \frac{\alpha}{N_0}\right)^{2(t-i)} \mathbb{E}(X_i)[1 - \mathbb{E}(X_i)] \prod_{j=i+1}^{t}\left(1 - \frac{1}{N_j}\right), \end{aligned} \tag{7}$$

and

$$\begin{aligned}[\mathbb{E}(X_t) - x_0]^2 &= \left\{ \left(1 - \frac{\alpha}{N_0}\right)^t x_0 + \gamma \left[1 - \left(1 - \frac{\alpha}{N_0}\right)^t\right] - x_0 \right\}^2 \\ &= \left[1 - \left(1 - \frac{\alpha}{N_0}\right)^t\right]^2 (\gamma - x_0)^2 ,\end{aligned} \tag{8}$$

where

$$\mathbb{E}(X_i)[1 - \mathbb{E}(X_i)] = -\left(1 - \frac{\alpha}{N_0}\right)^{2i} (x_0 - \gamma)^2 + \left(1 - \frac{\alpha}{N_0}\right)^i (x_0 - \gamma)(1 - 2\gamma) + \gamma(1 - \gamma). \tag{9}$$

Equation (8) only depends on the population size through N_0. In contrast, (7) depends on N_t. Specifically, the effect of N_t is to locally slow down or speed up (compared to N_0) the change in the variance: doubling the population size corresponds to slowing time by a factor of two, at least for large population sizes.

Furthermore, by scaling the generation (time) and the parameter b in units of N_0, we introduce the notation,

$$\beta := N_0 b,$$

such that

$$\gamma = \frac{b}{a} := \frac{\alpha}{\beta}$$

is independent of N_0 and $t = \lfloor N_0 u \rfloor$, where $\lfloor \cdot \rfloor$ is the rounding operation and $u \in (0, \infty)$. We allow the population size to vary in a slow way, that is, we assume that there is a positive continuous function $h : (0, \infty) \to (0, \infty)$, such that

$$\lim_{N_0 \to \infty} \frac{N_{\lfloor N_0 u \rfloor}}{N_0} = h(u).$$

Such assumptions are widely used in population genetics, for example, to derive the diffusion limit of the Wright-Fisher model (Ewens, 2004).

Defining two limits $D_2(x_0, Y_u) = \lim_{N_0 \to \infty} F_2(\lfloor u N_0 \rfloor)$ and $Y_u = \lim_{N_0 \to \infty} X_{\lfloor u N_0 \rfloor}$ (Ewens, 2004), then we denote

$$D_2(u) = D_2(x_0, Y_u) = Var(Y_u) + [\mathbb{E}(Y_u) - x_0]^2 , \tag{10}$$

where, by defining $\Lambda(u) = \int_0^u 1/h(s) ds$ as the population-size intensity function (Griffiths

and Tavaré, 1994),

$$\begin{aligned} Var(Y_u) = &-(x_0-\gamma)^2\left[e^{-2\alpha u}-e^{-\Lambda(u)-2\alpha u}\right] \\ &+(x_0-\gamma)(1-2\gamma)\left[e^{-\alpha u}-e^{-\Lambda(u)-2\alpha u}-\alpha e^{-\Lambda(u)-2\alpha u}\int_0^u e^{\alpha r+\Lambda(r)}dr\right] \\ &+\gamma(1-\gamma)\left[1-e^{-\Lambda(u)-2\alpha u}-2\alpha e^{-\Lambda(u)-2\alpha u}\int_0^u e^{2\alpha r+\Lambda(r)}dr\right], \end{aligned}$$

and

$$[\mathbb{E}(Y_u)-x_0]^2=(x_0-\gamma)^2(1-e^{-\alpha u})^2.$$

Through the above definitions, we obtain the explicit expression $F_2(t)$ in the general case and the limit expression $D_2(u)$ in the infinite size case. In the following part, we give some theoretical properties through the discussion of parameters.

2.2.2 *Properties for the* F_2

We first present results for a single population under pure drift in a population of any size. In pure drift case, $g(x)=x$ and the expectation of frequency is $\mathbb{E}(X_t)=\mathbb{E}(X_{t-1})=x_0$. Then

$$F_2(t)=Var(X_t)=\frac{1}{N_t}\mathbb{E}(X_t)[1-\mathbb{E}(X_t)]+\left(1-\frac{1}{N_t}\right)Var(X_{t-1}).$$

It follows that

$$\begin{aligned} Var(X_t)-Var(X_{t-1}) &= \frac{1}{N_t}\left[\mathbb{E}(X_t)-(\mathbb{E}(X_t))^2-Var(X_{t-1})\right] \\ &= \frac{1}{N_t}\left[\mathbb{E}(X_{t-1})-(\mathbb{E}(X_{t-1}))^2-Var(X_{t-1})\right] \\ &= \frac{1}{N_t}\mathbb{E}[X_{t-1}(1-X_{t-1})], \end{aligned}$$

hence due to $0\le X_t\le 1$,

$$Var(X_t)\ge Var(X_{t-1}).$$

Proposition below gives the performance of $F_2(t)$ under pure drift.

Proposition. *Under pure drift, the variance and* $F_2(t)$ *are gradually increasing over generations in a single population of any size (Barton and Turelli, 2004; Peter, 2016).*

In the generous linear evolutionary pressure model, the change in $F_2(t)$ is not as straightforward as the case of pure drift. We assume $N_t\equiv N_0=N$ to simplify the analysis.

Using (7), (8) and (9), we replace the expression for $F_2(t)$ in (6) by an expression without the recursion,

$$\begin{aligned} F_2(t) = {} & \gamma\,(1-\gamma)\,\frac{1-\left(1-\frac{\alpha}{N}\right)^{2t}\left(1-\frac{1}{N}\right)^{t}}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} \\ & + (1-2\gamma)\,(x_0-\gamma)\left(1-\frac{\alpha}{N}\right)^{t}\frac{1-\left(1-\frac{\alpha}{N}\right)^{t}\left(1-\frac{1}{N}\right)^{t}}{1+\alpha\left(1-\frac{1}{N}\right)} \\ & + (\gamma-x_0)^2\left[1-2\left(1-\frac{\alpha}{N}\right)^{t}+\left(1-\frac{\alpha}{N}\right)^{2t}\left(1-\frac{1}{N}\right)^{t}\right]. \end{aligned} \tag{11}$$

The expression above is consistent with the results in the literature (Siren, 2012; Tataru et al., 2015). It follows that the parameters (x_0, γ, α, N) and $(1-x_0, 1-\gamma, \alpha, N)$ result in the same $F_2(t)$ value. Thus, we might assume the parameter x_0 lies in $[0, 0.5]$.

Theorem 1 shows that $F_2(t)$ is not always increasing. Define Δ_1 and Δ_2 by

$$\Delta_1 = (x_0-\gamma)\left[\frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)}+2(\gamma-x_0)\right]\ln\left(1-\frac{\alpha}{N}\right),$$

$$\Delta_2 = \left[\frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)}+\frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)}-(\gamma-x_0)^2\right]\ln\left[\left(1-\frac{\alpha}{N}\right)^2\left(1-\frac{1}{N}\right)\right].$$

Theorem 1. *Suppose $N_t \equiv N_0 = N$, then $F_2(t)$ has an inflection point if and only if $\Delta_1 < 0$, and $F_2(t)$ is non-increasing for all $t > \widehat{t}$, where*

$$\widehat{t} = \frac{\ln\frac{\Delta_1}{\Delta_2}}{\ln\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]}.$$

We replace the condition $\Delta_1 < 0$ by

$$\Delta_{11}\Delta_{12} > 0,$$

where,

$$\Delta_{11} = x_0-\gamma, \quad \Delta_{12} = \frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)}+2(\gamma-x_0).$$

Thus, Δ_{11} and Δ_{12} must have the same sign for the condition to hold. According to this criterion, Figure 1 shows the region $\Delta_1 < 0$ for $x_0 = 0.2, 0.4$ and $N = 5, 10, 50, \infty$. By allowing the population size to vary over time, clearly more complicated behaviors of $F_2(t)$ should be expected.

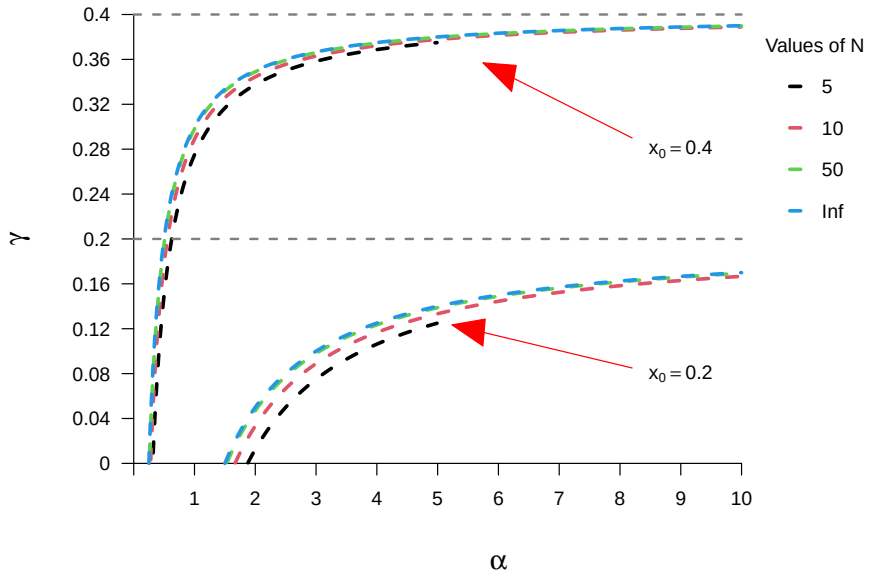


Figure 1: The region where $\Delta_1 < 0$. For $x_0 = 0.4$, the region is divided by a gray horizontal line $\gamma = 0.4$, axis of coordinates and a series of color lines at the top left corner of the figure; for $x_0 = 0.2$, the region is divided by a gray horizontal line $\gamma = 0.2$, axis of coordinates and a series of color lines at the bottom right corner of the figure.

Note that $\widehat{t}$ in Theorem 1 is defined as the inflection point of $F_2(t)$ from increase to decrease, whose visualization is analyzed in the following. Since the position of inflection point involves parameters N, x_0, α and γ, in order to display the change of $F_2(t)$ intuitively, we choose fixed $N = 100$ and $x_0 = 0.2$. By changing α and γ, we can give a heat map of inflection points, which shows the early and late appearance (see Figure 2).

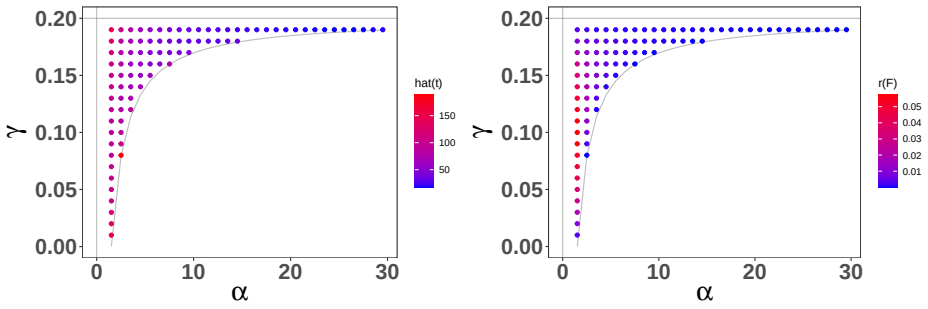


Figure 2: Shown is the heat map of $\widehat{t}$ and $r(F)$ under the condition of $N = 100$ and $x_0 = 0.2$.

As shown in Figure 2, we choose the inflection point is near 100 and get the parameters $\alpha = 1.5158, \gamma = 0.12$ that allow us to find the case where the inflection point should be (see Figure 3). As can be seen from Figure 3, $\widehat{t}$ is marked as the moment when the variation (compared to x_0) of population allele frequency is the greatest, and t at infinity represents a stationary state. The relative difference degree of $F_2(t)$ value between these two points (namely peak and plateau value) is also an indicator to understand the population genetic process. We give the following definition,

$$r(F) = \frac{F_2(\,\widehat{t}\,) - F_2\,(\infty)}{F_2\,(\infty)}.$$

Similarly, we show $r(F)$ for different α and γ in the heat map (see Figure 2). In general, we find $r(F)$ might be up to 0.1.

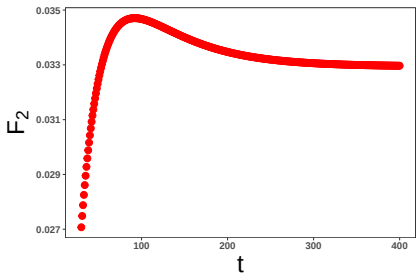


Figure 3: Shown is the population genetic evolution process with parameters $N = 100, x_0 = 0.2, \alpha = 1.5158$, and $\gamma = 0.12$. The F_2 has obvious peak and plateau value.

The case where N goes to infinity is defined as an infinite population model. In the

infinite population case, we take some steps to get similar results. Under pure drift, we have

$$D_2(u) = x_0(1-x_0)\left[1-e^{-\Lambda(u)}\right],$$

by setting $h \equiv 1$, $D_2(u)$ simplifies to

$$D_2(u) = x_0(1-x_0)\left(1-e^{-u}\right).$$

$D_2(u)$ in this case is increasing with u. In the generous linear evolutionary pressure model and by setting $h \equiv 1$, the expression (10) can be simplified to

$$\begin{aligned} D_2(u) = &(x_0-\gamma)^2\left[1-2e^{-\alpha u}+e^{-(2\alpha+1)u}\right] \\ &+\frac{1}{\alpha+1}(x_0-\gamma)(1-2\gamma)\left[e^{-\alpha u}-e^{-(2\alpha+1)u}\right] \\ &+\frac{1}{2\alpha+1}\gamma(1-\gamma)\left[1-e^{-(2\alpha+1)u}\right]. \end{aligned} \tag{12}$$

The parameters (x_0, γ, α) and $(1-x_0, 1-\gamma, \alpha)$ result in the same $D_2(u)$ value. Theorem 2 below gives a similar result to that of Theorem 1. Define Θ_1 and Θ_2 by

$$\Theta_1 = \frac{\alpha}{\alpha+1}(x_0-\gamma)(2x_0\alpha - 2\gamma\alpha + 2x_0 - 1), \tag{13}$$

$$\Theta_2 = \Theta_1 - x_0(1-x_0). \tag{14}$$

Theorem 2. *Suppose $h \equiv 1$, then $D_2(u)$ has an inflection point if and only if $\Theta_1 < 0$, and if this is the case then $D_2(u)$ is non-increasing for all $u > \widehat{u}$, where*

$$\widehat{u} = \frac{1}{\alpha+1}\ln\frac{\Theta_2}{\Theta_1}.$$

For consistency with the finite N case, we replace the condition $\Theta_1 < 0$ by

$$\Theta_{11}\Theta_{12} > 0,$$

where

$$\Theta_{11} = x_0 - \gamma, \quad \Theta_{12} = \frac{1-2\gamma}{1+\alpha} + 2(\gamma - x_0).$$

Thus, Θ_{11} and Θ_{12} must have the same sign for the condition to hold. According to this criterion, Figure 4 shows the region $\Theta_1 < 0$ for $x_0 = 0.2, 0.4, 0.5, 0.6$ and 0.8.

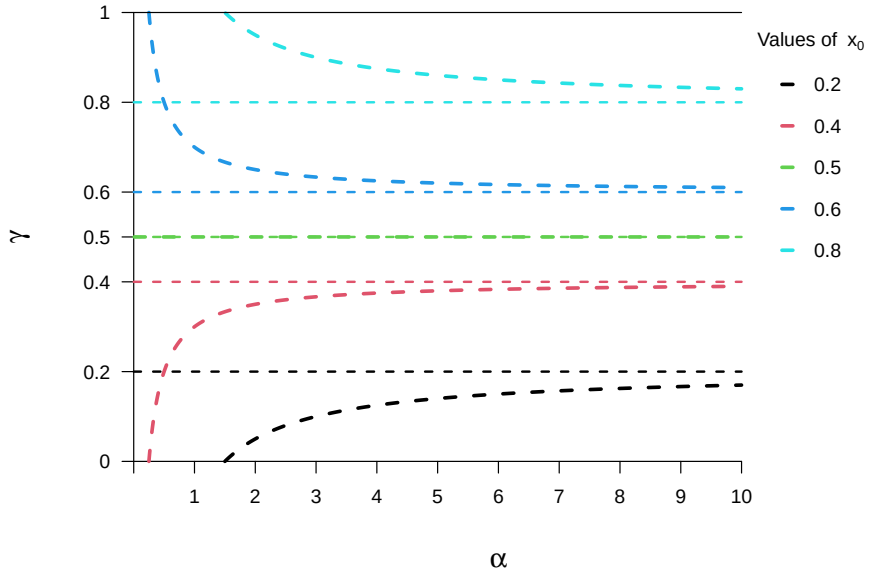


Figure 4: The region $\Theta_1 < 0$ for different values of x_0 in the case $N \to \infty$. For each given x_0, the region is divided by a colored horizontal line $\gamma = x_0$, a curve corresponding to the same color and axis of coordinates. For $x_0 = 0.5$, the region is empty.

As shown in the Figure 4, if $x_0 = 0.5$, then $\Theta_1 \geq 0$, that means $D_2(u)$ keeps increasing for any γ and α. In contrast, the initial value x_0 within a specified range can also determine that the inflection point must exist. The following corollary shows this and gives the specific location of the region.

Corollary. *Let $\alpha \geq 0, \gamma \in [0,1]$ be given, such that $\gamma \neq 0.5$, and let N be a natural number. Then $F_2(t)$ has an inflection point for any $x_0 \in (x_{0L}, x_{0R})$, where*

$$x_{0L} = \min\left(\gamma, \gamma + \frac{1-2\gamma}{2\left[1+\alpha\left(1-\frac{1}{N}\right)\right]}\right), \quad x_{0R} = \max\left(\gamma, \gamma + \frac{1-2\gamma}{2\left[1+\alpha\left(1-\frac{1}{N}\right)\right]}\right);$$

Simialrly, $D_2(u)$ has an inflection point for any $x_0 \in (x_{0L}, x_{0R})$, where

$$x_{0L} = \min\left(\gamma, \gamma + \frac{1-2\gamma}{2(1+\alpha)}\right), \quad x_{0R} = \max\left(\gamma, \gamma + \frac{1-2\gamma}{2(1+\alpha)}\right).$$

2.2.3 *The F_{st}*

For two or more populations, we study F_{st} which was first proposed by Sewall Wright (Wright, 1951) to measure the differences of allele frequencies among populations. Note that there are some different definitions of F_{st} (Nei, 1986; Holsinger and Weir, 2009; Durrett, 2008). For two different populations P_1 and P_2, we use the following definition of F_{st} in terms of probability,

$$F_{\mathrm{st}}\left(P_1, P_2\right) = \frac{q_1 - q_2}{1 - q_2},$$

where q_1 and q_2 represent the probability that two given reference alleles are the same from within and between populations, respectively. Suppose that the reference allele has frequency $X_{t_1,1}$, $X_{t_2,2}$ in population P_1 with population size $N_{t_1,1}$ at time t_1, and P_2 with population size $N_{t_2,2}$ at time t_2, respectively. Then, we define q_1 and q_2 as follows (Reich et al., 2009)

$$\begin{aligned} q_1 &= 1 - X_{t_1,1}(1 - X_{t_1,1}) - X_{t_2,2}(1 - X_{t_2,2}), \\ q_2 &= X_{t_1,1}X_{t_2,2} + (1 - X_{t_1,1})(1 - X_{t_2,2}) \\ &= 1 - X_{t_1,1} - X_{t_2,2} + 2X_{t_1,1}X_{t_2,2}. \end{aligned} \tag{15}$$

Using equation (15), we have

$$\begin{aligned} q_1 - q_2 &= \left(X_{t_1,1} - X_{t_2,2}\right)^2, \\ 1 - q_2 &= X_{t_1,1} + X_{t_2,2} - 2X_{t_1,1}X_{t_2,2}. \end{aligned} \tag{16}$$

As q_1, q_2 are stochastic variables, we adopt the following definition, replacing P_1, P_2, with t_1, t_2, respectively,

$$F_{st}\left(t_1, t_2\right) = \frac{\mathbb{E}\left(X_{t_1,1} - X_{t_2,2}\right)^2}{\mathbb{E}\left(X_{t_1,1} + X_{t_2,2} - 2X_{t_1,1}X_{t_2,2}\right)}.$$

We regard F_{st} as a function that changes over time. For two different populations, if we only consider the linear evolutionary pressure model, the two branching populations from a common ancestral population are independent of each other in the subsequent evolutionary process (Hansen and Martins, 1996), i.e., $\mathrm{Cov}(X_{t_1,1}, X_{t_2,2} \mid X_0) = 0$. If X_0 is fixed, indirectly, we have $\mathrm{Cov}(X_{t_1,1}, X_{t_2,2}) = 0$.

In the following, we set $t_1 = t_2 = t$, $N_{t_1,1} = N_{t_2,2} = N_t$, $N_t \equiv N_0 = N$ (indicating $h \equiv 1$), $F_{\mathrm{st}}(t) = F_{\mathrm{st}}(t,t)$ and using the previous notation, we denote $F_{\mathrm{st}}(t)$ as $D_{st}(u)$ when N is

large. To facilitate the application of the above results, according to the definitions (6) and (10), we split $F_{st}(t)$ and $D_{st}(t)$ as follows,

$$F_{\mathrm{st}}(t) = \frac{F_2\left(x_0, X_{t,1}\right) + F_2\left(x_0, X_{t,2}\right) - 2\mathbb{E}\left(X_{t,1} - x_0\right)\mathbb{E}\left(X_{t,2} - x_0\right)}{\mathbb{E}\left(X_{t,1} + X_{t,2}\right) - 2\mathbb{E}(X_{t,1})\mathbb{E}(X_{t,2})} \tag{17}$$

and

$$D_{st}(u) = \frac{D_2\left(x_0, Y_{u,1}\right) + D_2\left(x_0, Y_{u,2}\right) - 2\mathbb{E}\left(Y_{u,1} - x_0\right)\mathbb{E}\left(Y_{u,2} - x_0\right)}{\mathbb{E}\left(Y_{u,1} + Y_{u,2}\right) - 2\mathbb{E}(Y_{u,1})\mathbb{E}(Y_{u,2})}, \tag{18}$$

where $Y_{u,i} = \lim_{N\to\infty} X_{\lfloor uN \rfloor, i}$, $i = 1, 2$.

If pure drift only is considered in the evolution of two populations, then expression (17) and (18) degenerate to

$$F_{\mathrm{st}}(t) = \frac{\mathrm{Var}(X_t)}{x_0(1-x_0)} \quad \text{and} \quad D_{\mathrm{st}}(u) = \frac{\mathrm{Var}(Y_u)}{x_0(1-x_0)}.$$

Based on the results for $F_2(t)$ (and $D_2(u)$), then $F_{\mathrm{st}}(t)$ (and $D_{\mathrm{st}}(u)$) is gradually increasing over generations for two different populations of any size under the pure drift. In the previous section, we elaborated on the properties of $F_2(t)$ and $D_2(u)$, which are consistent, and in the following, we only focus on the infinite population size case of $D_{\mathrm{st}}(u)$ and the situation in which an inflection point occurs.

If we consider pure drift as the only evolutionary force factor for P_1 and the general linear evolutionary pressure model for P_2, then (18) degenerates to

$$D_{\mathrm{st}}(u) = \frac{Var(Y_{u,1}) + D_2\left(x_0, Y_{u,2}\right)}{x_0 + \mathbb{E}(Y_{u,2}) - 2x_0\mathbb{E}(Y_{u,2})},$$

where,

$$\mathbb{E}(Y_{u,2}) = e^{-\alpha u}\left(x_0 - \gamma\right) + \gamma,$$

$$Var(Y_{u,1}) = x_0(1-x_0)\left(1 - e^{-u}\right),$$

and using (13), (14), the expression (12) can be simplified to

$$D_2\left(x_0, Y_{u,2}\right) = -\frac{\Theta_1}{\alpha}e^{-\alpha u} + \frac{\Theta_2}{2\alpha+1}e^{-(2\alpha+1)u} + (x_0-\gamma)^2 + \frac{\gamma(1-\gamma)}{2\alpha+1}.$$

Similar to D_2, in order to find out whether D_{st} has an inflection point (from increasing to decreasing), we make the following analysis. Consider two non-negative functions f_1 and f_2 are differentiable, the chain rule says,

$$\left(\frac{f_1}{f_2}\right)' = \frac{f_1' f_2 - f_1 f_2'}{f_2^2}.$$

We set

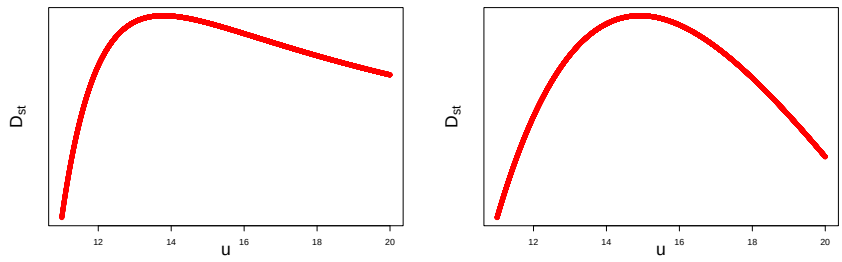
$$f_1 = Var(Y_{u,1}) + D_2\left(x_0, Y_{u,2}\right),$$

$$f_2 = (1-2x_0)\mathbb{E}(Y_{u,2}) + x_0.$$

If $f_1^{'}f_2 - f_1 f_2^{'} < 0$, then $(f_1/f_2)' < 0$. We extract the sign of $(f_1/f_2)'$ by $f_1^{'}f_2 - f_1 f_2^{'}$. In the first step, the chain rule is used to find the approximate range of each parameter when the inflection point exists. In the second step, the parameters are selected according to the range to determine the approximate position of the inflection point. In the first step we consider the following limiting form as a case and give a result under $\alpha < 1$ (see **Appendix** B.1),

$$\begin{aligned}\frac{f_1' f_2 - f_1 f_2'}{e^{-\alpha u}} \to \Theta_1\left[x_0 + \gamma\left(1-2x_0\right)\right] &+ \alpha x_0\left(1-x_0\right)\left(1-2x_0\right)\left(x_0-\gamma\right) \\ &+ \alpha\left(1-2x_0\right)\left(x_0-\gamma\right)^3 \\ &+ \frac{\alpha}{2\alpha+1}\left(1-2x_0\right)\left(x_0-\gamma\right)\gamma(1-\gamma), \quad as\ u\to\infty.\end{aligned}$$

The above limit result contains α (Θ_1 also contains α) and Θ_1 contains $(x_0 - \gamma)$, which means that $(1-2x_0)(x_0-\gamma)$ determines the sign of all terms except $\Theta_1 x_0$ in the limit result. Consider fixing a positive value of α, then take the parameters x_0 and γ, s.t. $(1-2x_0)(x_0-\gamma) < 0$, and check the case where the limit is negative. Based on this process, we get the parameters $\alpha = 0.1, \gamma = 0.31$ and $x_0 = 0.3$ as a case where the inflection point should be (see Figure 5(a)). The above parameters were used to draw the curve of D_{st} with u. As shown in Figure 5(a), the inflection point does exist and is near 14.



(a) (b)

Figure 5: Shown is the curve of D_{st} with u. The inflection point does exist.

If we consider the same linear evolutionary pressure model for P_1 and P_2, then (18) can be simplified as

$$D_{st}(u) = \frac{D_2\left(x_0, Y_u\right) - \left(\mathbb{E}(Y_u) - x_0\right)^2}{\mathbb{E}(Y_u)\left(1 - \mathbb{E}(Y_u)\right)},$$

where,

$$\mathbb{E}(Y_u) = e^{-\alpha u}\left(x_0 - \gamma\right) + \gamma,$$

$$D_2\left(x_0, Y_u\right) = -\frac{\Theta_1}{\alpha}e^{-\alpha u} + \frac{\Theta_2}{2\alpha+1}e^{-(2\alpha+1)u} + \left(x_0 - \gamma\right)^2 + \frac{\gamma(1-\gamma)}{2\alpha+1}.$$

We set

$$f_1 = D_2\left(x_0, Y_u\right) - \left(\mathbb{E}(Y_u) - x_0\right)^2,$$

$$f_2 = \mathbb{E}(Y_u)\left(1 - \mathbb{E}(Y_u)\right),$$

and take the limit to extract the sign part (see **Appendix** B.2),

$$\frac{f_1' f_2 - f_1 f_2'}{e^{-\alpha u}} \to \left(x_0 - \gamma\right)\left(1 - 2\gamma\right)\gamma(1-\gamma)\frac{-\alpha^2}{(2\alpha+1)(\alpha+1)}, \quad as\ u \to \infty.$$

Obviously, we only need to consider x_0 and γ, s.t. $\left(x_0 - \gamma\right)\left(1 - 2\gamma\right) > 0$. We give the parameters, $\alpha = 0.1, \gamma = 0.4$ and $x_0 = 0.6$, to support our judgment (see Figure 5(b)). As shown in Figure 5(b), the inflection point is near 15.

3 Random migration rates

In the case of migration, we study the simplest linear form,

$$g(X_t) = (1 - m)X_t + mX^*,$$

where the migration probability m is assumed to be fixed. In nature, however, populations migrate differently over time. Environmental climate, population size and other factors always affect the probability of migration in and out. Taking time dependence and randomness into account, we denote m_t as the migration probability, s.t.,

$$m_t \overset{i.i.d}{\sim} (m, \sigma^2), \text{ and } m_t \perp X_t.$$

Note that the distribution of m_t is ignored and only its first two moments are marked as $m, \sigma^2 > 0$. Assume $X_t^* \equiv x^*$, then

$$g(X_t) = (1 - m_t)X_t + m_t x^*, \quad 0 \le m_t, x^* \le 1.$$

Following the Wright-Fisher model and approach taken in the previous section, we also give explicit formula for $F_2(t)$ and $D_2(u)$. The first two moments of X_t can be obtained,

$$\begin{aligned}\mathbb{E}(X_t) = \mathbb{E}[g(X_{t-1})] &= \mathbb{E}(1-m_{t-1})\mathbb{E}(X_{t-1}) + \mathbb{E}(m_{t-1}x^*) \\ &= (1-m)\mathbb{E}(X_{t-1}) + mx^* \\ &= ... \\ &= (1-m)^t(x_0-x^*) + x^*,\end{aligned} \tag{19}$$

and

$$\begin{aligned}Var(X_t) &= \frac{1}{N_t}\mathbb{E}(X_t)(1-\mathbb{E}(X_t)) + (1-\frac{1}{N_t})Var[g(X_{t-1})] \\ &= \frac{1}{N_t}\mathbb{E}(X_t)(1-\mathbb{E}(X_t)) + (1-\frac{1}{N_t})\{\mathbb{E}[g^2(X_{t-1})] - [\mathbb{E}[g(X_{t-1})]]^2\} \\ &= \frac{1}{N_t}\mathbb{E}(X_t)(1-\mathbb{E}(X_t)) + (1-\frac{1}{N_t})\sigma^2[\mathbb{E}(X_{t-1}) - x^*]^2 \\ &\quad + (1-\frac{1}{N_t})[\sigma^2 + (1-m)^2]Var(X_{t-1}) \\ &= \sum_{i=1}^{t}\frac{1}{N_i}[\sigma^2 + (1-m)^2]^{t-i}\mathbb{E}(X_i)(1-\mathbb{E}(X_i))\prod_{j=i+1}^{t}(1-\frac{1}{N_j}) \\ &\quad + \sigma^2\sum_{i=1}^{t}[\sigma^2 + (1-m)^2]^{t-i}[\mathbb{E}(X_{i-1}) - x^*]^2\prod_{j=i}^{t}(1-\frac{1}{N_j}) \\ &= \sum_{i=1}^{t}\frac{1}{N_i}[\sigma^2 + (1-m)^2]^{t-i}[(1-m)^i(x_0-x^*) + x^*][1-(1-m)^i(x_0-x^*) - x^*]\prod_{j=i+1}^{t}(1-\frac{1}{N_j}) \\ &\quad + \sigma^2\sum_{i=1}^{t}[\sigma^2 + (1-m)^2]^{t-i}(1-m)^{2i}(x_0-x^*)^2\prod_{j=i}^{t}(1-\frac{1}{N_j}),\end{aligned} \tag{20}$$

then by (19) we get

$$[\mathbb{E}(X_t) - x_0]^2 = [1-(1-m)^t]^2(x^*-x_0)^2. \tag{21}$$

Combining (20) and (21), we follow the definition (6)

$$\begin{aligned}F_2(t) &= [1-(1-m)^t]^2(x^*-x_0)^2 \\ &\quad + \sigma^2\sum_{i=1}^{t}[\sigma^2 + (1-m)^2]^{t-i}(1-m)^{2i}(x_0-x^*)^2\prod_{j=i}^{t}(1-\frac{1}{N_j}) \\ &\quad + \sum_{i=1}^{t}\frac{1}{N_i}[\sigma^2 + (1-m)^2]^{t-i}[(1-m)^i(x_0-x^*) + x^*][1-(1-m)^i(x_0-x^*) - x^*]\prod_{j=i+1}^{t}(1-\frac{1}{N_j}).\end{aligned} \tag{22}$$

By scaling parameters m and σ^2 in units of N_0, we introduce

$$\epsilon_1 := N_0 m, \quad \epsilon_2 := N_0 \sigma^2.$$

Following the previous definitions, we have

$$\mathbb{E}\left(Y_u\right) = e^{-\epsilon_1 u}\left(x_0 - x^*\right) + x^*, \tag{23}$$

and

$$\begin{aligned}\mathrm{Var}\left(Y_u\right) = &-\left(x_0 - x^*\right)^2\left[e^{-2\epsilon_1 u} - e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\right] \\ &+ \left(x_0 - x^*\right)\left(1 - 2x^*\right)\left[e^{-\epsilon_1 u} - e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u} - (\epsilon_1 - \epsilon_2)e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\int_0^u e^{(\epsilon_1-\epsilon_2)r+\Lambda(r)}dr\right] \\ &+ x^*(1 - x^*)\left[1 - e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u} - (2\epsilon_1 - \epsilon_2)e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\int_0^u e^{(2\epsilon_1-\epsilon_2)r+\Lambda(r)}dr\right],\end{aligned} \tag{24}$$

then by (23) we get

$$\left[\mathbb{E}\left(Y_u\right) - x_0\right]^2 = \left(x_0 - x^*\right)^2\left(1 - e^{-\epsilon_1 u}\right)^2. \tag{25}$$

Combining (24) and (25), we follow the definition (10)

$$\begin{aligned}D_2(u) = &\left(x_0 - x^*\right)^2\left[1 - 2e^{-\epsilon_1 u} + e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\right] \\ &+ \left(x_0 - x^*\right)\left(1 - 2x^*\right)\left[e^{-\epsilon_1 u} - e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u} - (\epsilon_1 - \epsilon_2)e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\int_0^u e^{(\epsilon_1-\epsilon_2)r+\Lambda(r)}dr\right] \\ &+ x^*(1 - x^*)\left[1 - e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u} - (2\epsilon_1 - \epsilon_2)e^{-\Lambda(u)-2\epsilon_1 u+\epsilon_2 u}\int_0^u e^{(2\epsilon_1-\epsilon_2)r+\Lambda(r)}dr\right],\end{aligned} \tag{26}$$

With the expressions of $F_2(t)$ and $D_2(u)$, we can use the method in the paper to reasonably discuss the parameters and the existence of inflection points. And the study of F_{st} with respect to F_2, further research results will be carried out in our future work.

4 Discussion

We found expressions for the F_2-statistic and the F_{st}-statistic and how they vary over time. In general, we find that even for small population sizes, the behavior of the two statistics are well approximated by large population scaled expressions, considering time and paprameters scaled in units of population size.

Of particular interest is that migration might give rise to non-monotonic behavior. As real world most populations are subjected to migration, then this points to the conclusion that the behavior of the F-statistics for real world population in most cases will be non-monotonic.

It is worth mentioning that our proposed method only considers the first two moments of the allele frequency X_t, which is also applicable to other cases as long as the second moment exists. In such a case, we can still give reasonable results for the linear representation of evolutionary forces, the expression of F-statistics and related parameters analysis. In population genetic studies, nonlinear factors such as natural selection are usually considered in order to explore the process of allele frequency change caused by evolutionary forces. In such a complex study, it is not hard to imagine that there would be no explicit expression for F-statistics. Therefore, the diffusion approximation of this setting is a mean of conducting similar analysis (Ewens, 2004; Lacerda and Seoighe, 2014; Tataru et al., 2015). We consider these as future research directions.

Acknowledgements

The authors acknowledge the financial support from the funding agency of China Scholarship Council. The authors are supported by the Independent Research Fund Denmark (grant number: 8021-00360B) and the University of Copenhagen through the Data+ initiative.

References

R.A. Fisher. *The genetical theory of natural selection*. Oxford: Clarendon, 1930.

S Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

J.F. Crow and M. Kimura. *An introduction to population genetics theory*. New York, Evanston and London: Harper and Row, 1970.

W.J. Ewens. *Mathematical Population Genetics 1: I. Theoretical Introduction*. Springer Science and Business Media, 2004.

S. Karlin and H. M. Taylor. *A Course in Stochastic Processes*. Wiley, New York, NY, 1980.

C Neuhauser. Mathematical models in population genetics. *Handbook of statistical genetics*, pages 153–178, 2001.

D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995. doi: 10.1007/BF01441146.

J. K. Nicholson, J. Connelly, J. C. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature reviews. Drug discovery*, 1(2):153–161, 2002. doi: 10.1038/nrd728.

Matthieu Foll and Oscar Gaggiotti. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2):977–993, 2008. doi: 10.1534/genetics.108.092221.

G Coop, D Witonsky, A Di Rienzo, and JK. Pritchard. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423, 2010. doi: 10.1534/genetics.110.114819.

M Gautier. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4):1555–1579, 2015. doi: 10.1534/genetics.115.181453.

R. J. Haasl and B. A. Payseur. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular ecology*, 25(1):5–23, 2016. doi: 10.1111/mec.13339.

D. Reich, K. Thangaraj, N. Patterson, A. LPrice, and L. Singh. Reconstructing Indian population history. *Nature*, 461:489–494, 2009.

BM Peter. Admixture, Population Structure, and F-Statistics. *Genetics*, 202(4):1485–1501, 2016. doi: 10.1534/genetics.115.183913.

K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics*, 10(9):639–650, 2009. doi: 10.1038/nrg2611.

Richard Durrett. *Probability Models for DNA Sequence Evolution*. Springer New York, NY, 2008. doi: 10.1007/978-0-387-78168-6.

L.L. Cavalli-Sforza and A.W. Edwards. Phylogenetic analysis: Models and estimation procedures. *Am J Hum Genet*, 19(3 Pt 1):233–257, 1967.

S Wright. The genetical structure of populations. *Annals of eugenics*, 15(4):323–354, 1951.

N. J. Patterson, P. Moorjani, Y. Luo, S. Mallick, and N. Rohland et al. Ancient admixture in human history. *Genetics*, 192:1065–1093, 2012.

Samuele Soraggi and Carsten Wiuf. General theory for stochastic admixture graphs and F-statistics. *Theoretical Population Biology*, 125:56–66, 2019. doi: 10.1016/j.tpb.2018.12.002.

L.L. Cavalli-Sforza. Analytic review: some current problems of human population genetics. *Am J Hum Genet*, 25(1):82–104, 1973.

Jukka Siren. *Statistical models for inferring the structure and history of populations from genetic data*. PhD thesis, Finland, 2012.

Paula Tataru, Thomas Bataillon, and Asger Hobolth. Inference Under a Wright-Fisher Model Using an Accurate Beta Approximation. *Genetics*, 201, 08 2015. doi: 10.1534/genetics.115.179606.

Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. *Systematic Biology*, 66(1): e30–e46, 08 2016. ISSN 1063-5157. doi: 10.1093/sysbio/syw056.

R.C. Griffiths and S Tavaré. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*, 344(1310):403–410, 06 1994. doi: 10.1098/rstb.1994.0079.

N. H. Barton and Michael Turelli. Effects of genetic drift on variance components under a general model of epistasis. *Evolution*, 58(10):2111–2132, 2004.

M. Nei. Definition and Estimation of Fixation Indices. *Evolution*, 40(3):643–645, 1986. doi: 10.2307/2408586.

T. F. Hansen and E. P. Martins. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4):1404–1417, 08 1996. doi: 10.1111/j.1558-5646.1996.tb03914.x.

M. Lacerda and C. Seoighe. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, 198(3):1237–1250, 2014. doi: 10.1534/genetics.114.167957.

Appendix A Proofs

In appendix, we prove the theorems and corollary stated in the main text.

Random mating leads to a count of A alleles in generation $t+1$ that is binomially distributed,

$$Z_{t+1}|X_t = x_t \sim \mathcal{B}in(N_{t+1}, g(x_t)). \tag{27}$$

The goal is to account for effects of evolutionary and demographic forces on allele frequencies over time. We first calculate the first two moments of the allele frequencies.

$$\mathbb{E}(X_t) = \mathbb{E}[\mathbb{E}(X_t|X_{t-1})] = \mathbb{E}[g(X_{t-1})] \tag{28}$$

$$Var(X_t) = \frac{1}{N_t}\mathbb{E}(X_t)[1-\mathbb{E}(X_t)] + \left(1-\frac{1}{N_t}\right) Var\left[g(X_{t-1})\right]. \tag{29}$$

In the following, we treat the general linear case

$$g(x) = (1-a)x + b,$$

the mean and variance expression (28, 29) may be replaced by

$$\begin{aligned}
\mathbb{E}(X_t) = \mathbb{E}[g(X_{t-1})] &= (1-a)\mathbb{E}(X_{t-1}) + b \\
&= ... \\
&= (1-a)^t x_0 + b\sum_{i=0}^{t-1}(1-a)^i \\
&= (1-a)^t x_0 + \frac{b}{a}[1-(1-a)^t] \\
&= (1-a)^t\left(x_0 - \frac{b}{a}\right) + \frac{b}{a},
\end{aligned} \tag{30}$$

$$\begin{aligned}
Var(X_t) &= \frac{1}{N_t}\mathbb{E}(X_t)[1-\mathbb{E}(X_t)] + \left(1-\frac{1}{N_t}\right)(1-a)^2 Var(X_{t-1}) \\
&= \sum_{i=1}^{t}\frac{1}{N_i}(1-a)^{2(t-i)}\mathbb{E}(X_i)[1-\mathbb{E}(X_i)]\prod_{j=i+1}^{t}\left(1-\frac{1}{N_j}\right).
\end{aligned} \tag{31}$$

To consider approximations resulting from the infinite population limit, we take some appropriate variable transformations,

$$u = \frac{t}{N_0}, \quad r = \frac{i}{N_0}, \quad s = \frac{j}{N_0}, \tag{32}$$

and

$$h(u) = h(\frac{t}{N_0}) = \frac{N_t}{N_0}, \tag{33}$$

where $u, r, s \in \mathbb{R}_+,\ i, j = 1, \ldots, t \in \mathbb{N}$ and $h \in L^1(\mathbb{R}_+)$. Using the Riemann sum and Taylor approximation, let N_0 be large enough, then

$$\begin{aligned}
\prod_{j=i+1}^{t}\left(1-\frac{1}{N_j}\right) &= \exp\left[\log\prod_{j=i+1}^{t}\left(1-\frac{1}{N_j}\right)\right] \\
&= \exp\left[\sum_{j=i+1}^{t}\log\left(1-\frac{1}{N_j}\right)\right] \\
&= \exp\left[\sum_{j=i+1}^{t}\log\left(1-\frac{1}{N_0 h(s)}\right)\right] \\
&= \exp\left[N_0\sum_{j=i+1}^{t}\frac{1}{N_0}\log\left(1-\frac{1}{N_0 h(s)}\right)\right] \\
&= \exp\left[N_0\int_{\frac{i+1}{N_0}}^{\frac{t}{N_0}}\log\left(1-\frac{1}{N_0 h(s)}\right)ds\right] \\
&= \exp\left(-\int_r^u \frac{1}{h(s)}ds\right) + o(1),
\end{aligned} \tag{34}$$

where $o(1)$ is an infinitesimally small quantity, which is negligible in the limit. Using previous notations $\alpha = N_0 a$ and $\beta = N_0 b$, we have

$$(1-a)^t = \left(1-\frac{\alpha}{N_0}\right)^t = \left(1-\frac{\alpha}{N_0}\right)^{-\frac{N_0}{\alpha}\left(-\frac{t}{N_0}\right)\alpha} = e^{-\alpha u} + o(1).$$

Defining $Y_u = \lim_{N_0\to\infty} X_{\lfloor uN_0\rfloor}$ (Ewens, 2004), using (30),(31) and the Riemann sum, we obtain the mean and variance as a function of the scaled time

$$\mathbb{E}(Y_u) = e^{-\alpha u}\left(x_0-\frac{\beta}{\alpha}\right)+\frac{\beta}{\alpha}, \tag{35}$$

$$\begin{aligned}
Var(Y_u) &= e^{-2\alpha u}\int_0^u \frac{1}{h(r)}e^{2\alpha r}\mathbb{E}(Y_r)[1-\mathbb{E}(Y_r)]\exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr \\
&= e^{-2\alpha u}\int_0^u \frac{1}{h(r)}e^{2\alpha r}\left[e^{-\alpha r}\left(x_0-\frac{\beta}{\alpha}\right)+\frac{\beta}{\alpha}\right]\left[1-e^{-\alpha r}\left(x_0-\frac{\beta}{\alpha}\right)-\frac{\beta}{\alpha}\right]\exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr \\
&= e^{-2\alpha u}\int_0^u \frac{1}{h(r)}e^{2\alpha r}\left[-e^{-2\alpha r}\left(x_0-\frac{\beta}{\alpha}\right)^2\right]\exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr \\
&\quad + e^{-2\alpha u}\int_0^u \frac{1}{h(r)}e^{2\alpha r}\left[e^{-\alpha r}\left(x_0-\frac{\beta}{\alpha}\right)\left(1-\frac{2\beta}{\alpha}\right)\right]\exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr \\
&\quad + e^{-2\alpha u}\int_0^u \frac{1}{h(r)}e^{2\alpha r}\left[\frac{\beta}{\alpha}\left(1-\frac{\beta}{\alpha}\right)\right]\exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr \\
&= -e^{-2\alpha u}\left(x_0-\frac{\beta}{\alpha}\right)^2\left[1-exp\left(-\int_0^u\frac{1}{h(s)}ds\right)\right] \\
&\quad + e^{-2\alpha u}\left(x_0-\frac{\beta}{\alpha}\right)\left(1-\frac{2\beta}{\alpha}\right)\left[e^{\alpha u}-exp\left(-\int_0^u\frac{1}{h(s)}ds\right)-\int_0^u \alpha e^{\alpha r}exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr\right] \\
&\quad + e^{-2\alpha u}\frac{\beta}{\alpha}\left(1-\frac{\beta}{\alpha}\right)\left[e^{2\alpha u}-exp\left(-\int_0^u\frac{1}{h(s)}ds\right)-\int_0^u 2\alpha e^{2\alpha r}exp\left(-\int_r^u\frac{1}{h(s)}ds\right)dr\right],
\end{aligned} \tag{36}$$

where, the last step in (36) was obtained using integration by parts.

A.1 Proof of Theorem 1

Proof. Following expression (11), we consider the first derivative of F_2 with respect to t,

$$\frac{\partial F_2}{\partial t} := \Delta_1 \left(1-\frac{\alpha}{N}\right)^t - \Delta_2 \left[\left(1-\frac{\alpha}{N}\right)^2 \left(1-\frac{1}{N}\right)\right]^t,$$

where

$$\Delta_1 = (x_0-\gamma)\left[\frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)} + 2(\gamma - x_0)\right] \ln\left(1-\frac{\alpha}{N}\right), \tag{37}$$

and

$$\Delta_2 = \left[\frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} + \frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} - (\gamma-x_0)^2\right] \ln\left[\left(1-\frac{\alpha}{N}\right)^2\left(1-\frac{1}{N}\right)\right]. \tag{38}$$

We want to observe where the inflection point of F_2 occurs, so the following analysis is introduced,

$$\begin{aligned} \frac{\partial F_2}{\partial t} &< 0 \\ \Delta_1 \left(1-\frac{\alpha}{N}\right)^t &< \Delta_2 \left[\left(1-\frac{\alpha}{N}\right)^2\left(1-\frac{1}{N}\right)\right]^t \\ \Delta_1 &< \Delta_2 \left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t. \end{aligned} \tag{39}$$

When $\Delta_2 = 0$, the last step in (39) indicates $\Delta_1 < 0$. However, by the definition (38),

$$\begin{aligned} \frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} + \frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} - (\gamma-x_0)^2 &= 0 \\ \frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} + \frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} &= (\gamma-x_0)^2 \\ \frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} &\leq 2(\gamma-x_0)^2. \end{aligned}$$

The last step was obtained using the following,

$$\frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} \geq 0. \tag{40}$$

Hence, the definition (37) indicates $\Delta_1 \geq 0$, that is a contradiction.

When $\Delta_2 \neq 0$, (39) indicates

$$\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t > \frac{\Delta_1}{\Delta_2}, \quad \Delta_2 > 0, \tag{41}$$

or

$$\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t < \frac{\Delta_1}{\Delta_2}, \quad \Delta_2 < 0. \tag{42}$$

For (41), if $\Delta_1 \leq 0$, by (40) and the definition (37),

$$\begin{aligned} &\frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} - 2(\gamma-x_0)^2 \geq 0, \\ &\frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} \geq 2(\gamma-x_0)^2 \geq (\gamma-x_0)^2, \end{aligned} \tag{43}$$

we get $\Delta_2 \leq 0$, this contradicts $\Delta_2 > 0$. If $\Delta_1 > 0$ and $0 < \Delta_2 < \Delta_1$, then $1 < \Delta_1/\Delta_2$. However,

$$\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t < 1,$$

it contradicts (41). So, there may be an inflection point when $0 < \Delta_1 < \Delta_2$, which makes F_2 have a decreasing trend. In this case, using (41), we can get the range

$$t < \frac{\ln\frac{\Delta_1}{\Delta_2}}{\ln\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]} := \hat{t}. \tag{44}$$

We know when $t = 0$ then $F_2 = 0$ and $F_2 \geq 0$ for all t. $\hat{t}$ all depends on N, α, γ, x_0 and $\hat{t} > 0$ then $\forall t \in [0, \hat{t}\,)$, by (41) $\partial F_2/\partial t < 0$, $F_2\left(X_0, X_t\right) < F_2\left(X_0, X_0\right) = 0$, that is a contradiction.

From the above analysis, the existence of the inflection point of F_2 means that (42) holds. For some $t > 0$, we prove

$$(42)\ \textit{holds} \iff \Delta_1 < 0. \tag{45}$$

1) "$\Longrightarrow$"

Consider proof by contradiction. Clearly, if $\Delta_1 \geq 0$ then we get

$$\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t > 0 \geq \frac{\Delta_1}{\Delta_2}.$$

2) "$\Longleftarrow$"

Clearly,

$$\ln\left[\left(1-\frac{\alpha}{N}\right)^2\left(1-\frac{1}{N}\right)\right] < 2\ln\left(1-\frac{\alpha}{N}\right) < \ln\left(1-\frac{\alpha}{N}\right) < 0. \tag{46}$$

Using $\Delta_1 < 0$ and the definition (37), then (43) and (40) are satisfied. We obtain

$$\begin{aligned} 0 < (x_0-\gamma)&\left[\frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)} + 2(\gamma-x_0)\right] < \\ &\left[\frac{\gamma(1-\gamma)}{1+\alpha\left(2-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)} + \frac{(1-2\gamma)(x_0-\gamma)}{1+\alpha\left(1-\frac{1}{N}\right)} - (\gamma-x_0)^2\right], \end{aligned} \tag{47}$$

Combining (46) and (47),

$$\Delta_2 < (x_0-\gamma)\left[\frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)}+2(\gamma-x_0)\right]\ln\left[\left(1-\frac{\alpha}{N}\right)^2\left(1-\frac{1}{N}\right)\right] < \Delta_1 < 0. \quad (48)$$

(48) indicates $\hat{t} > 0$ and for all $t > \hat{t}$, we have

$$\begin{aligned} t &> \frac{\ln\frac{\Delta_1}{\Delta_2}}{\ln\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]} \\ \ln\left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t &< \ln\frac{\Delta_1}{\Delta_2} \\ \left[\left(1-\frac{\alpha}{N}\right)\left(1-\frac{1}{N}\right)\right]^t &< \frac{\Delta_1}{\Delta_2}, \end{aligned} \quad (49)$$

(42) holds.

All the above proof take into account that F_2 is a continuous function of t, and even though we only use $t \in \mathbb{N}$, the conclusion still holds. □

A.2 Proof of Theorem 2

Proof. Following expression (12), we consider the first derivative of $D_2(u)$ with respect to u,

$$\begin{aligned} \frac{\partial D_2(u)}{\partial u} &= \left[2\alpha(x_0-\gamma)^2-\frac{\alpha}{\alpha+1}(x_0-\gamma)(1-2\gamma)\right]e^{-\alpha u} \\ &\quad -\left[(2\alpha+1)(x_0-\gamma)^2-\frac{2\alpha+1}{\alpha+1}(x_0-\gamma)(1-2\gamma)-\gamma(1-\gamma)\right]e^{-(2\alpha+1)u} \\ &:= \Theta_1 e^{-\alpha u}-\Theta_2 e^{-(2\alpha+1)u}, \end{aligned}$$

where,

$$\Theta_1 = (x_0-\gamma)\frac{\alpha}{\alpha+1}(2x_0\alpha-2\gamma\alpha+2x_0-1),$$

and

$$\Theta_2 = \Theta_1 - x_0(1-x_0).$$

Refer to the proof of Theorem 1 (A.1), the following analysis is introduced,

$$\begin{aligned} \frac{\partial D_2(u)}{\partial u} &< 0 \\ \Theta_1 e^{-\alpha u}-\Theta_2 e^{-(2\alpha+1)u} &< 0 \\ \Theta_1 e^{-\alpha u} &< \Theta_2 e^{-(2\alpha+1)u} \\ \Theta_1 &< \Theta_2 e^{-(\alpha+1)u}. \end{aligned} \quad (50)$$

When $\Theta_2 = 0$, then $\Theta_1 = x_0(1-x_0) \geq 0$, that contradicts the last step indicating $\Theta_1 < 0$ in (50). When $\Theta_2 \neq 0$, then (50) can be further transformed into

$$e^{-(\alpha+1)u} > \frac{\Theta_1}{\Theta_2}, \quad \Theta_2 > 0, \quad (51)$$

or

$$e^{-(\alpha+1)u} < \frac{\Theta_1}{\Theta_2}, \quad \Theta_2 < 0. \tag{52}$$

If $\Theta_2 > 0$ then $\Theta_1 > \Theta_2 > 0$, we have $e^{-(\alpha+1)u} < 1 < \Theta_1/\Theta_2$. For (51), $D_2(u)$ cannot have an inflection point in its decreasing trend. If $\Theta_2 < 0$, $\Theta_1 > 0$, then $\Theta_1/\Theta_2 < 0$, that is not possible; if $\Theta_1, \Theta_2 < 0$ and obviously, $\Theta_2 < \Theta_1 < 0$ indicates $\Theta_2/\Theta_1 > 1$, that is possible. For (52), we can get the range,

$$u > \frac{1}{\alpha+1} \ln \frac{\Theta_2}{\Theta_1} := \hat{u} > 0, \quad \Theta_1 < 0. \tag{53}$$

The process details are similar to the proof of Theorem 1 (A.1) and are partially omitted here. □

A.3 Proof of Corollary

Proof. According to Theorem 1 and 2, the necessary and sufficient condition for the existence of the inflection point is $\Delta_1 < 0$ and $\Theta_1 < 0$, respectively. To prove the corollary in 2.2.2, we introduce the following analysis.

For Δ_1 or Θ_1, we set α, N, γ as constants, then Δ_1 or Θ_1 is a parabolic function of x_0. And for a parabolic function, the fact that existence of the roots depends on the sign of the discriminant, which is exactly that the discriminant $\Delta \geq 0$. In addition, since $x_0 \in [0, 1]$, we consider the range of the roots to complete the proof.

For Δ_1, we know

$$\Delta_1 \propto -\Delta_{11}\Delta_{12},$$

where, $\Delta_{11} = x_0 - \gamma$ and

$$\Delta_{12} = \frac{1-2\gamma}{1+\alpha\left(1-\frac{1}{N}\right)} + 2(\gamma - x_0).$$

Then

$$\Delta_1 < 0 \Longleftrightarrow \Delta_{11}\Delta_{12} > 0.$$

Defining $\Delta^* = \Delta_{11}\Delta_{12}$, obviously,

$$\begin{aligned}\Delta^* &= \frac{1}{1+\alpha\left(1-\frac{1}{N}\right)} \left\{ (x_0-\gamma)(1-2\gamma) - 2(x_0-\gamma)^2 \left[1+\alpha\left(1-\frac{1}{N}\right)\right] \right\} \\ &:\propto -2\left[1+\alpha\left(1-\frac{1}{N}\right)\right] x_0^2 + \left[2\gamma+1+4\gamma\alpha\left(1-\frac{1}{N}\right)\right] x_0 - \left[\gamma+2\gamma^2\alpha\left(1-\frac{1}{N}\right)\right].\end{aligned} \tag{54}$$

Using the last expression in (54), we calculate the discriminant

$$\begin{aligned}\Delta &= \left[2\gamma+1+4\gamma\alpha\left(1-\frac{1}{N}\right)\right]^2 - 8\left[1+\alpha\left(1-\frac{1}{N}\right)\right]\left[\gamma+2\gamma^2\alpha\left(1-\frac{1}{N}\right)\right] \\ &= (2\gamma-1)^2.\end{aligned} \tag{55}$$

For $\forall \alpha \geq 0, N \in \mathbb{N}, \gamma \in [0,1]$ and $\gamma \neq 0.5$, $\Delta > 0$. And the expressions for the two roots are

$$\gamma, \quad \gamma + \frac{1-2\gamma}{2\left[1+\alpha\left(1-\frac{1}{N}\right)\right]} \in [0,1].$$

Hence, for a downward opening parabola, $\forall x_0 \in (x_{0L}, x_{0R})$, where

$$x_{0L} = \min\left(\gamma, \gamma + \frac{1-2\gamma}{2\left[1+\alpha\left(1-\frac{1}{N}\right)\right]}\right), \quad x_{0R} = \max\left(\gamma, \gamma + \frac{1-2\gamma}{2\left[1+\alpha\left(1-\frac{1}{N}\right)\right]}\right),$$

then $\Delta^* > 0$ and $\Delta_1 < 0$. According Theorem 1, $F_2(t)$ has an inflection point.

For Θ_1, the process details are similar to the above steps and are omitted here. □

Appendix B The derivative and limit

For the $D_{st}(u)$ in 2.2.3, the inflection point is sought by the derivative and limit. We describe the calculations used in the formation of ideas as follows.

B.1 Pure drift for P_1, the linear pressure model for P_2

We have

$$D_{\mathrm{st}}(u) = \frac{Var(Y_{u,1}) + D_2\left(x_0, Y_{u,2}\right)}{x_0 + \mathbb{E}(Y_{u,2}) - 2x_0\mathbb{E}(Y_{u,2})},$$

where

$$\mathbb{E}(Y_{u,2}) = e^{-\alpha u}\left(x_0 - \gamma\right) + \gamma,$$

$$D_2\left(x_0, Y_{u,2}\right) = -\frac{\Theta_1}{\alpha}e^{-\alpha u} + \frac{\Theta_2}{2\alpha+1}e^{-(2\alpha+1)u} + (x_0-\gamma)^2 + \frac{\gamma(1-\gamma)}{2\alpha+1},$$

and

$$Var(Y_{u,1}) = x_0(1-x_0)\left(1-e^{-u}\right).$$

When two non-negative functions f_1 and f_2 are differentiable, the chain rule says,

$$\left(\frac{f_1}{f_2}\right)' = \frac{f_1' f_2 - f_1 f_2'}{f_2^2},$$

if ${f_1}' < 0$ and ${f_2}' > 0$, then $({f_1}/{f_2})' < 0$. We set

$$f_1 = Var(Y_{u,1}) + D_2\left(x_0, Y_{u,2}\right),$$

$$f_2 = (1-2x_0)\mathbb{E}(Y_{u,2}) + x_0.$$

Using the above expressions, we get

$${f_2}' = -\alpha(1-2x_0)(x_0-\gamma) > 0 \iff (1-2x_0)(x_0-\gamma) < 0,$$

and

$$\begin{aligned} {f_1}' = \frac{df_1}{du} &= x_0(1-x_0)e^{-u} + \Theta_1 e^{-\alpha u} - \Theta_2 e^{-(2\alpha+1)u} \\ &= (\Theta_1 - \Theta_2)e^{-u} + \Theta_1 e^{-\alpha u} - \Theta_2 e^{-(2\alpha+1)u} \\ &= \Theta_1 e^{-u}\left(1 + e^{(1-\alpha)u}\right) - \Theta_2 e^{-u}(1 + e^{-2\alpha u}), \end{aligned}$$

then

$$f_1{}' < 0 \iff \Theta_1\left(1+e^{(1-\alpha)u}\right) < \Theta_2(1+e^{-2\alpha u}),$$

so,

$$\frac{1+e^{-2\alpha u}}{1+e^{(1-\alpha)u}} > \frac{\Theta_1}{\Theta_2}, \quad \Theta_2 > 0, \tag{56}$$

or

$$\frac{1+e^{-2\alpha u}}{1+e^{(1-\alpha)u}} < \frac{\Theta_1}{\Theta_2}, \quad \Theta_2 < 0. \tag{57}$$

But from the previous analysis, we know that $\Theta_2 < \Theta_1$ and $0 < e^{-2\alpha u} < e^{(1-\alpha)u}$, (56) is not possible; $(1-2x_0)(x_0-\gamma) < 0$ and $\Theta_1 < 0$ also cannot be held together in the areas delineated by the Figure 4, (57) is also undesirable. The above preliminary judgment inspires us to consider

$$\begin{aligned} f_1' f_2 - f_1 f_2' =& e^{-u} x_0 (1-x_0) [x_0 + \gamma (1-2x_0)] \\ &+ e^{-\alpha u} \{\Theta_1 [x_0 + \gamma (1-2x_0)] + \alpha (1-2x_0)(x_0-\gamma) x_0 (1-x_0) \\ &\quad + \alpha (1-2x_0)(x_0-\gamma)^3 + \frac{\alpha}{2\alpha+1}(1-2x_0)(x_0-\gamma)\gamma(1-\gamma)\Big\} \\ &+ e^{-(\alpha+1)u} x_0 (1-x_0)(1-2x_0)(x_0-\gamma)(1-\alpha) \\ &- e^{-(2\alpha+1)u}\Theta_2 [x_0 + \gamma (1-2x_0)] \\ &- e^{-(3\alpha+1)u}\Theta_2 (1-2x_0)(x_0-\gamma)\frac{\alpha+1}{2\alpha+1}. \end{aligned} \tag{58}$$

Letting $e^{-u} = \lambda$, then (58) can be transformed into

$$\begin{aligned} f_1' f_2 - f_1 f_2' =& \lambda x_0 (1-x_0) [x_0 + \gamma (1-2x_0)] \\ &+ \lambda^{\alpha} \{\Theta_1 [x_0 + \gamma (1-2x_0)] + \alpha (1-2x_0)(x_0-\gamma) x_0 (1-x_0) \\ &\quad + \alpha (1-2x_0)(x_0-\gamma)^3 + \frac{\alpha}{2\alpha+1}(1-2x_0)(x_0-\gamma)\gamma(1-\gamma)\Big\} \\ &+ \lambda^{(\alpha+1)} x_0 (1-x_0)(1-2x_0)(x_0-\gamma)(1-\alpha) \\ &- \lambda^{(2\alpha+1)}\Theta_2 [x_0 + \gamma (1-2x_0)] \\ &- \lambda^{(3\alpha+1)}\Theta_2 (1-2x_0)(x_0-\gamma)\frac{\alpha+1}{2\alpha+1}. \end{aligned} \tag{59}$$

Using (59), we introduce the following limit.

1) If $\alpha > 1, u \to \infty$, then

$$f_1' f_2 - f_1 f_2' = x_0 (1-x_0) [x_0 + \gamma (1-2x_0)] \lambda + o(\lambda).$$

2) If $\alpha = 1, u \to \infty$, then

$$\begin{aligned} f_0' f_2 - f_1 f_2' =& \{2x_0^2 (1-x_0)^2 + \Theta_1 [x_0 + \gamma (1-2x_0)] + (1-2x_0)(x_0-\gamma)^3 \\ &+ \frac{1}{3}(1-2x_0)(x_0-\gamma)\gamma(1-\gamma)\Big\} \lambda + o(\lambda). \end{aligned}$$

3) If $\alpha < 1, u \to \infty$, then

$$\begin{aligned} f_1' f_2 - f_1 f_0' =& \{\Theta_1 [x_0 + \gamma (1-2x_0)] + \alpha (1-2x_0)(x_0-\gamma) x_0 (1-x_0) \\ &+ \alpha (1-2x_0)(x_0-\gamma)^3 + \frac{\alpha}{2\alpha+1}(1-2x_0)(x_0-\gamma)\gamma(1-\gamma)\Big\} \lambda^{\alpha} + o(\lambda^{\alpha}). \end{aligned}$$

Obviously, in the **1)** $f_1'f_2 - f_1f_0' \geq 0$, and $f_1'f_2 - f_1f_0'$ is more likely to be negative in the **3)** than in the **2)**. We consider the **3)** limiting form as a case and give a result under $\alpha < 1$ in the main text.

B.2 The same linear pressure model for P_1 and P_2

We consider

$$D_{st}(u) = \frac{D_2\left(x_0, Y_{u,1}\right) + D_2\left(x_0, Y_{u,2}\right) - 2\mathbb{E}\left(Y_{u,1} - x_0\right)\mathbb{E}\left(Y_{u,2} - x_0\right)}{\mathbb{E}\left(Y_{u,1} + Y_{u,2}\right) - 2\mathbb{E}(Y_{u,1})\mathbb{E}(Y_{u,2})},$$

and the same linear evolutionary pressure model for P_1 and P_2, then $D_{st}(u)$ can be simplified as

$$D_{st}(u) = \frac{D_2\left(x_0, Y_u\right) - \left(\mathbb{E}(Y_u) - x_0\right)^2}{\mathbb{E}(Y_u)\left(1 - \mathbb{E}(Y_u)\right)},$$

where,

$$\mathbb{E}(Y_u) = e^{-\alpha u}\left(x_0 - \gamma\right) + \gamma,$$

$$D_2\left(x_0, Y_u\right) = -\frac{\Theta_1}{\alpha}e^{-\alpha u} + \frac{\Theta_2}{2\alpha+1}e^{-(2\alpha+1)u} + \left(x_0 - \gamma\right)^2 + \frac{\gamma(1-\gamma)}{2\alpha+1}.$$

We set

$$f_1 = D_2\left(x_0, Y_u\right) - \left(\mathbb{E}(Y_u) - x_0\right)^2,$$
$$f_2 = \mathbb{E}(Y_u)\left(1 - \mathbb{E}(Y_u)\right),$$

then in details,

$$\begin{aligned}
f_1 =& \left(x_0 - \gamma\right)^2\left[e^{-(2\alpha+1)u} - e^{-2\alpha u}\right] \\
&+ \frac{1}{\alpha+1}\left(x_0 - \gamma\right)(1 - 2\gamma)\left[e^{-\alpha u} - e^{-(2\alpha+1)u}\right] \\
&+ \frac{1}{2\alpha+1}\gamma(1-\gamma)\left[1 - e^{-(2\alpha+1)u}\right] \\
=& e^{-\alpha u}\cdot\frac{1}{\alpha+1}\left(x_0 - \gamma\right)(1 - 2\gamma) \\
&- e^{-2\alpha u}\left(x_0 - \gamma\right)^2 \\
&+ \frac{1}{2\alpha+1}\gamma(1-\gamma) \\
&+ e^{-(2\alpha+1)u}\cdot\frac{\Theta_2}{2\alpha+1}(\text{using}(13),(14)), \\
f_2 =& e^{-\alpha u}\left(x_0 - \gamma\right)(1 - 2\gamma) - e^{-2\alpha u}\left(x_0 - \gamma\right)^2 + \gamma(1-\gamma).
\end{aligned}$$

We calculate two derivatives,

$$\begin{aligned}
f_1' =& e^{-\alpha u}\frac{-\alpha}{\alpha+1}\left(x_0 - \gamma\right)(1 - 2\gamma) + e^{-2\alpha u}\cdot 2\alpha\left(x_0 - \gamma\right)^2 + e^{-(2\alpha+1)u}\left(-\Theta_2\right) \\
f_2' =& e^{-\alpha u}(-\alpha)\left(x_0 - \gamma\right)(1 - 2\gamma) + e^{-2\alpha u}\cdot 2\alpha\left(x_0 - \gamma\right)^2.
\end{aligned}$$

Combining f_1' and f_2', we obtain

$$\begin{aligned} f_1'f_2 - f_1f_2' =& e^{-\alpha u}\alpha\left(x_0-\gamma\right)\left(1-2\gamma\right)\gamma(1-\gamma)\left(\frac{1}{2\alpha+1}-\frac{1}{\alpha+1}\right) \\ &+ e^{-2\alpha u}\left(x_0-\gamma\right)^2\gamma(1-\gamma)\frac{4\alpha^2}{2\alpha+1} \\ &+ e^{-(2\alpha+1)u}\left(-\theta_2\right)\gamma(1-\gamma) \\ &+ e^{-3\alpha u}\left(x_0-\gamma\right)^3\left(1-2\gamma\right)\left(\alpha-\frac{\alpha}{\alpha+1}\right) \\ &+ e^{-(3\alpha+1)u}\theta_2\left(x_0-\gamma\right)\left(1-2\gamma\right)\left(\frac{\alpha}{2\alpha+1}-1\right) \\ &+ e^{-(4\alpha+1)u}\theta_2\left(x_0-\gamma\right)^2\left(1-\frac{2\alpha}{2\alpha+1}\right). \end{aligned} \tag{60}$$

Letting $e^{-u}=\lambda$ and $u\to\infty$, then (60) can be transformed into

$$\frac{f_1'f_2-f_1f_2'}{\lambda^\alpha} = \left(x_0-\gamma\right)\left(1-2\gamma\right)\gamma(1-\gamma)\frac{-\alpha^2}{(2\alpha+1)(\alpha+1)}+o(1).$$

Obviously, we only need to consider x_0 and γ, s.t. $\left(x_0-\gamma\right)\left(1-2\gamma\right)>0$ and give a case in the main text.

Chapter 3

Manuscript 2

Evaluation of population structure inferred by principal component analysis or the admixture model

Jan van Waaij[1,3], Song Li[1], Genís Garcia-Erill[2], Anders Albrechtsen[2] and Carsten Wiuf[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark.

[2]Department of Biology, University of Copenhagen, Denmark.

[3]Department of Health Technology, Danish Technical University, Denmark.

Publication details: Submitted for publication in *Genetics*.

OXFORD GENETICS

Evaluation of population structure inferred by principal component analysis or the admixture model

Jan van Waaij[1,3,†], Song Li[1,†], Genís Garcia-Erill[2], Anders Albrechtsen[2] and Carsten Wiuf[1,*]

[1]Department of Mathematical Science, University of Copenhagen, 2100 Copenhagen, Denmark
[2]Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[3]Current address: Department of Health Technology, Danish Technical University, 2800 Kgs. Lyngby, Denmark
[†]These authors contributed equally to this work

[*]Corresponding author: Department of Mathematical Sciences, Universitetsparken 5, 2100 Copenhagen, Denmark. Email: wiuf@math.ku.dk

Abstract

Principal component analysis (PCA) is commonly used in genetics to infer and visualize population structure and admixture between populations. PCA is often interpreted in a way similar to inferred admixture proportions, where it is assumed that individuals belong to one of several possible populations or are admixed between these populations. We propose a new method to assess the statistical fit of PCA (interpreted as a model spanned by the top principal components) and to show that violations of the PCA assumptions affect the fit. Our method uses the chosen top principal components to predict the genotypes. By assessing the covariance (and the correlation) of the residuals (the differences between observed and predicted genotypes), we are able to detect violation of the model assumptions. Based on simulations and genome wide human data we show that our assessment of fit can be used to guide the interpretation of the data and to pinpoint individuals that are not well represented by the chosen principal components. Our method works equally on other similar models, such as the admixture model, where the mean of the data is represented by linear matrix decomposition.

Keywords: PCA; residuals; population modelling; ancient DNA; statistical fit

Introduction

Principal component analysis (PCA) and model-based clustering methods are popular ways to disentangle the ancestral genetic history of individuals and populations. One particular model, the admixture model (Pritchard *et al.* 2000), has played a prominent role because of its simple structure and, in some cases, easy interpretability. PCA is often seen as being model free but as noted by Engelhardt and Stephens (2010), the two approaches are very similar. The interpretation of the results of a PCA analysis is often based on assumptions similar to those of the admixture model, such that admixed individuals are linear combinations of the eigenvectors representing unadmixed individuals. In this way, the admixed individuals lie in-between the unadmixed individuals in a PCA plot. As shown for the admixture model, there are many demographic histories that can lead to the same result (Lawson *et al.* 2018a) and many demographic histories that violate the assumptions of the admixture model (Garcia-Erill and Albrechtsen 2020). As we will show, this is also the case for PCA, since it has a similar underlying model (Engelhardt and Stephens 2010).

The admixture model states that the genetic material from each individual is composed of contributions from k distinct ancestral homogeneous populations. However, this is often contested in real data analysis, where the ancestral population structure might be much more complicated than that specified by the admixture model. For example, the k ancestral populations might be heterogeneous themselves, the exact number of ancestral populations might be difficult to assess due to many smaller contributing populations, or the genetic composition of an individual might be the result of continuous migration or recent backcrossing, which also violates the assumptions of the admixture model. Furthermore, the admixture model assumes individuals are unrelated, which naturally might not be the case. This paper is concerned with assessing the fit of PCA building on the special relationship with the admixture model (Engelhardt and Stephens 2010). In particular, we are interested in quantifying the model fit and assessing the validity of the model at the level of the sample as well as at the level of the individual. Using real and simulated data we show that the fit from a PCA analysis is affected by violations of the admixture model.

We consider genotype data G from n individuals and m SNPs, such that $G_{si} \in \{0,1,2\}$ is the number of reference alleles for individual i and SNP s. Typically, G_{si} is assumed to be binomially distributed with parameter Π_{si}, where Π_{si} depends on the number of ancestral populations, k, their admixture proportions and the ancestral population allele frequencies. For clustering based analysis such as ADMIXTURE (Alexander and Lange 2011), k is the number of clusters while in PCA, it is the $k-1$ top principal components. We give the specifics of the admixture model in the next section and show its relationship to PCA in the Material and methods section.

Several methods aim to estimate the best k in some sense (Alexander and Lange 2011; Evanno *et al.* 2005; Pritchard *et al.* 2000; Raj *et al.* 2014; Wang 2019), but finding such k does not imply the data fit the model (Lawson *et al.* 2018b; Janes *et al.* 2017). In statistics, it is standard to use residuals and distributional summaries of the residuals to assess model fit (Box *et al.* 2005). The residual of an observation is defined as the difference between the observed and the predicted value (estimated under some model). Visual trends in the residuals (for example, differences between populations) are indicative of model

misfit, and large absolute values of the residuals are indicative of outliers (for example due to experimental errors, or kinship). If the model is correct, a histogram of the residuals is expected to be monomodal centered around zero (Box *et al.* 2005).

In our context, Garcia-Erill and Albrechtsen (2020) argue that trends in the residual correlation matrix carries information about the underlying model and might be used for visual model evaluation. A method is designed to assess whether the correlation structure agrees with the proposed model, in particular, whether it agrees with the proposed number of homogeneous ancestral populations (Garcia-Erill and Albrechtsen 2020). However, even in the case the model is correctly specified, the residuals are in general correlated (Box *et al.* 2005), and therefore, trends might be observed even if the model is true, leading to incorrect model assessment. To adjust for this correlation, a leave-one-out procedure, based on maximum likelihood estimation of the admixture model parameters, is developed that removes the correlation between residuals in the case the model is correct, but not if the model is misspecified (Garcia-Erill and Albrechtsen 2020). This approach could also be applied to PCA, where expected genotypes could be calculated using probabilistic PCA (Meisner *et al.* 2021). This leave-one-out procedure is, however, computationally expensive.

To remedy the computational difficulties, we take a different approach to investigate the correlation structure. We suggest two different ways of calculating the correlation matrix of the residuals. The first is simply the empirical correlation matrix of the residuals. The second might be considered an estimated correlation matrix, based on a model. Both are simple to compute. Under mild regularity assumptions, these two measures agree if the model is correct and the number of SNPs is large. Hence, their difference is expected to be close to zero, when the admixture model is not violated. If the difference is considerably different from zero, then this is proof of model misfit.

To explore the adequacy of the proposed method, we investigate different ways to calculate the predicted values of the genotype (hence, the residuals), using Principal Component Analysis (PCA) in different ways. However, we also show that this approach can be used on estimated admixture proportions. Specifically, we use 1) an uncommon but very useful PCA approach (here, named PCA 1) based on unnormalized genotypes (Cabreros and Storey 2019; Chen and Storey 2015), 2) PCA applied to mean centred data (PCA 2), see Patterson *et al.* (2006), and 3) PCA applied to mean and variance normalised data (PCA 3) (Patterson *et al.* 2006). All three approaches are computationally fast and do not require separate estimation of ancestral allele frequencies and population proportions, as in Garcia-Erill and Albrechtsen (2020). Hence, the computation of the residuals are computationally inexpensive. Additionally, we show that this approach can also be applied to output from, for example, the software ADMIXTURE (Alexander *et al.* 2009) to estimate Π_{si} for each s and i, and to calculate the residuals from these estimates. An overview of PCA can be found in Jolliffe and Cadima (2016).

We demonstrate that our proposed method works well on simulated and real data, when the predicted values (and the residuals) are calculated in any of the four mentioned ways. Furthermore, we back this up mathematically by showing that the two correlation measures agree (if the number of SNPs is large) under the correct admixture model for PCA 1 and PCA 2. For the latter, a few additional assumptions are required. The estimated covariance (and correlation coefficient) under the proposed model might be seen as a correction term for population structure. Subtracting it from the empirical covariance, thus gives a covariance estimate with baseline zero under the correct model, independent of the population structure. It is natural to suspect that similar can be done in models with population structure and kinship, which we will pursue in a subsequent study.

In the next section, we describe the model, the statistical approach to compute the residuals, and how we evaluate model fit. In addition, we give mathematical statements that show how the method performs theoretically. In the 'Results' section, we provide analysis of simulated and real data, respectively. We end with a discussion. Mathematical proofs are collected in the appendix.

Materials and methods

Notation

For an $\ell_1 \times \ell_2$ matrix $A = (A_{ij})_{i,j}$, $A_{\star i}$ denotes the i-th column of A, $A_{i\star}$ the i-th row, A^T the transpose matrix, and $\operatorname{rank}(A)$ the rank. The Frobenius norm of a square $\ell \times \ell$ matrix A is

$$\|A\|_F = \sqrt{\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} A_{ij}^2}.$$

A square matrix A is an orthogonal projection if $A^2 = A$ and $A^T = A$. A symmetric matrix has n real eigenvalues (with multiplicity) and the eigenvectors can be chosen such that they are orthogonal to each other. If the matrix is positive (semi-)definite, then the eigenvalues are positive (non-negative).

For a random variable/vector/matrix X, its expectation is denoted $\mathbb{E}[X]$ (provided it exist). The variance of a random variable X is denoted $\operatorname{var}(X)$, and covariance between two random variables X, Y is denoted $\operatorname{cov}(X,Y)$ (provided they exist). Similarly, for a random vector $X = (X_1, \ldots, X_n)$, the covariance matrix is denoted $\operatorname{cov}(X)$. For a sequence X_m, $m = 0, \ldots,$ of random variables/vectors/matrices, if $X_m \to X_0$ as $m \to \infty$ almost surely (convergence for all realisations but a set of zero probability), we leave out 'almost surely' and write $X_m \to X_0$ as $m \to \infty$ for convenience.

The PCA and the admixture model

We consider a model with genotype observations from n individuals, and m biallelic sites (SNPs), where m is assumed to be (much) larger than n, $m \geq n$. The genotype G_{si} of SNP s in individual i is assumed to be a binomial random variable

$$G_{si} \sim \text{binomial}(2, \Pi_{si}).$$

In matrix notation, we have $G \sim \text{binomial}(2, \Pi)$ with expectation $\mathbb{E}(G \mid \Pi) = 2\Pi$, where G and Π are $m \times n$ dimensional matrices. Conditional on Π, we assume the entries of G are independent random variables.

Furthermore, we assume the matrix Π takes the form $\Pi = FQ$, where Q is a (possibly unconstrained) $k \times n$ matrix of rank $k \leq n$, and F is a (possibly unconstrained) $m \times k$ matrix, also of rank k (implying Π likewise is of rank k, Lemma 13). Entry-wise, this amounts to

$$\Pi_{si} = (FQ)_{si} = \sum_{k=1}^{k} F_{sk} Q_{ki}, \quad s = 1, \ldots, n, \quad i = 1, \ldots, m.$$

For the binomial assumption to make sense, we must require the entries of Π to be between zero and one.

In the literature, this model is typically encountered in the form of an admixture model with k ancestral populations, see for example, Pritchard *et al.* (2000); Garcia-Erill and Albrechtsen (2020). The general unconstrained setting which applies to PCA has also been discussed (Cabreros and Storey 2019). In the case of an admixture model, Q is a matrix of ancestral admixture proportions, such that the proportion of individual i's genome originating from population j is Q_{ji}. Furthermore, F is a matrix of ancestral SNP frequencies, such

that the frequency of the reference allele of SNP s in population j is F_{sj}. In many applications, the columns of Q sum to one.

While we lean towards an interpretation in terms of ancestral population proportions and SNP frequencies, our approach does not enforce or assume the columns of Q (the admixture proportions) to sum to one, but allow these to be unconstrained. This is advantageous for at least two reasons. First, a proposed model might only contain the major ancestral populations, leaving out older or lesser defined populations. Hence, the sum of ancestral proportions might be smaller than one. Secondly, when fitting a model with fewer ancestral populations than the true model, one should only require the admixture proportions to sum to at most one.

The residuals

Our goal is to design a strategy to assess the hypothesis that Π is a product of two matrices. As we do not know the true k, we suggest a number k' of ancestral populations and estimate the model parameters under this constraint. That is, we assume a model of the form

$$G \sim \text{binomial}(2, \Pi_{k'}), \quad \Pi_{k'} = F_{k'} Q_{k'},$$

where each entry of G follows a binomial distribution. $Q_{k'}$ has dimension $k' \times n$, $F_{k'}$ has dimension $m \times k'$, and $\text{rank}(Q_{k'}) = \text{rank}(F_{k'}) = k'$, hence also $\text{rank}(\Pi_{k'}) = k'$. Throughout, we use the index k' to indicate the imposed rank condition, and assume $k' \leq k$ unless otherwise stated. The latter assumption is only to guarantee the mathematical validity of certain statements, and is not required for practical use of the method.

Our approach is build on the residuals, the difference between observed and predicted data. To define the residuals, we let $P\colon\ \mathbb{R}^n \to \mathbb{R}^n$ be the orthogonal projection onto the k-dimensional subspace spanned by the k rows of (the true) Q, hence $P = Q^T(QQ^T)^{-1}Q$, and $QP = Q$. Let $\widehat{P}_{k'}$ be an estimate of P based on the data G, and assume $\widehat{P}_{k'}$ is an orthogonal projection onto a k'-dimensional subspace. Later in this section, we show how an estimate $\widehat{P}_{k'}$ can be obtained from an estimate of $Q_{k'}$ or an estimate of $\Pi_{k'}$. Estimates of these parameters might be obtained using existing methods, based on for example, maximum likelihood analysis (Wang 2003; Alexander *et al.* 2009; Garcia-Erill and Albrechtsen 2020). Furthermore, for the three PCA approaches, an estimate of the projection matrix can simply be obtained from eigenvectors of a singular value decomposition (SVD) of the data matrix.

We define the $m \times n$ matrix of residuals by

$$R_{k'} = G - 2\widehat{\Pi} = G(I - \widehat{P}_{k'}),$$

where G is the observed data and $G\widehat{P}_{k'}$, the predicted values. The latter might also be considered an estimate of 2Π, the expected value of G. This definition of residuals is in line with how the residuals are defined in a multilinear regression model as the difference between the observed data (here, G) and the projection of the data onto the subspace spanned by the regressors (here, $G\widehat{P}_{k'}$). The essential difference being that in a multilinear regression model, the regressors are known and does not depend on the observed data, while $\widehat{P}_{k'}$ is estimated from the data.

We assess the model fit by studying the correlation matrix of the residuals in two ways. First, we consider the *empirical covariance matrix* $\widehat{B}$ with entries

$$\begin{aligned}\widehat{B}_{ij} &= \frac{1}{m-1}\sum_{s=1}^{m}(R_{k',si} - \bar{R}_{k',i})(R_{k',sj} - \bar{R}_{k',j}) \\ &= \frac{1}{m-1}\sum_{s=1}^{m}(R_{k',si}R_{k',sj} - \bar{R}_{k',i}\bar{R}_{k',j}),\end{aligned}$$

where

$$\bar{R}_{k'i} = \frac{1}{m}\sum_{s=1}^{m} R_{k',si},$$

and the corresponding *empirical correlation matrix* with entries

$$\widehat{b}_{ij} = \frac{\widehat{B}_{ij}}{\sqrt{\widehat{B}_{ii}\widehat{B}_{jj}}},$$

$i, j = 1, \ldots, n$. Secondly, we consider the *estimated covariance matrix*

$$\widehat{C} = (I - \widehat{P}_{k'})\widehat{D}(I - \widehat{P}_{k'})$$

with corresponding *estimated correlation matrix*,

$$\widehat{c}_{ij} = \frac{\widehat{C}_{ij}}{\sqrt{\widehat{C}_{ii}\widehat{C}_{jj}}},$$

$i, j = 1, \ldots, n$. Here, $\widehat{D}$ is the $n \times n$ diagonal matrix containing the average heterozygosities of each individual,

$$\widehat{D}_{ii} = \frac{1}{m}\sum_{s=1}^{m} G_{si}(2 - G_{si}), \quad i = 1, \ldots, n.$$

Under reasonable regularity conditions, we can quantify the behaviour of $\widehat{B}$ and $\widehat{C}$ as the number of SNPs become large. Specifically, we assume the rows of F are independent and identically distributed with distribution $\text{Dist}(\mu, \Sigma)$, where μ denote the k-dimensional mean vector of the distribution, and Σ the $k \times k$-covariance matrix, that is,

$$F_{s\star} = (F_{s1}, \ldots, F_{sk}) \overset{\text{iid}}{\sim} \text{Dist}(\mu, \Sigma),$$

$s = 1, \ldots, m$. The matrix Q is assumed to be non-random, that is, fixed. These assumptions are standard and typically used in simulation of genetic data, see for example, Pickrell and Pritchard (2012); Cabreros and Storey (2019); Garcia-Erill and Albrechtsen (2020). Often $\text{dist}(\mu, \Sigma)$ is taken to be the product of k independent uniform distributions in which case $\mu = 0.5(1, 1, \ldots, 1)$ and Σ is a diagonal matrix with entries $1/12$, though other choices have been applied, see for example Balding and Nichols (1995); Conomos *et al.* (2016).

Let D be the diagonal matrix with entries

$$D_{ii} = 2\,\mathbb{E}[\Pi_{si}(1 - \Pi_{si})], \quad i = 1, \ldots, n. \tag{1}$$

It follows from Lemma 7 in the appendix, that $\widehat{D}$ converges to D as $m \to \infty$. Furthermore, as D_{ii} is the variance of G_{si} (it is binomial), then $\widehat{D}_{ii}$ might be considered an estimate of this variance. The proofs of the statements are in the appendix.

Theorem 1. *Let $k' \leq k$. Under the given assumptions, suppose further that $\widehat{P}_{k'} \to P_{k'}$ as $m \to \infty$, for some matrix $P_{k'}$. Then, $P_{k'}$ is an orthogonal projection. Furthermore, the following holds,*

$$\begin{aligned}\widehat{B} &\to (I - P_{k'})(D + 4Q^T\Sigma Q)(I - P_{k'}), \\ \widehat{C} &\to (I - P_{k'})D(I - P_{k'}),\end{aligned}$$

as $m \to \infty$. Hence, also

$$\begin{aligned}\widehat{B} - \widehat{C} &\to 4(I - P_{k'})Q^T\Sigma Q(I - P_{k'}) \\ &= 4(P - P_{k'})Q^T\Sigma Q(P - P_{k'}),\end{aligned}$$

as $m \to \infty$. For $k' = k$, if $P_k = P$, then the right hand side is the zero matrix, whereas this is not the case in general for $k' < k$.

Theorem 2. *Assume $k' = k$ and $P_k = P$. Furthermore, suppose as in Theorem 1 and that the vector with all entries equal to one is in the space spanned by the rows of Q (this is, for example, the case if the admixture proportions sum to one for each individual). Then,*

$$\frac{\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n}\widehat{B}_{ij}}{\sum_{i=1}^{n}\widehat{B}_{ii}} \to -1, \quad \text{as} \quad m \to \infty. \tag{2}$$

In addition, if Q takes the form

$$Q = \begin{pmatrix} Q_1 & 0 & \cdots & 0 \\ 0 & Q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_r \end{pmatrix}$$

where Q_ℓ has dimension $k_\ell \times n_\ell$, $\sum_{\ell=1}^r k_\ell = k$ and $\sum_{\ell=1}^r n_\ell = n$, then (2) holds for each component of n_ℓ individuals. If $Q_\ell = (1 \dots 1)$, then

$$\widehat{b}_{ij} \to -\frac{1}{n_\ell - 1}, \quad as \quad m \to \infty,$$

for all individuals i, j in the ℓ-th component, irrespective the form of $Q_{\ell'}$, $\ell' \neq \ell$.

Theorem 3. *Assume $k' = k$ and $P_k = P$. Furthermore, suppose as in Theorem 1 and that Q takes the form*

$$Q = \begin{pmatrix} Q_1 & Q_2 \\ 0 & Q_3 \end{pmatrix},$$

where $Q_1 = (1 \dots 1)$ has dimension $1 \times n_1$, $n_1 \leq n$. Then, $\widehat{b}_{ij}$ converges as $m \to \infty$ to a value larger than or equal to $-\frac{1}{n_1 - 1}$, for all $i, j = 1, \dots, n_1$.

The same statements in the last two theorems hold with $\widehat{B}$ and $\widehat{b}$ replaced by $\widehat{C}$ and $\widehat{c}$, respectively.

The three theorems provide means to evaluate the model. In particular, Theorem 1 might be used to assess the correctness (or appropriateness) of the proposed k', while Theorem 2 and Theorem 3 might be used to assess whether data from a group of individuals (e.g., a modern day population) originates from a single ancestral population, irrespective, the origin of the remaining individuals. We give examples in the Results section.

The work flow is shown in Algorithm 1. We process real and simulated genotype data using PCA 1, PCA 2, PCA 3, and the software ADMIXTURE, and evaluate the fit of the model.

Algorithm 1 Work flow of the proposed method

1. Choose k',
2. Compute an estimate $\widehat{P}_{k'}$ of the projection P,
3. Calculate the residuals $R_{k'} = G(I - \widehat{P}_{k'})$,
4. Calculate the correlation coefficients, $\widehat{b}$ and $\widehat{c}$,
5. Plot $\widehat{b}$ and the difference, the corrected correlation coefficients, $\widehat{b} - \widehat{c}$,
6. Assess visually the fit of the model.

Estimation of $P_{k'}$

Estimation of Q, F, and Π has received considerable interest in the literature, using for example, maximum likelihood (Wang 2003; Alexander *et al.* 2009), Bayesian approaches (Pritchard *et al.* 2000) or PCA (Engelhardt and Stephens 2010).

We discuss different ways to obtain an estimate $\widehat{P}_{k'}$ of P.

Using an estimate $\widehat{Q}_{k'}$ of $Q_{k'}$ An estimate $\widehat{P}_{k'}$ might be obtained by projecting onto the subspace spanned by the k' rows of $\widehat{Q}_{k'}$,

$$\widehat{P}_{k'} = \widehat{Q}_{k'}^T (\widehat{Q}_{k'} \widehat{Q}_{k'}^T)^{-1} \widehat{Q}_{k'},$$

assuming $\operatorname{rank}(\widehat{Q}_{k'}) = k'$ for the calculation to be valid.

We apply this approach to estimate the projection matrix using output from the software ADMIXTURE.

Using an estimate $\widehat{\Pi}_{k'}$ of $\Pi_{k'}$ Let $\widetilde{\Pi}_{k'}$ be k' linearly independent rows chosen from $\widehat{\Pi}_{k'}$ (out of m rows). Then, an estimate $\widehat{P}_{k'}$ of $P_{k'}$ is

$$\widehat{P}_{k'} = \widetilde{\Pi}_{k'}^T (\widetilde{\Pi}_{k'} \widetilde{\Pi}_{k'}^T)^{-1} \widetilde{\Pi}_{k'},$$

assuming $\operatorname{rank}(\widehat{\Pi}_{k'}) = k'$ for the calculation to be valid. Alternatively, one might apply the Gram-Schmidt method in which case the vectors are orthonormal by construction and $\widehat{P}_{k'} = \widetilde{\Pi}_{k'}^T \widetilde{\Pi}_{k'}$. The estimate $\widehat{P}_{k'}$ is independent of the choice of the k' rows, provided $\operatorname{rank}(\widehat{\Pi}_{k'}) = k'$.

Using PCA 1 We consider a PCA approach, originally due to Chen and Storey (2015), to estimate the space spanned by the rows of Q. We follow the procedure laid out in Cabreros and Storey (2019).

Let $\widehat{H}$ be the symmetric matrix

$$\widehat{H} = \frac{1}{m} G^T G - \widehat{D}.$$

Since $\widehat{H}$ is symmetric, all eigenvalues are real and the matrix is diagonalisable. Furthermore, $\widehat{H}$ is a variance adjusted version of $\frac{1}{m} G^T G$, see (1). Let $u_1, \dots, u_{k'}$ be $k' \leq k$ orthogonal eigenvectors belonging to the k' largest eigenvalues of $\widehat{H}$, counted with multiplicities. Define the $n \times k'$ matrix $U_{k'} = (u_1, \dots, u_{k'})$ and the $n \times n$ orthogonal projection matrix

$$\widehat{P}_{k'} = U_{k'} (U_{k'}^T U_{k'})^{-1} U_{k'}^T = U_{k'} U_{k'}^T$$

onto the subspace given by the span of the vectors $u_1, \dots, u_{k'}$.

In this particular case, convergence of $\widehat{P}_{k'}$ can be made precise. Define the matrix $H = 4Q^T(\Sigma + \mu\mu^T)Q$. Then, H is symmetric and positive semi-definite because Σ and $\mu\mu^T$ both are positive semi-definite. Hence, H has non-negative eigenvalues. Furthermore, according to Lemma 8 in the appendix, $\widehat{H}$ converges to H as $m \to \infty$.

Theorem 4. *Assume $k' \leq k$. Let $\lambda_1 \geq \dots \geq \lambda_n \geq 0$ be the eigenvalues of H, with corresponding orthogonal eigenvectors $v_1, \dots, v_n$. In particular, $\lambda_{k+1} = \dots = \lambda_n = 0$, as Q has rank k. Let $P_{k'}$ be the orthogonal projection onto the span of $v_1, \dots, v_{k'}$, that is,*

$$P_{k'} = V_{k'} (V_{k'}^T V_{k'})^{-1} V_{k'}^T = V_{k'} V_{k'}^T,$$

where $V_{k'} = (v_1, \dots, v_{k'})$.

Assume $k' = n$ or $\lambda_{k'} > \lambda_{k'+1}$, referred to as the eigenvalue condition. Then, $\widehat{P}_{k'} \to P_{k'}$ as $m \to \infty$. If the eigenvalue condition is fulfilled for $k' = k$, then $P_k = P$, that is, P_k is the orthogonal projection onto the span of the row vectors of Q. In particular, the eigenvalue condition is fulfilled for $k' = k$ if and only if $\Sigma + \mu\mu^T$ is positive definite. The latter is the case if Σ is positive definite.

For $k' = k$, the correct row space of Q is found eventually, but not Q itself. If $k' < k$, then a subspace of this row space is found, corresponding to the k' largest eigenvalues. As the data is not mean centred, we discard the first principal component, and use the subsequent $k' - 1$ eigenvectors and eigenvalues.

Using PCA 2 (mean centred data) A popular approach to estimation of Π in the admixture model is PCA based on mean centred data, or mean and variance normalised data (Pritchard *et al.* 2000; Engelhardt and Stephens 2010; Patterson *et al.* 2006).

Let $G_1 = G - \frac{1}{n}GE = G(I - \frac{1}{n}E)$ be the SNP-wise mean centred genotypes, where E is an $n \times n$ matrix with all entries equal to one. Following the exposition and notation in Cabreros and Storey (2019), let $G_1 = U\Delta V^T$ be the SVD of G_1, where ΔV^T consists of the row-wise principal components of G_1, ordered according to the singular values. Define

$$S_{k'} = \begin{pmatrix} U^T_{1:(k'-1)} \\ e \end{pmatrix},$$

where $e = (1\,1\,\ldots\,1)$ is a vector with all entries one, and $U^T_{1:(k'-1)}$ contains the top $k'-1$ rows of U^T. Then, an estimate of the projection is

$$\widehat{P}_{k'} = S^T_{k'}(S_{k'}S^T_{k'})^{-1}S_{k'}.$$

The squared singular values in the SVD decomposition of G_1 are the same as the eigenvalues of

$$\widehat{H}_1 = \frac{1}{m}G_1^T G_1 = \frac{1}{m}\left(I - \frac{1}{n}E\right)G^T G\left(I - \frac{1}{n}E\right)$$

(Jolliffe 2002). We have

$$\begin{aligned}\mathbb{E}[\widehat{H}_1] &= \frac{1}{m}\left(I - \frac{1}{n}E\right)\mathbb{E}[G^T G]\left(I - \frac{1}{n}E\right) \\ &= \left(I - \frac{1}{n}E\right)(D + 4Q^T(\Sigma + \mu\mu^T)Q)\left(I - \frac{1}{n}E\right). \quad (3)\end{aligned}$$

Let H_1 denote the right hand side of (3).

Theorem 5. *Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues of H_1, with corresponding orthogonal eigenvectors $v_1, \ldots, v_n$. In particular, $v_n = e$ and $\lambda_n = 0$. If D has all diagonal entries positive, then $\lambda_{n-1} > 0$.*

Let $k' \leq n$ and let $P_{k'}$ be the orthogonal projection onto the span of $v_1, \ldots, v_{k'-1}, e$, that is,

$$P_{k'} = V_{k'}(V_{k'}^T V_{k'})^{-1}V_{k'}^T,$$

where $V_{k'} = (v_1, \ldots, v_{k'-1}, e)$. If $k' = n$ or $\lambda_{k'} > \lambda_{k'+1}$, then $\widehat{P}_{k'} \to P_{k'}$ as $m \to \infty$.

There are no guarantees that for $k' = k$, we have $P_k = P$ and that the difference between $\widehat{B}$ and $\widehat{C}$ converges to zero for large m. However, this is the case under some extra conditions, and appears to be the case in many practical situations, see the Results section.

Theorem 6. *Assume $D = dI$ for some $d > 0$. Furthermore, assume the vector e is in the row space of Q (this is, for example, the case if the admixture proportions sum to one for each individual). Then, $\lambda_k = \ldots = \lambda_{n-1} = d$, and $\lambda_n = 0$.*

If $\Sigma + \mu\mu^T$ is positive definite, then $\lambda_{k+1} > \lambda_k$ and $P_k = P$, where P_k is as in Theorem 4. As a consequence, with $k' = k$ in Theorem 1, $\widehat{B} - \widehat{C} \to 0$ as $m \to \infty$.

Using PCA 3 (mean and variance normalised data) Let $G_2 = W^{-1}G_1$ be the SNP mean and variance normalised genotypes, where W is an $m' \times m'$ diagonal matrix with s-th entry being the observed standard deviation of the genotypes of SNP s. All SNPs for which no variation are observed are removed, hence the number of SNPs might be smaller than the original number, $m' \leq m$. Following the same procedure as for PCA 2, let $G_2 = U\Delta V^T$ be the SVD of G_2, where ΔV^T consists of the row-wise principal components of G_2, ordered according to the singular values. Define

$$S_{k'} = \begin{pmatrix} V^T_{1:(k'-1)} \\ e \end{pmatrix},$$

where $e = (1\,1\,\ldots\,1)$, and $V^T_{1:(k'-1)}$ contains the top $k'-1$ rows of V^T. Then, an estimate of the projection is $\widehat{P}_{k'} = S^T_{k'}(S_{k'}S^T_{k'})^{-1}S_{k'}$.

We are not aware of any theoretical justification of this procedure similar to Theorem 1, but it appears to perform well in many practical situations, according to our simulations.

Simulation of genotype data

We simulated genotype data from different demographic scenarios using different sampling strategies. We deliberately choose different sampling strategies to challenge the method. We first made simple simulations that illustrate the problem of model fit as well as to demonstrate the theoretical and practical properties of the residual correlations that arise from having data from a finite number of individuals and a large number of SNPs. An overview of the simulations are given in Table 1.

In the first two scenarios, the ancestral allele frequencies are simulated independently for each ancestral population from a uniform distribution, $F_{si} \sim \text{Unif}(0,1)$ for each site $s = 1, \ldots, m$ and each ancestral population $i = 1, \ldots, k$. In scenario 1, we simulated unadmixed individuals from three populations with either an equal or an unequal number of sampled individuals from each population. In scenario 2, we simulated two ancestral populations and a population that is admixed with half of its ancestry coming from each of the two ancestral populations.

In scenario 3, we set $F_{si} \sim \text{Unif}(0.01, 0.99)$ and simulated spatial admixture in a way that resembles a spatial decline of continuous gene flow between populations living in a long narrow island. We first simulated a single population in the middle of the long island. From both sides of the island, we then recursively simulated new populations from a Balding-Nichols distribution with parameter $F_{st} = 0.001$ using the R package 'bnpsd' (Ochoa and Storey 2019). In this way, each pair of adjacent populations along the island has an F_{st} of 0.001. Additional details on the simulation and an schematic visualization can be found in Figure 2 of Garcia-Erill and Albrechtsen (2020).

In scenario 4, we first simulated allele frequencies for an ancestral population from a symmetric beta distribution with shape parameter 0.03, $F_{si} \sim \text{Beta}(0.3, 0.3)$, which results in an allele frequency spectrum enriched for rare variants, mimicking the human allele frequency spectrum. We then sampled allele frequencies from a bifurcating tree (((pop1:0.1,popGhost:0.2):0.05,pop2:0.3):0.1,pop3:0.5), where pop1 and popGhost are sister populations and pop3 is an outgroup. Using the Balding-Nichols distribution and the F_{st} branch lengths of the tree (see Figure 5), we sampled allele frequencies in the four leaf nodes. Then, we created an admixed population with 30% ancestry from popGhost and 70% from pop2. We sampled 10 million genotypes for 50 individuals from each population except for the ghost population which was not included in the analysis, and subsequently removed sites with a sample minor allele frequency below 0.05, resulting in a total of 694,285 sites.

In scenario 5, we simulated an ancestral population with allele frequencies from a uniform distribution $F_{si} \sim \text{Unif}\,(0.05, 0.95)$, from which we sampled allele frequencies for two daughter populations from a Balding Nicholds distributions with $F_{st} = 0.3$ from the ancestral population, using 'bnpsd'. We then created recent hybrids based on a pedigree where all but one founder has ancestry from the first population. The number of generations in the pedigree then determines the admixture proportions and the age of the admixture where F1

individuals have one unadmixed parent from each population and backcross individuals have one unadmixed parent and the other F1. Double backcross individuals have one unadmixed parent and the other is a backcross. We continue to quadruple backcross with one unadmixed parent and the other triple backcross. Note that for the recent hybrids the ancestry of the pair of alleles at each loci is no longer independent which is a violation of the admixture model.

Results

Scenario 1

In this first set-up, we demonstrate the method using PCA 1 only. We simulated unadmixed individuals from $k = 3$ ancestral populations

$$Q = \begin{pmatrix} 1_{n_1} & 0 & 0 \\ 0 & 1_{n_2} & 0 \\ 0 & 0 & 1_{n_3} \end{pmatrix},$$

where 1_{n_i} is a row vector with all elements being one, and $n_1 + n_2 + n_3 = n$. We simulated genotypes for $n = 60$ individuals with sample sizes n_1, n_2 and n_3, respectively, as detailed in the previous section. In Figure 1(A), we show the residual correlation coefficients for $k' = 2, 3$ and plot the corresponding major PCs. For the PCA 1 approach, the first principal component does not relate to population structure as the data is not mean centered, and we use the following $k' - 1$ principal components.

When assuming that there are only two populations, $k' = 2$, we note that the empirical correlation coefficients appear largely consistent within each population sample, but the corrected correlation coefficients are generally non-zero with different signs, which points to model misfit. In contrast, when assuming the correct number of populations is $k' = 3$, the empirical correlation coefficients match nicely the theoretical values of $-\frac{1}{n_i - 1}$, which comply with Theorem 2 (see Table 2). A fairly homogeneous pattern in the corrected correlation coefficients appears around zero across all samples. This is a good indication that the model fits well and that the PCA plots using principal components 2 and 3 reflex the data well.

Scenario 2

In this set-up we also include admixed individuals. We simulated samples from two ancestral populations and individuals that are a mix of the two. We then applied all three PCA procedures and the software ADMIXTURE to the data. Specifically, we choose

$$Q = \begin{pmatrix} 1_{n_1} & \frac{1}{2}1_{n_2} & 0 \\ 0 & \frac{1}{2}1_{n_2} & 1_{n_3} \end{pmatrix},$$

with $k = 2$ true ancestral populations, and $(n_1, n_2, n_3) = (20, 20, 20)$ or $(n_1, n_2, n_3) = (10, 20, 30)$, see the previous section for details. We analysed the data with $k' = 1, 2, 3$, and obtained the correlation structure shown in Figures 2 and 3, and Table 2. The two standard approaches PCA 2 and PCA 3 show almost identical results, hence only PCA 2 is shown in the figures. Both PCA 2 and PCA 3 use the top principal components, while PCA 1 disregards the first, hence the discrepancy in the axis labeling in Figures 2(b) and 3/b). For $k' = 1$ none of the principal components are used and the predicted normalized genotypes is simply 0. All four methods show consistent results, in particular, for the correct k' (= 2), while there are smaller discrepancies between the methods for wrong $k' = 1, 3$. This is most pronounced for PCA 1 and ADMIXTURE. We note that the average correlation coefficient of $\widehat{b}$ within each population sample comply with Theorem 1 (see Table 2).

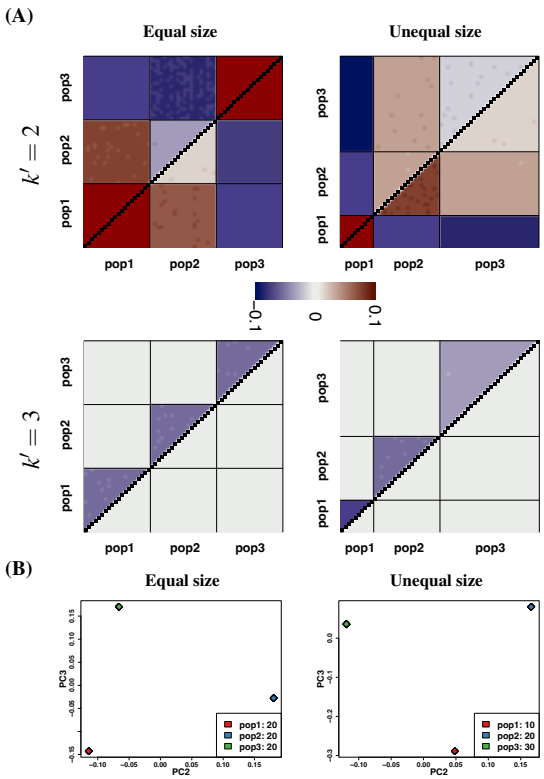


Figure 1 Results for simulated Scenario 1. (A) The upper triangle in the plots shows the empirical correlation coefficients $\hat{b}$ and the lower triangle shows the corrected correlation coefficients $\hat{b} - \hat{c}$. (B) The major principal components ($k' = 3$) result in a clear separation of the three samples (all data points within each sample are almost identical).

A fairly homogeneous pattern in the corrected correlation coefficients appears around zero across all samples for $k' = 2$, as in scenario 1, which shows that the model fits well. However, unlike in scenario 1 the bias for the empirical correlation coefficient is not a simple function of the sample size (see Table 2).

In this case, and similarly in all other investigated cases, we don't find any big discrepancies between the four methods. Therefore, we only show the results of PCA 1 for which we have theoretical justification for the results.

Scenario 3

We simulated genotypes for $n = 500$ individuals at $m = 88,082$ sites with continuous genetic flow between individuals, thus there is not a true k. We analysed the data assuming $k' = 2, 3$, see Figure 4. In the figure, the individuals are ordered according to the estimated proportions of the ancestral populations, hence it appears there is a color wave pattern in the empirical and the corrected correlation coefficients, see Figure 4(A). As expected, the corrected correlation coefficients are closer to zero for $k' = 3$ than $k' = 2$, though the deviations from zero are still large. We thus find no support for the model for either value of k'. This is consistent with the plots of the major PCs, that show

Table 1 Overview of simulations.

Scenario	k	n	m	Description	F_{is}[a]
1	3	20,20,20	$500K$	Unadmixed	$\text{Unif}(0,1)$
1	3	10,20,30	$500K$	Unadmixed	$\text{Unif}(0,1)$
2	2	20,20,20	$500K$	Admixed	$\text{Unif}(0,1)$
2	2	10,20,30	$500K$	Admixed	$\text{Unif}(0,1)$
3		500	$100K$[b]	Spatial with $F_{st} = 0.001$ between adjacent populations	$\text{Unif}(0.01,0.99)$
4	4	50,50,50,50,0[c]	$10M$[d]	Ghost admixture	$\text{Beta}(0.3,0.3)$
5	2	20,20,50	$500K$	Recent hybrids	$\text{Unif}(0.05,0.95)$

[a] Ancestral allele frequencies, $i = 1,\ldots,k$
[b] after applying MAF> 5% filtering, 88,082 remained.
[c] No reference samples are provided on the ghost population.
[d] after applying MAF> 5% filtering, 694,285 remained.

Table 2 The mean (standard deviation) of $\hat{b}$ and $\hat{b}-\hat{c}$ within each population using PCA 1.

Scenario 1	k'	n		pop1	pop2	pop3
	3	$(20,20,20)$	$\hat{b}$[a]	-0.0526 (0.0015)	-0.0526 (0.0016)	-0.0526 (0.0016)
				-0.0526	-0.0526	-0.0526
			$\hat{b}-\hat{c}$	0e-04 (0.0015)	0e-04 (0.0016)	0e-04 (0.0016)
		$(10,20,30)$	$\hat{b}$	-0.1111 (0.0011)	-0.0526 (0.0016)	-0.0345 (0.0016)
				-0.1111	-0.0526	-0.0345
			$\hat{b}-\hat{c}$	0e-04 (0.0012)	0e-04 (0.0016)	0e-04 (0.0016)
Scenario 2	k'	n		**pop1**	**admixed**	**pop3**
	2	$(20,20,20)$	$\hat{b}$	-0.0419 (0.0015)	-0.0192 (0.0015)	-0.0420 (0.0015)
				-0.0420	-0.0193	-0.0420
			$\hat{b}-\hat{c}$	0e-04 (0.0015)	0e-04 (0.0015)	0e-04 (0.0015)
		$(10,20,30)$	$\hat{b}$	-0.0701 (0.0018)	-0.0228 (0.0014)	-0.0304 (0.0016)
				-0.0701	-0.0229	-0.0304
			$\hat{b}-\hat{c}$	0e-04 (0.0017)	0e-04 (0.0014)	0e-04 (0.0016)
Scenario 4	k'	n		**pop1**	**pop2**	**pop3**	**pop4**
	3	$(50,50,50,50)$	$\hat{b}$	-0.0190 (0.0015)	0.0027 (0.0015)	-0.0204 (0.0017)	0.0122 (0.0013)
			$\hat{b}-\hat{c}$	0.0009 (0.0015)	0.0147 (0.0015)	0e-04 (0.0017)	0.0208 (0.0013)
	4		$\hat{b}$	-0.0204 (0.0015)	-0.0204 (0.0015)	-0.0204 (0.0017)	-0.0204 (0.0014)
			$\hat{b}-\hat{c}$	0e-04 (0.0015)	0e-04 (0.0015)	0e-04 (0.0017)	0e-04 (0.0013)

[a] The second line of $\hat{b}$ in each case shows the theoretical value obtained from the limit in Theorem 1.

continuous change without grouping the data into two or three clusters, see Figure 4(B).

Scenario 4

This case is based on the tree in Figure 5, which include an unsampled (so-called) ghost population, popGhost. The popGhost is sister population to pop1.

We simulated genotypes for $n = 200$ individuals: 150 unadmixed samples from pop1, pop2, and pop3; and 50 samples admixed with 0.3 ancestry from popGhost and 0.7 ancestry from pop2 (as pop4), as detailed in the previous section. As there is drift between the populations and hence genetic differences, the correct $k = 4$ (pop1, pop2, pop3, popGhost). This is picked up by our method that clearly shows $k' = 3$ is wrong with large deviation from zero in the corrected correlation coefficients. In contrast, for $k' = 4$, the corrected correlation coefficients are almost zero (Figure 6).

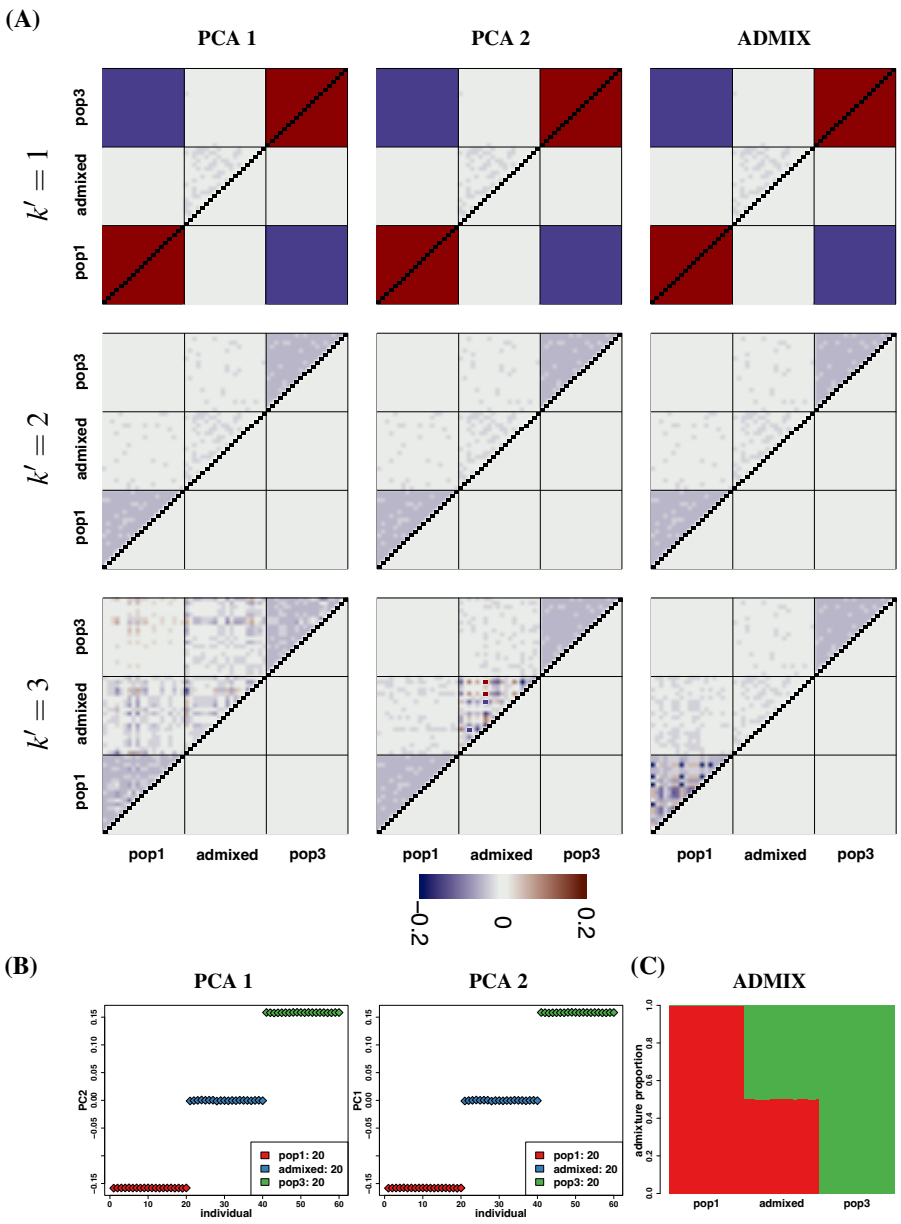


Figure 2 Results for simulated Scenario 2 with equal sample sizes. (A) For each of PCA 1, PCA 2 and ADMIXTURE, the upper left triangle in the plots shows the empirical correlation $\hat{b}$ and the lower right triangle shows the difference $\hat{b} - \hat{c}$ with sample sizes $(n_1, n_2, n_3) = (20, 20, 20)$. (B) The major principal component for the PCA based methods for $k' = 2$ (in which case there is only one principal component). Individuals within each sample have the same color. (C) The estimated admixture proportions in the case of ADMIXTURE.

Scenario 5

In the last example, we simulated two populations (originating from a common ancestral population) and created admixed populations by backcrossing, as detailed in the previous section. Thus, the model does not fulfil the assumptions of the admixture model in that the number of reference alleles are not binomially distributed, but depends on the particular backcross and the frequencies of the parental populations.

We simulate genotypes for $n = 90$ individuals at $m = 500,000$ sites.

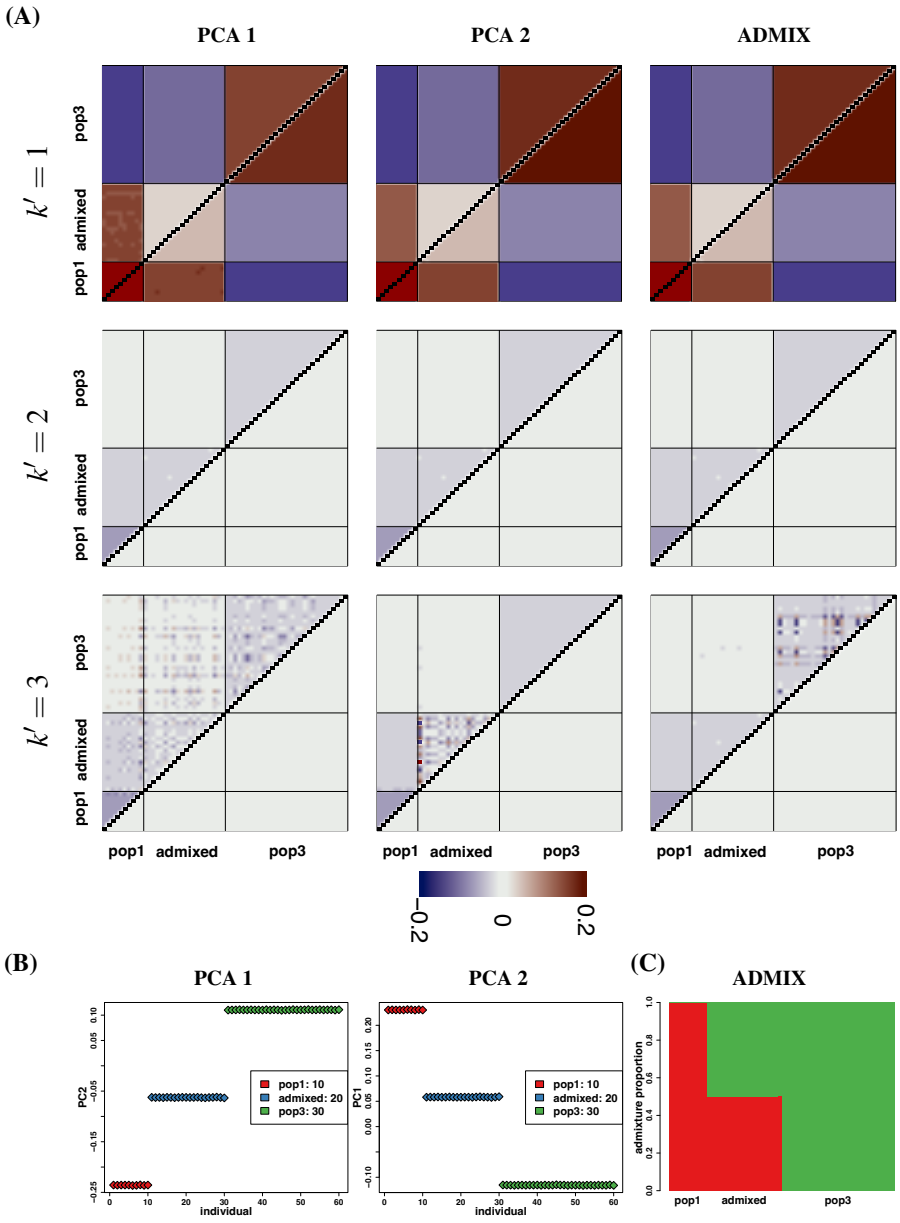


Figure 3 Results for simulated Scenario 2 with unequal sample sizes. (A) For each of PCA 1, PCA 2 and ADMIXTURE, the upper left triangle in the plots shows the empirical correlation $\hat{b}$ and the lower right triangle shows the difference $\hat{b} - \hat{c}$ with sample sizes $(n_1, n_2, n_3) = (20, 20, 20)$. (B) The major principal component for the PCA based methods for $k' = 2$ (in which case there is only one principal component). Individuals within each sample have the same color. (C) The estimated admixture proportions in the case of ADMIXTURE.

There are 20 homogeneous individuals from each parental population, and 10 different individuals from each of the different recent admixture classes. Then, we analysed the data with $k' = 2$ and found the corrected correlation coefficients deviated consistently from zero, in particular for one of the parental populations (Figure 7). We are thus able to say the admixture model does not provide a reasonable fit.

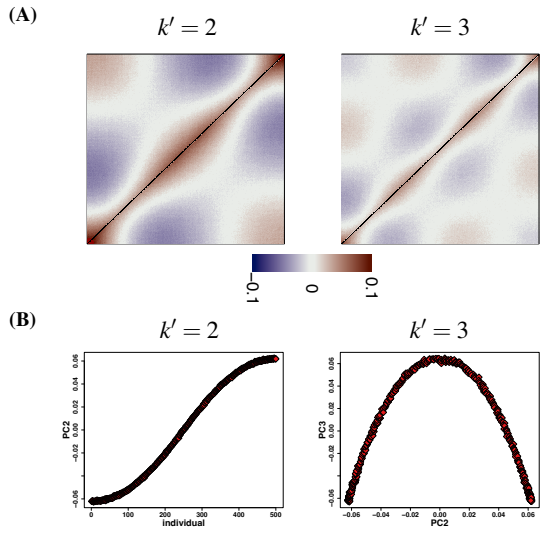


Figure 4 Results for simulated scenario 3. (A) The upper triangle in the plots shows the empirical correlation $\hat{b}$ and the lower triangle shows the difference $\hat{b} - \hat{c}$. (B) The major principal components (only one in the case of $k' = 2$).

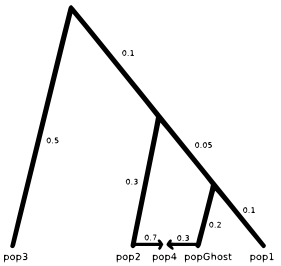


Figure 5 Schematic of the tree used to simulate population allele frequencies for Scenario 4, including 5 populations: pop1, pop2, pop3, pop4 and popGhost. The pop4 population is the result of admixture between pop2 and popGhost, for which there are no individuals sampled and is therefore a ghost population. The values in the branches indicate the drift in units of F_{ST}. The values along the two admixture edges are the admixture proportions coming from each population.

Real data

We analysed a whole genome sequencing data set from the 1000 Genomes Project (Auton *et al.* 2015), see also Garcia-Erill and Albrechtsen (2020) where the same data is used. It consists of data from five groups of different descent: a Yoruba group from Ibadan, Nigeria (YRI), residents from Southwest US with African ancestry (ASW), Utah residents with Northern and Western European ancestry (CEU), a group with Mexican ancestry from Los Angeles, California (MXL), and a group of Han Chinese from Beijing, China (CHB) with sample sizes 108, 61, 99, 63 and 103, respectively, in total, $n = 434$. We kept only sites present in the Human Origins SNP panel (Lazaridis *et al.* 2014), with a total of $m = 406,279$ SNPs were left after a MAF filter of 0.05.

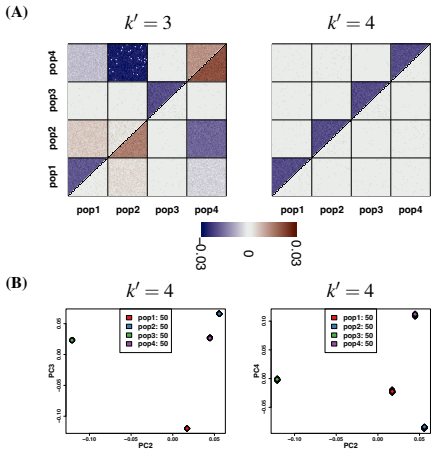


Figure 6 Results for simulated scenario 4. (A) The upper triangle in the plots shows the empirical correlation $\hat{b}$ and the lower triangle shows the difference $\hat{b} - \hat{c}$. (B) The major principal components for $k' = 4$, that result in a clear separation of the four samples (all data points within each sample are almost identical).

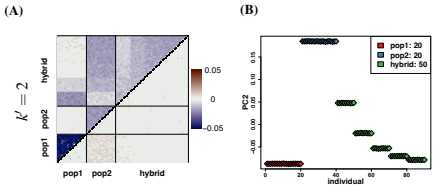


Figure 7 Results for simulated scenario 5 (recent admixture). (A) The upper triangle in the plots shows the empirical correlation $\hat{b}$ and the lower triangle shows the difference $\hat{b} - \hat{c}$. (B) The major principal component for $k' = 2$.

We analyzed the data with $k' = 3, 4$. For $k' = 3$, Figure 8 shows that it is not possible to explain the relationship between MXL, CEU and CHB, indicating that MXL is not well explained as a mixture of the two. For $k' = 4$, the color shades of the corrected correlation coefficients are almost negligible within each population, pointing at a contribution from a native american population. This is further corroborated in Figure 8(D) that shows estimated proportions from the four ancestral populations using the software ADMIXTURE.

Discussion

We have developed a novel approach to assess model fit of PCA and the admixture model based on structure of the residual correlation matrix. We have shown that it performs well for simulated and real data, using a suit of different PCA methods, commonly used in the literature, and the ADMIXTURE software to estimate model parameters. By assessing the residual correlation structure visually, one is able to detect model misfit and violation of modelling assumptions.

The model fit is assessed by comparing visually two matrices of residual correlation coefficients. The theoretical and practical advantage of our approach lie in three aspects. First, our approach is computationally simple and fast. Calculation of the two residual correlation matrices and their difference is computationally inexpensive. Secondly,

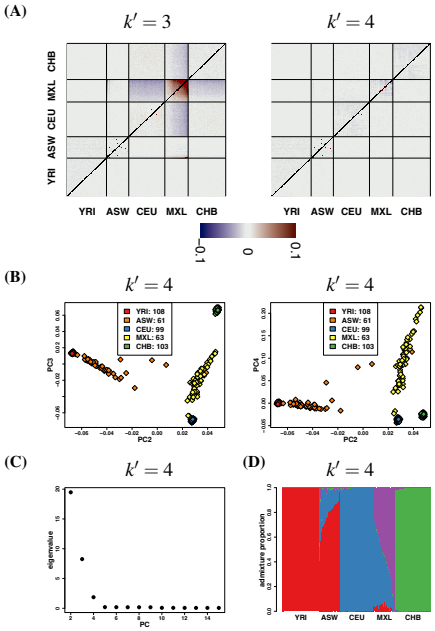


Figure 8 The residual correlation coefficient, the inferred population structure and the admixture proportions of a real human data from 1000 Genomes project. (A) The upper triangle in the plots shows the empirical correlation coefficient $\hat{b}$ and the lower triangle shows the difference $\hat{b} - \hat{c}$. (B) The three major principal component for PCA 1 for $k' = 4$. (C) The eigenvalue for the first PC is removed and the eigenvalues corresponding to the remaining PCs are close to 0 after the forth PC. (D) The admixture proportions as estimated with ADMIXTURE.

our approach provides a unified approach to model fitting based on PCA and clustering methods (like ADMIXTURE). In particular, it provides simple means to assess the adequacy of the chosen number of top principal components to describe the structure of the data. Assessing the adequacy by plotting the principal components against each other might lead to false confidence. In contrast, our approach exposes model misfit by plotting the difference between two matrices of the residual correlation coefficients. Thirdly, it comes with theoretical guarantees in some cases. These guarantees are further back up by simulations in cases, we cannot provide theoretical validity. Finally, our approach might be adapted to work on NGS data without estimating genotypes first, but working directly on genotype likelihoods.

Data availability

The data sets used in this study are all publicly available, including simulated and real data. Information about the R code used to analyze and simulate data is available at https://github.com/Ginwaitthreebody/evalPCA. The variant calls for the 1000 Genomes Project data used are publicly available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Acknowledgements

The authors are supported by the Independent Research Fund Denmark (grant number: 8021-00360B) and the University of Copenhagen through the Data+ initiative. SL acknowledges the financial support from the funding agency of China Scholarship Council. GGE and AA are supported by the Independent Research Fund Denmark (grant numbers: 8049-00098B and DFF-0135-00211B respectively).

Appendix A

We first state the expectation and covariance matrix of $G_{s\star}$ and $\Pi_{s\star}$, respectively, under the given distributional assumptions,

$$\mathbb{E}[\Pi_{s\star}] = \mu^T Q, \qquad \operatorname{cov}(\Pi_{s\star}) = Q^T \Sigma Q,$$
$$\mathbb{E}[G_{s\star}] = 2\mathbb{E}[F_{s\star} Q] = 2\mu^T Q,$$
$$\operatorname{cov}(G_{s\star}) = \mathbb{E}[\operatorname{cov}(G_{s\star} \mid \Pi)] + 4\operatorname{cov}(\Pi_{s\star}) = D + 4Q^T \Sigma Q,$$

for $s = 1, \ldots, m$, where

$$D = 2\mathbb{E}[\operatorname{diag}(\Pi_{s1}(1-\Pi_{s1}), \ldots, \Pi_{sn}(1-\Pi_{sn}))],$$

and

$$\mathbb{E}[\Pi_{si}(1-\Pi_{si})] = \mu^T Q_{\star i} - (\mu^T Q_{\star i})^2 - (Q^T \Sigma Q)_{ii}.$$

The unconditional columns $G_{s\star}$, $s = 1, \ldots, m$, of G are independent random vectors by construction.

The above implies that

$$\frac{1}{m}\mathbb{E}[G^T G] = D + 4Q^T(\Sigma + \mu\mu^T)Q. \tag{4}$$

Auxiliary results are in appendix B.

Lemma 7. *The estimator $\widehat{D}$ is an unbiased estimator of D, that is, $\mathbb{E}[\widehat{D}] = D$. Furthermore, it holds that $\widehat{D} \to D$ as $m \to \infty$.*

Proof. Conditional on Π_{si}, using binomiality, we have $\mathbb{E}[G_{si}(2 - G_{si}) \mid \Pi_{si}] = 2\Pi_{si}(1-\Pi_{si})$, and the first result follows. For convergence, note that $G_{si}(2-G_{si})$, $s = 1, \ldots, m$, unconditionally, form a sequence of iid random variables with finite variance, hence the convergence statement follows from the strong Law of Large Numbers (Jacod and Protter 2004). □

Lemma 8. *The estimator $\widehat{H} = \frac{1}{m}G^T G - \widehat{D}$ is an unbiased estimator of $H = 4Q^T(\Sigma + \mu\mu^T)Q$, that is, $\mathbb{E}[\widehat{H}] = H$. Furthermore, it holds that $\widehat{H} \to 4Q^T(\Sigma + \mu^T\mu)Q$ as $m \to \infty$, and*

$$\mathbb{E}\left[\|\widehat{H} - 4Q^T(\Sigma + \mu\mu^T)Q\|_F^2\right] \leq \frac{16n^2}{m}.$$

Proof. Unbiasedness follows from (4) and Lemma 7. Consider the (i,j)-th entry of $\frac{1}{m}G^T G$, namely, $\frac{1}{m}\sum_{s=1}^m G_{si}G_{sj}$. The sequence $G_{si}G_{sj}$, $s = 1, \ldots, m$, is iid with finite variance, hence $\frac{1}{m}G^T G$ converges to $\mathbb{E}[G^T G]$ as $m \to \infty$ by the strong Law of Large Numbers (Jacod and Protter 2004). Combined with Lemma 7 gives convergence of $\widehat{H}$ to H as $m \to \infty$.

It remains to prove the inequality. Define

$$A_{s,ij} = \begin{cases} G_{si}G_{sj} - 4(Q^T(\Sigma + \mu\mu^T)Q)_{ij} & \text{if } i \neq j, \\ 2G_{si}(G_{si}-1) - 4(Q^T(\Sigma + \mu\mu^T)Q)_{ii} & \text{if } i = j. \end{cases}$$

Then,

$$(\widehat{H}_{ij} - \mathbb{E}[\widehat{H}_{ij}])^2 = \left(\frac{1}{m}\sum_{s=1}^m A_{s,ij}\right)^2 = \frac{1}{m^2}\sum_{s=1}^m\sum_{t=1}^m A_{s,ij}A_{t,ij},$$
$$\|\widehat{H} - \mathbb{E}[\widehat{H}]\|_F^2 = \frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m\sum_{t=1}^m A_{s,ij}A_{t,ij}.$$

Using $\mathbb{E}[A_{s,ij}] = 0$, independence of $A_{s,ij}$ and $A_{t,ij}$ for $s \neq t$, and $|A_{s,ij}| \leq 4$, we have

$$\mathbb{E}[\|\widehat{H} - \mathbb{E}[\widehat{H}]\|_F^2] = \frac{1}{m^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{s=1}^m \mathbb{E}[A_{s,ij}^2] \leq \frac{1}{m^2}16mn^2 = \frac{16n^2}{m},$$

which proves the claim. □

The convergence result is also in Chen and Storey (2015, theorem 2). The second part provides the rate of convergence of $\widehat{H}$ in the L^2-norm. Convergence is contingent on large m, rather than large n, and requires m to increase at least like the square of n.

Proof of Theorem 1. Since $\widehat{P}_{k'}$ is assumed to be an orthogonal projection, that is, $\widehat{P}_{k'}^2 = \widehat{P}_{k'}$ and $\widehat{P}_{k'}^T = \widehat{P}_{k'}$, then also the limit is an orthogonal projection, $P_{k'}^2 = P_{k'}$ and $P_{k'}^T = P_{k'}$.

Consider the empirical covariance $\widehat{B}$. Define the variables $T_{k'} = G(I - P_{k'})$ with $\widehat{P}_{k'}$ replaced by $P_{k'}$, and the empirical covariance

$$\begin{aligned}\widetilde{B}_{ij} &= \frac{1}{m-1}\sum_{s=1}^{m}(T_{k',si}T_{k',sj} - \overline{T}_{k',i}\overline{T}_{k',j}) \\ &= \frac{1}{m-1}\sum_{s=1}^{m} T_{k',si}T_{k',sj} - \frac{m}{m-1}\overline{T}_{k',i}\overline{T}_{k',j},\end{aligned}$$

defined similarly to $\widehat{B}_{ij}$, with $\overline{T}_{k',i} = \frac{1}{m}\sum_{s=1}^{m} T_{k',si}$. The sequences $T_{k',si}T_{k',sj}$, $s = 1,2,\ldots$, and $T_{k',si}$, $s = 1,2,\ldots$, are iid random variables, by the distributional assumptions on G. Furthermore, since $P_{k'}$ is an orthogonal projection, then $\|I - P_{k'}\|_F^2 \le n$ is bounded (Lemma 12). Therefore, also $T_{k',si}$ is bounded uniformly in s,i by $2\sqrt{n} \le 2n$.

Using boundedness, independence and the strong Law of Large Numbers (Jacod and Protter 2004),

$$\widetilde{B}_{ij} \to \mathbb{E}[T_{k',1i}T_{k',1j}] - \mathbb{E}[T_{k',1i}]\,\mathbb{E}[T_{k',1j}] = \operatorname{cov}(T_{k',1i}, T_{k',1j}), \quad (5)$$

for $m \to \infty$, and $\operatorname{cov}(T_{k',1i}, T_{k',1j}) = (I - P_{k'})(D + 4Q^T\Sigma Q)(I - P_{k'})$. The latter equality follows from (4).

Consider $R = G(I - \widehat{P}_{k'}) = G(I - P_{k'}) + G(P_{k'} - \widehat{P}_{k'}) = T + G(P_{k'} - \widehat{P}_{k'})$. Hence,

$$\begin{aligned}|\overline{R}_{k',i} - \overline{T}_{k',i}| &\le \frac{1}{m}\sum_{s=1}^{m}\sum_{i'=1}^{n}\sum_{j'=1}^{n} 2|(P_{k'} - \widehat{P}_{k'})_{i'j'}| \\ &= 2\sum_{i'=1}^{n}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{i'j'}| \to 0,\end{aligned}$$

as $m \to \infty$ by assumption of the theorem. It follows that $\overline{R}_{k',i}$ converges to $\mathbb{E}[T_{k',1i}]$ as $m \to \infty$. Furthermore,

$$\begin{aligned}&\frac{1}{m-1}\sum_{s=1}^{m} R_{k',si}R_{k',sj} - \frac{1}{m-1}\sum_{s=1}^{m} T_{k',si}T_{k',sj} \\ &= \frac{1}{m-1}\sum_{s=1}^{m}(T_{si} + (G(P_{k'} - \widehat{P}_{k'}))_{si})(T_{sj} + (G(P_{k'} - \widehat{P}_{k'}))_{sj}) \\ &\quad - \frac{1}{m-1}\sum_{s=1}^{m} T_{k',si}T_{k',sj} \\ &= \frac{1}{m-1}\sum_{s=1}^{m} T_{si}(G(P_{k'} - \widehat{P}_{k'}))_{sj} + \frac{1}{m-1}\sum_{s=1}^{m}(G(P_{k'} - \widehat{P}_{k'}))_{si}T_{sj} \\ &\quad + \frac{1}{m-1}\sum_{s=1}^{m}(G(P_{k'} - \widehat{P}_{k'}))_{si}(G(P_{k'} - \widehat{P}_{k'}))_{sj}.\end{aligned}$$

The absolute value of the first term in the last line above is bounded by

$$\frac{4nm}{m-1}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{j'j}|,$$

and similarly for the second term. The third is bounded by

$$\frac{4m}{m-1}\sum_{i'=1}^{n}\sum_{j'=1}^{n}|(P_{k'} - \widehat{P}_{k'})_{i'i}||(P_{k'} - \widehat{P}_{k'})_{j'j}|.$$

All three terms converge to zero as $m \to \infty$, hence we conclude from (5) that $\widehat{B}_{ij} \to \operatorname{cov}(T_{k',1i}, T_{k',1j})$ as $m \to \infty$.

The result for the estimated covariance $\widehat{C}$ follows from convergence of $\widehat{D}$ and by assumption of the theorem. The remaining part follows from the convergence of $\widehat{B}$ and $\widehat{C}$. Note that $QP = Q$, hence the second equation holds. The last statement of the theorem follows directly.

Proof of Theorem 2. Consider $T_k = G(I - P_k) = G(1 - P)$, where $P = Q^T(QQ^T)^{-1}Q$ is the projection onto the row space of Q. Then, T_k contains the residuals under multiple regression of the m rows of G on the k rows of Q (Box *et al.* 2005). Since e is in the row space of Q, then the sum of the residuals is zero for each $s = 1,\ldots,m$: $\sum_{i=1}^{n} T_{k,si} = 0$ (the assumption that e is in the row space is equivalent to having an intercept in the regression model) (Box *et al.* 2005). We have, for $s = 1,\ldots,m$,

$$\begin{aligned}0 = \operatorname{var}\left(\sum_{i=1}^{n} T_{k,si}\right) &= \sum_{i=1}^{n}\operatorname{var}(T_{k,si}) + \sum_{i=1}^{n}\sum_{j=1,i\ne j}^{n}\operatorname{cov}(T_{k,si}, T_{k,sj}) \\ &= \sum_{i=1}^{n}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n}\sum_{j=1,i\ne j}^{n}\operatorname{cov}(T_{k,1i}, T_{k,1j}),\end{aligned}$$

since the distribution of $T_{k,si}$ is independent of s. From the proof of Theorem 4, it follows that $\widehat{B}$ converges to $\operatorname{cov}(T_{k,1\star})$ as $m \to \infty$. Hence,

$$\sum_{i=1}^{n}\widehat{B}_{ii} + \sum_{i=1}^{n}\sum_{j=1,i\ne j}^{n}\widehat{B}_{ij} \to 0, \quad \text{as} \quad m \to \infty,$$

and the desired result follows by rearrangement.

If Q takes the given form, then the residuals under multiple regression are independent between compartments, as the projection is

$$P = \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_r \end{pmatrix},$$

where $P_\ell = Q_\ell^T(Q_\ell Q_\ell^T)^{-1}Q_\ell$ has dimension $n_\ell \times n_\ell$. It follows that the computation above holds for each compartment. Finally, if $Q_\ell = (1\ldots1)$, then the distribution of the random vector $T_{k,1\star}$ is exchangeable, resulting in

$$\begin{aligned}0 = \operatorname{var}\left(\sum_{i=1}^{n_\ell} T_{k,1i}\right) &= \sum_{i=1}^{n_\ell}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n_\ell}\sum_{j=1,i\ne j}^{n_\ell}\operatorname{cov}(T_{k,1i}, T_{k,1j}) \\ &= n_\ell\operatorname{var}(T_{k,11}) + n_\ell(n_\ell - 1)\operatorname{cov}(T_{k,11}, T_{k,12})\end{aligned}$$

assuming the individuals in the ℓ-th compartment are numbered 1 to n_ℓ. Rearranging terms and substituting $\widehat{b}_{ij}$ for the moments of $T_{k,i\star}$ yields the desired result.

Proof of Theorem 3. Consider $T_k = G(I - P_k) = G(1 - P)$, where $P = Q^T(QQ^T)^{-1}Q$ is the projection onto the row space of Q. If $Q_1 = (1\ldots1)$, then the distribution of the random variables $T_{k,11},\ldots,T_{k,1n_1}$ are exchangeable, resulting in

$$\begin{aligned}0 \le \operatorname{var}\left(\sum_{i=1}^{n_1} T_{k,1i}\right) &= \sum_{i=1}^{n_1}\operatorname{var}(T_{k,1i}) + \sum_{i=1}^{n_1}\sum_{j=1,i\ne j}^{n_1}\operatorname{cov}(T_{k,1i}, T_{k,1j}) \\ &= n_1\operatorname{var}(T_{k,11}) + n_1(n_1 - 1)\operatorname{cov}(T_{k,11}, T_{k,12}).\end{aligned}$$

Rearranging terms and substituting $\widehat{b}_{ij}$ for the moments of $T_{k,i\star}$ yields the desired result.

Proof of Theorem 4. The convergence statement of the theorem is a special case of Theorem 9 in Appendix B. Take $A_m = \widehat{H}$ (that depends

on the number of SNPs m, and the particular realization), $A = H$, and $k = k'$ in the theorem (k is used as a generic index in Theorem 9). Then, $E_k E_k^T = P_{k'}$ and $F_{m,k'} F_{m,k}^T = \widehat{P}_{k'}$, and the conclusion of Theorem 4 holds. Convergence in Frobenius norm is equivalent to pointwise convergence (as n is fixed) $\widehat{P}_{k'} \to P_{k'}$ as $m \to \infty$ by definition.

If $\Sigma + \mu\mu^T$ is positive definite, then it has rank k. As $\operatorname{rank}(Q) = k$ by assumption, it follows from Lemma 13 that $\operatorname{rank}(H) = k$. Consequently, there are k positive eigenvalues of H and $\lambda_{k+1} = 0$, and the eigenvalue condition holds. Conversely, assume the eigenvalue condition holds. By definition $\operatorname{rank}(H) \le k$. As $\lambda_k > \lambda_{k+1} \ge 0$ by assumption, then also $\operatorname{rank}(H) \ge k$ and we conclude $\operatorname{rank}(H) = k$. It follows that the rank of $\Sigma + \mu\mu^T$ is k; consequently, it is positive definite.

If $k' = k = n$, then $P_k = V_k V_k^T = I$ and $P = I$ (as $k = n$), and $P_k = P$. So assume $k' = k < n$. Since the eigenvalue condition is fulfilled, then from the above, we have $\operatorname{rank}(Q^T(\Sigma + \mu\mu^T)) = k$, and Lemma 13 yields that the row space of H and Q agree. Similarly, we have $H = V_k \operatorname{diag}(\lambda_1, \ldots, \lambda_k) V_k^T$ and Lemma 13 yields that the row space of H and V_k^T agree. This implies the row space of Q and V_k^T agree. Consequently, $P_k = Q^T(QQ^T)^{-1}Q = P$, and the statement holds.

Proof of Theorem 5. It follows trivially that e is an eigenvector of H_1 with eigenvalue 0. If D has all entries positive, then it is positive definite and $D + 4Q^T(\Sigma + \mu\mu^T)Q$ is also positive definite, hence has rank n. It follows from Lemma 13 that H_1 has rank $n-1$, hence $\lambda_{n-1} > 0$.

Similarly to the proof of Lemma 8 in Appendix B, one can show $\mathbb{E}[\widehat{H}_1] = H_1$ and $\widehat{H}_1 \to H_1$ as $m \to \infty$, where H_1 denotes the right hand side of (3). The remaining part of the theorem is proven similarly to Theorem 4.

Proof of Theorem 6. Note that e is an eigenvector of H_1 with eigenvalue 0. Consider an eigenvector v of H_1, orthogonal to e with eigenvalue λ. Then, the following two equations are equivalent,

$$\left(I - \frac{1}{n}E\right)(D + 4Q^T(\Sigma + \mu\mu^T)Q)\left(I - \frac{1}{n}E\right)v = \lambda v,$$
$$4\left(I - \frac{1}{n}E\right)Q^T(\Sigma + \mu\mu^T)Q\left(I - \frac{1}{n}E\right)v = (\lambda - d)v, \quad (6)$$

where it is used that $D = dI$ and $v \perp e$. It shows that v is an eigenvector of $K = 4(I - \frac{1}{n}E)Q^T(\Sigma + \mu\mu^T)Q(I - \frac{1}{n}E)$ with eigenvalue $\mu = \lambda - d$. Since Q has rank k and the vector e is in the space spanned by the rows of Q, then $Q(I - \frac{1}{n}E)$ has rank $k-1$. It follows that there are at most $k-1$ positive eigenvalues of K, that is, at most $k-1$ eigenvalues of H_1 such that $\lambda > d$. Furthermore, there are precisely $k-1$ positive eigenvalues, provided $\Sigma + \mu\mu^T$ is positive definite (Lemma 13). The remaining eigenvalues of K are zero, that is, the corresponding eigenvalues of H_1 are $\lambda = d$.

Assume $\Sigma + \mu\mu^T$ is positive definite, then by the above argument there precisely are $k-1$ eigenvalues of H_1 such that $\lambda > d$ with corresponding orthogonal eigenvectors $v_1, \ldots, v_{k-1}$. It follows from (6) that $v_1, \ldots, v_{k-1}$ are in the space spanned by the rows of $Q(I - \frac{1}{n}E)$, hence the eigenvectors are in the space spanned by the rows of Q. By assumption e is also in that row span. Hence, $v_1, \ldots, v_{k-1}, e$ forms an orthogonal basis of the row span of Q, as Q has rank k. Thus, $P_k = P$.

Appendix B

Theorem 9. *Let A_m be a sequence of symmetric $n \times n$-matrices that converges to a symmetric $n \times n$-matrix A in the Frobenius norm, that is $\|A_m - A\|_F \to 0$, as $m \to \infty$. Let $\lambda_1 \ge \ldots \ge \lambda_n$ be the eigenvalues of A (with multiplicity, and not necessarily non-negative). Let $k \le n$ be given and assume either $k = n$ or $\lambda_k > \lambda_{k+1}$. Furthermore, let $e_1, \ldots, e_k$ be orthogonal eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$, respectively, and let $f_{m,1}, \ldots, f_{m,k}$ be orthogonal eigenvectors corresponding to the k largest eigenvalues of A_m (with multiplicity). Then, the orthogonal projection onto the span of $f_{m,1}, \ldots, f_{m,k}$ converges to the orthogonal projection onto the span of $e_1, \ldots, e_k$ in the Frobenius norm. That is, define $E_k = (e_1, \ldots, e_k)$ and $F_{m,k} = (f_{m,1}, \ldots, f_{m,k})$, then $\|F_{m,k}F_{m,k}^T - E_k E_k^T\|_F \to 0$ as $m \to \infty$.*

Proof. If $k = n$, then $E_n E_n^T = I$ and $F_{m,n}F_{m,n}^T = I$, and the statement is trivial. Hence, assume $k < n$. Let $e_1, \ldots, e_n$ be eigenvectors of A corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$, respectively. Let $f_{m,1}, \ldots, f_{m,n}$ be the eigenvectors of A_m corresponding to the eigenvalues $\mu_{m,1} \ge \ldots \ge \mu_{m,n}$. All eigenvectors can be asssumed to be orthonormal.

As $\|A - A_m\|_F^2 \to 0$ for $m \to \infty$, then every entry of A_m converges to the corresponding entry of A. Consequently, the characteristic function of A_m converges to that of A, and the eigenvalues of A_m converges to those of A, that is, $\mu_{m,j} \to \lambda_j$ for $j = 1, \ldots, n$, and $m \to \infty$. Let T_m be such that $E_n = F_{m,n}T_m$. As E_n and $F_{m,n}$ are orthogonal matrices, hence also T_m is orthogonal. Applying Lemma 10 in the first and third line gives

$$\begin{aligned}
\|A - A_m\|_F^2 &= \|AE - A_m E_n\|_F^2 = \|E\operatorname{diag}(\lambda_1, \ldots, \lambda_n) - A_m F_{m,n} T_m\|_F^2 \\
&= \|F_{m,n}T_m \operatorname{diag}(\lambda_1, \ldots, \lambda_n) - F_{m,n}\operatorname{diag}(\mu_{m,1}, \ldots, \mu_{m,n})T_m\|_F^2 \\
&= \|T_m \operatorname{diag}(\lambda_1, \ldots, \lambda_n) - \operatorname{diag}(\mu_{m,1}, \ldots, \mu_{m,n})T_m\|_F^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n (\lambda_j T_{m,ij} - \mu_{m,i} T_{m,ij})^2 = \sum_{i=1}^n \sum_{j=1}^n T_{m,ij}^2 (\lambda_j - \mu_{m,i})^2.
\end{aligned}$$

By assumption, $\lambda_k > \lambda_{k+1}$. Hence, by convergence of eigenvalues, for $j \le k$, $i \ge k+1$, or $j \ge k+1$, $i \le k$, we have $T_{m,ij} \to 0$ as $m \to \infty$.

Furthermore,

$$\begin{aligned}
E_k E_k^T - F_{m,k}F_{m,k}^T &= \sum_{\ell=1}^k \left(e_\ell e_\ell^T - f_{m,\ell} f_{m,\ell}^T\right) \\
&= \sum_{\ell=1}^k \left(\left(\sum_{a=1}^n f_{m,a} T_{m,a\ell}\right)\left(\sum_{a=1}^n f_{m,a} T_{m,a\ell}\right)^T - f_{m,\ell} f_{m,\ell}^T\right) \\
&= \sum_{\ell=1}^k \left(\sum_{a=1}^n \sum_{b=1}^n T_{m,a\ell} T_{m,b\ell} f_{m,a} f_{m,b}^T - f_{m,\ell} f_{m,\ell}^T\right) \\
&= \sum_{a=1}^n \sum_{b=1}^n \sum_{\ell=1}^k T_{m,a\ell} T_{m,b\ell} f_{m,a} f_{m,b}^T - \sum_{\ell=1}^k f_{m,\ell} f_{m,\ell}^T. \\
&= \sum_{(a,b) \in \{1,\ldots,n\}^2 \setminus A_{1,k}} \left(\sum_{i=1}^k T_{m,ai} T_{m,bi}\right) f_{m,a} f_{m,b}^T \\
&\quad + \sum_{(a,a) \in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai} T_{m,ai} - 1\right) f_{m,a} f_{m,a}^T,
\end{aligned}$$

where $A_{i,j} = \{(a,a) : i \le a \le j\}$.

From Lemma 11, we have $f_{m,a} f_{m,b}^T \perp f_{m,c} f_{m,d}^T$ for $(a,b) \ne (c,d)$ in the Frobenius inner product. Moreover, $\|f_{m,a} f_{m,b}^T\|_F = 1$ for all a, b. Hence,

$$\begin{aligned}
&\|E_k E_k^T - F_{m,k}F_{m,k}^T\|_F^2 \\
&= \sum_{(a,b) \in \{1,\ldots,n\}^2 \setminus A_{1,k}} \left(\sum_{i=1}^k T_{m,ai} T_{m,bi}\right)^2 + \sum_{(a,a) \in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai} T_{m,ai} - 1\right)^2. \\
&= \sum_{(a,b) \in \{1,\ldots,n\}^2 \setminus A_{1,n}} \left(\sum_{i=1}^k T_{m,ai} T_{m,bi}\right)^2 + \sum_{(a,a) \in A_{k+1,n}} \left(\sum_{i=1}^k T_{m,ai} T_{m,ai}\right)^2 \\
&\quad + \sum_{(a,a) \in A_{1,k}} \left(\sum_{i=1}^k T_{m,ai} T_{m,ai} - 1\right)^2. \quad (7)
\end{aligned}$$

As noted above, $T_{m,ij} \to 0$ as $m \to \infty$ for $j \leq k$, $i \geq k+1$, or $j \geq k+1$, $i \leq k$. Using this and orthogonality of T_m gives

$$\sum_{i=1}^{k} T_{m,ai}T_{m,bi} = \sum_{i=1}^{n} T_{m,ai}T_{m,bi} - \sum_{i=k+1}^{n} T_{m,ai}T_{m,bi} \to \begin{cases} 0 & \text{if } a \neq b, \\ 1 & \text{if } a = b, \end{cases}.$$

Inserting into (7) results in $\|E_k E_k^T - F_{m,k}F_{m,k}^T\|_F^2 \to 0$, as $m \to \infty$. □

Lemma 10. *Let A be an $a \times b$ matrix. Let U be a $b \times b$ orthogonal matrix and V an $a \times a$ orthogonal matrix. Then,*

$$\|A\|_F = \|VA\|_F = \|AU\|_F = \|VAU\|_F.$$

Proof. See Golub and Loan (2013). □

Lemma 11. *Let $w,x,y,z \in \mathbb{R}^b$. For $a \times b$-matrices A and B, let $\langle A,B\rangle_F = \sum_{i=1}^{a}\sum_{j=1}^{b} A_{ij}B_{ij}$ be the Frobenius inner product of A and B, and let $\langle \cdot,\cdot \rangle$ be the standard inner product on $\mathbb{R}^b$. Then, $\langle wx^T, yz^T\rangle_F = \langle w,y\rangle \langle x,z\rangle$. In particular, $\|wx^T\|_F = \|w\|_2\|x\|_2$ and $wx^T \perp yz^T$ if $w \perp y$ or $x \perp z$.*

Proof. Note that

$$\left\langle wx^T, yz^T \right\rangle = \sum_{i=1}^{b}\sum_{j=1}^{b} w_i x_j y_i z_j = \sum_{i=1}^{b} w_i y_i \sum_{j=1}^{b} x_j z_j = \langle w,y\rangle \langle x,z\rangle.$$

Hence, if either $w \perp y$ or $x \perp z$, then $wx^T \perp yz^T$, and $\|wx^T\|_F^2 = \langle wx^T, wx^T\rangle = \langle w,w\rangle\langle x,x\rangle = \|w\|_2^2\|x\|_2^2$, such that $\|wx^T\|_F = \|w\|_2\|x\|_2$. □

Lemma 12. *Let $v_1,\ldots,v_\ell$ be linearly independent vectors. An orthogonal projection matrix on span$(v_1,\ldots,v_\ell)$ has Frobenius norm $\sqrt{\ell}$.*

Proof. We may assume that $v_1,\ldots,v_\ell$ are orthonormal. Then, we can write the projection matrix as $P = v_1v_1^T + \ldots + v_\ell v_\ell^T$. By Lemma 11, $v_iv_i^T \perp_F v_jv_j^T$ for $i \neq j$. So, again by Lemma 11, $\|P\|_F^2 = \|v_1v_1^T\|_F^2 + \ldots + \|v_\ell v_\ell^T\|_F^2 = \ell$. □

Lemma 13. *Let A be an $a \times b$ matrix and B an $b \times c$ matrix, both of rank b, such that $a,c \geq b$. Let $C = AB$. Then, C is of rank b, and the row space of C coincides with the row space of B.*

Proof. First we show that rank$(C) = b$. Note that A has b linearly independent rows $1 \leq i_1 < \ldots < i_b \leq b$, and B has b linearly independent columns $1 \leq j_1 < \ldots < j_b \leq b$. Let $\widetilde{A}$ and $\widetilde{B}$ be the $b \times b$ matrices with $\widetilde{A}_{cd} = A_{i_c d}$ and $\widetilde{B}_{cd} = B_{c j_d}$. Then $\widetilde{A}$ and $\widetilde{B}$ are invertible matrices. Hence, also $\widetilde{C} = \widetilde{A}\widetilde{B} = (C_{i_a j_b})_{a,b}$ is invertible and has rank k. It follows that C has rank k. As $\widetilde{A}$ is invertible, then the span of the rows of $\widetilde{A}B$ is equal to the span of the rows of B. That is, the span of the rows of AB is equal to the span of the rows of B. □

Literature cited

Alexander DH, Lange K. 2011. Enhancement of the admixture algorithm for individual ancestry estimation. BMC Bioinformatics. 12:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR *et al*. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 96:3–12.

Box G, Hunter J, Hunter W. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley Series in Probability and Statistics. Wiley.

Cabreros I, Storey J. 2019. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. Genetics. 212:1009–1029.

Chen X, Storey J. 2015. Consistent estimation of low-dimensional latent structure in high-dimensional data.

Conomos M, Reiner A, Weir B, Thornton T. 2016. Model-free estimation of recent genetic relatedness. Am J Hum Genet. 98:127–148.

Engelhardt B, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genetics. 6.

Evanno G, Regaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: A simulation study. Mol Ecol. 14:2622–2620.

Garcia-Erill G, Albrechtsen A. 2020. Evaluation of model fit of inferred admixture proportions. Molecular Ecology Resources. 20:936–949.

Golub GH, Loan CF. 2013. *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences. JHU Press.

Jacod J, Protter P. 2004. *Probability Essentials*. Universitext. Springer.

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017. The *k*02 conundrum. Mol Ecol. 26:3594–3602.

Jolliffe IT. 2002. *Principle Component Analysis (2nd Ed.)*. Springer Series in Statistics. Springer.

Jolliffe T, Cadima J. 2016. Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A. 374:0150202.

Lawson D, van Dorp L, Falush D. 2018a. A tutorial on how not to overinterpret structure and admixture bar plots. Nature Communications. 9.

Lawson DJ, van Dorp L, Falush D. 2018b. A tutorial on how not to over-interpret structure and admisture bar plots. Nat Comm. 19:3258.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M *et al*. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 513:409–413.

Meisner J, Liu S, Huang M, Albrechtsen A. 2021. Large-scale inference of population structure in presence of missingness using PCA. Bioinformatics. 37:1868–1875.

Ochoa A, Storey JD. 2019. f_{ST} and kinship for arbitrary population structures i: Generalized definitions. bioRxiv. .

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genetics. 2:e190.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genetics. 8:1–17.

Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Raj A, Stephens M, Pritchard J. 2014. Faststructure: Variational inference of populations structure in large snp data sets. Genetics. 197:573–589.

Wang J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics. 154:747 –765.

Wang J. 2019. A parsimony estimator of the number of populations froma structure-like analysis. Mol Ecol Res. 19:970 –981.

Chapter 4

Manuscript 3

Estimating Allele Frequency and Recent Kinship Coefficient

Song Li[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark.

Publication details: Both manuscript and results are preliminary.

Estimating Allele Frequency and Recent Kinship Coefficient

Song Li*
Department of Mathematical Sciences, University of Copenhagen, Denmark

Abstract

The kinship of individuals inferred from genetic data have important applications in population genetics. The kinship coefficient between individuals in a pedigree is a fundamental concept that quantified the probability of sharing ancestral alleles potentially in the presence of complex population structure. This paper provides a new estimation formula for the kinship coefficient and designs a procedure to obtain the data set corresponding to the studied individual from the given sample, which can be used in the principal component analysis (PCA) method to estimate the allele frequency of the studied individual present in the formula. For the case of full-siblings, the validity of the estimation is verified in simulated scenarios.

Keywords: Population genetics; Kinship coefficient; Population structure; PCA; Allele frequency.

1 Introduction

Advances in array-based genotyping technology have enable researchers to obtain a large amount of genotype data. The information contained in the data has led to the development of a number of methods for inferring the ancestral origin of genes. Trying to establish the relationship between individuals and their ancestors through the genetic information they carry is a central issue. In population genetics, mathematical tools are used to infer whether individuals come from a homogeneous or structured population by looking for evidence in the data. However, existing population structure inference tools are mainly developed for individuals that are assumed to be unrelated. Genetic studies include related individuals, which motivates efforts to quantify relatedness between individuals. Any alleles that are inherited copies of a common ancestral allele are said to be identical by descent (IBD). The coefficients developed with the concept of IBD become measures of

*Corresponding author.
E-mail address: song.li@math.ku.dk

relatedness, where kinship coefficients are fundamental and depend on the choice of specific pedigree (Speed and Balding, 2015). When pedigree information is limited or unavailable, some kinship coefficient estimation formulas based on genotype data are proposed instead. The estimators obtained by considering the likelihood function for unlinked loci show good results in terms of bias in cases such as parent-offspring and full-siblings (Anderson and Weir, 2007). The estimators obtained by calculating the moments are widely used for the simplicity and effectiveness in large datasets. An estimator called KING-robust is capable of deriving inference about the relationship of any pair of individuals, independent of sample composition or population structure (Manichaikul et al., 2010). Different from the KING-robust "one-step" estimation, some methods have been proposed to consider using the genotype data to obtain the estimated values such as individual-specific allele frequency, and then to calculate the moments, which can be said to be "two-step" estimation. The estimated values of the individual-specific allele frequency is derived from the inference of population structure, which is commonly done by the clustering and PCA. Clustering algorithms, such as STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009), basically estimate the admixture proportion and allele frequencies in ancestral populations. The two clustering algorithms require appropriate reference population panels for the ancestries, which is also applicable to the REAP estimator of the kinship coefficient. The REAP estimator uses a model-based population structure analysis method to make individual-specific allele frequency estimable and thus realize the calculation of the moments (Thornton et al., 2012). The PCA method is proposed to utilize the first few principal components (PCs) to estimate allele frequencies (Hao et al., 2016). The usual practice of PCs extraction is for unrelated individuals, which is also critical to the PC-Relate estimator. Instead of using external reference population panels, the PC-Relate method estimates allele frequencies for all individuals using the top PCs extracted from a set of unrelated individuals separated from the sample according to the KING-robust estimator, and then computes the moments (Conomos et al., 2016).

In PC-Relate, all individuals which are unrelated to each other are separated from the sample containing related individuals and then are used to extract their top PCs, and next the PC values of the remaining individuals in the sample are predicted based on these top PCs, thus the complete top PCs of all individuals is obtained, so as to estimate individual-

specific allele frequency through the linear regression of the complete top PCs. Inspired by this, we design a procedure for each studied individual to obtain the corresponding data set similar to that in PC-Relate, and use PCA to estimate allele frequencies.

In this paper, we propose a formula to estimate the recent kinship coefficient in the presence of population structure. Similar to PC-Relate, we consider a strategy for building the data set of each studied individual, followed by a PCA method proposed by Chen and Storey (2015) on the data set to directly estimate individual-specific allele frequency present in the the recent kinship coefficient formula. The validity of the estimate is verified in full-siblings case by simulated data given different population structure contexts. The paper is structured as follows. The mathematical expression and estimation method of recent kinship coefficient are described in Section 2. In Section 3, we provide the process of generating the simulated sample data and analyze some future work directions. Proof, derivative and calculation are collected in the Appendix.

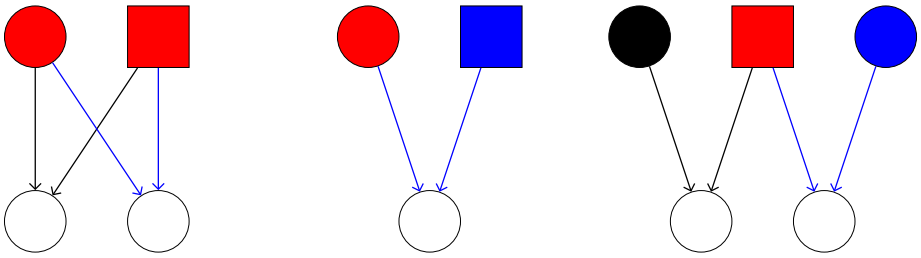
2 Material and Methods

2.1 Recent Kinship

We introduce the concept of IBD to describe possible scenarios of two alleles within and between individuals. For a diploid individual i, we define the probability that a pair of alleles from i are IBD as ψ_i, also known as the coefficient of inbreeding; and for two individuals i and j, we define the probability that a randomly selected allele from i and a randomly selected allele from j are IBD as φ_{ij}, also known as the coancestry of two individuals or the coefficient of kinship. If two alleles are randomly selected from the same individual i at one locus, we denote the probability that such two alleles are IBD as φ_{ii}, which is called the coefficient of self-kinship, i.e., $\varphi_{ii} = (1 + \psi_i)/2$.

Suppose N individuals and M loci are studied. A specific SNP site on the homologous chromosome is labeled as $s \in \{1, 2, \ldots, M\}$ and $y_i^s \in \{0, 1, 2\}$ $(i = 1, 2, \ldots, N)$ is a stochastic variable counting the number of given reference alleles in the ith individual at locus s, which is called the genotype value. We break y_i^s down into two dichotomous variables, namely, $y_i^s = y_{i,1}^s + y_{i,2}^s$, where $y_{i,a}^s \in \{0, 1\}$ $(a = 1, 2)$ indicates whether the first or second allele is the reference allele or not.

In the common pedigree-based structure, there are some basic cases such as full-siblings, parent-offspring and half-siblings (see Figure 1). The allele at each locus is passed to the child by a parent who is also the founder of the basic pedigree. More complex pedigrees are composed of these basic structures, and a more general sketch is formed when all the founders of a pedigree are grouped together (see Figure 2). We give the following mathematical framework.



(a) Full-siblings **(b)** Parent-offspring **(c)** Half-siblings

Figure 1: Three basic pedigree-based structures. Red, black and blue represent the shared founders of two studied individuals, the other founders belonging to corresponding individuals, respectively. For parent-offspring, red also represents one of the individuals studied.

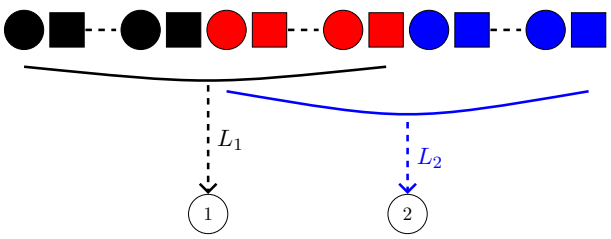


Figure 2: Red, black and blue represent the shared founders of individual 1 and 2, the other founders of individual 1 and individual 2, respectively. Individual 1 has L_1 founders through generation forward, and individual 2 has L_2 founders.

Assume that the pedigree founder l_i who is the recent ancestor of individual i has the studied allele with probability $\pi_{l_i}^{(i)s} \in [0,1]$ at locus s, where $l_i = 1, \ldots, L_i$, L_i is the number of i's founders. We define the following probability

$$\mathbf{Pr}\left(y_{i,a}^{s} = 1\right) = \sum_{l_i=1}^{L_i} w_{il_i} \pi_{l_i}^{(i)s}, \quad a = 1, 2, \tag{1}$$

where w_{il_i} represents the probability that individual i copied the studied allele from founder l_i and is called the weight. The sum of non-negative weights is required to be 1, i.e., $\sum_{l_i=1}^{L_i} w_{il_i} = 1$. (1) can be interpreted to mean that individual-specific allele frequency is the weighted average value across frequencies of founders corresponding to the individual. The generalization of individual-specific allele frequency has two intuitive examples: **i)** if each founder l_i has $\pi_{l_i}^{(i)s} = \pi^s$ as its allele frequency, then (1) becomes $\mathbf{Pr}\left(y_{i,a}^s = 1\right) = \sum_{l_i=1}^{L_i} w_{il_i}\pi^s = \pi^s$; **ii)** if founders are weighted equally, then (1) becomes

$$\mathbf{Pr}\left(y_{i,a}^s = 1\right) = \frac{1}{L_i}\sum_{l_i=1}^{L_i} \pi_{l_i}^{(i)s} := \pi^{(i)s}. \tag{2}$$

In general, weights allow adjustment for samples in various realistic scenarios. In the following, we use equal weights to explore the recent kinship.

As shown in Figure 2, we can set that the pedigree founders of individuals 1 and 2 have the studied allele with probabilities $\pi_{l_1}^{(1)s}$ and $\pi_{l_2}^{(2)s}$ at s SNP, respectively, where $l_1 \in \{1, \ldots, L_1\}$, $l_2 \in \{1, \ldots, L_2\}$. Let C be the number of shared founders from whom individuals 1 and 2 obtain the IBD allele, where $0 \le C \le \mathbf{min}[L_1, L_2]$. For $C > 0$, assume that $\pi_c^{(1,2)s}$ is defined as the allele frequency for the shared founder c, $c = 1, \ldots, C$; for $C = 0$, one says that two individuals are unrelated, i.e., $\varphi_{12} = 0$, $\pi_0^{(1,2)s} = 0$.

The kinship coefficient formula of individuals 1 and 2 at site s is obtained (details see Appendix A),

$$\varphi_{12}^s = \frac{1}{4} \cdot \frac{\mathbf{Cov}\left(y_1^s, y_2^s\right)}{\frac{1}{C}\sum_{c=1}^{C} \pi_c^{(1,2)s}\left(1 - \pi_c^{(1,2)s}\right)}. \tag{3}$$

Considering the fact that the kinship coefficient is constant across all SNPs, then the equation becomes,

$$\begin{aligned} \varphi_{12} &= \frac{1}{4M} \cdot \sum_{s=1}^{M} \frac{\mathbf{Cov}\left(y_1^s, y_2^s\right)}{\frac{1}{C}\sum_{c=1}^{C} \pi_c^{(1,2)s}\left(1 - \pi_c^{(1,2)s}\right)} \\ &= \frac{1}{4M} \cdot \sum_{s=1}^{M} \frac{\mathbb{E}\left[\left(y_1^s - 2\pi^{(1)s}\right)\left(y_2^s - 2\pi^{(2)s}\right)\right]}{\frac{1}{C}\sum_{c=1}^{C} \pi_c^{(1,2)s}\left(1 - \pi_c^{(1,2)s}\right)}. \end{aligned} \tag{4}$$

Equation 4 is presented as a mathematical expression of the kinship coefficient based on pedigree and genotype. When the parameters in the equation are estimated, we can measure the kinship coefficient of a pair of individuals. For parameters $\pi^{(1)s}$ and $\pi^{(2)s}$, we can use the PCA method introduced in the Section 2.2 to estimate. For the denominator part

consisting of parameters $\pi_c^{(1,2)s}$ and C, due to the absence of genotype information from ancestors or founders, it is difficult to estimate this part. In the Section 2.3, we consider a substitution for the denominator part in the **Full-siblings** case to present the estimate.

2.2 Estimation of Individual-specific Allele Frequency

In this section, we elaborate on the method of the consistent estimation of the latent linear space introduced by Chen and Storey (2015) to estimate individual-specific allele frequency.

Denote $\mathbf{Y} = \left(y_i^s\right)_{M\times N}$ as a matrix of observed variables and $\mathbf{\Theta} = \mathbb{E}\left[\mathbf{Y} \mid \mathbf{Q}\right] = 2\mathbf{FQ} = \left(\theta_i^s\right)_{M\times N}$ as the expectation, where $\mathbf{Q} = \left(q_{ki}\right)_{K\times N}$ is a matrix of K latent variables and $\mathbf{F} = \left(f_k^s\right)_{M\times K}$ is a matrix of parameters relating the latent variables to the observed variables. We define $\mathbf{D}_M = M^{-1}(\mathbf{Y}-\mathbf{\Theta})^T(\mathbf{Y}-\mathbf{\Theta})$ and denote $d_{M,ij} = M^{-1}\sum_{s=1}^{M}\left(y_i^s-\theta_i^s\right)\left(y_j^s-\theta_j^s\right)$ as the (i,j)th entry of $\mathbf{D}_M$. Let $\bar{\delta}_{M,ij} = M^{-1}\sum_{s=1}^{M}\mathbf{Cov}\left[y_i^s, y_j^s \mid \mathbf{Q}\right]$ be the column-wise average covariance. Theorem 1 shows a convergence result.

Theorem 1. *Assume independence of genotypes at different SNPs and uniformly bounded 4th conditional moments of $y_i^s - \theta_i^s$ are satisfied. Then*

$$\lim_{M\to\infty}\left|d_{M,ij} - \bar{\delta}_{M,ij}\right| = 0 \ \ a.s. \ .$$

Theorem 1 implies that the estimation of $\bar{\delta}_{M,ij}$ can be applied equally to $d_{M,ij}$, resulting in an estimate of $\mathbf{D}_M$. We introduce the concept of admixture and define individual-specific allele frequency π_i^s as follows,

$$\pi_i^s = \sum_{k=1}^{K} f_k^s q_{ki}. \tag{5}$$

In admixture model, $q_{ki}, f_k^s \in [0,1]$ are the admixture proportion of individual i from the ancestral population $k \in \{1,2,\cdots,K\}$ and the the reference allele frequency of the ancestral population k at SNP s, respectively. Let K be the number of ancestral populations and $\mathbf{\Pi} = \left(\pi_i^s\right)_{M\times N} = \mathbf{FQ}$ be a matrix of individual-specific allele frequencies. Note that similar to the weight, $\sum_{k=1}^{K} q_{ki} = 1$. We are interested in considering a binomial distribution with the parameter π_i^s to describe the genotype variable y_i^s,

$$y_i^s \mid \pi_i^s \sim \mathbf{Bi}\left(2, \pi_i^s\right).$$

If the genotypes of two individuals i and j on the same SNP are unrelated, then $\bar{\delta}_{M,ij} = 0$ and $\bar{\delta}_{M,ii} = M^{-1}\sum_{s=1}^{M}\mathbf{Var}\left[y_i^s \mid \pi_i^s\right]$. According to Lemma 7 and 8 in Chen and Storey (2015), we set

$$\hat{\delta}_{M,i} := \frac{1}{M}\sum_{s=1}^{M} y_i^s\left(2 - y_i^s\right)$$

as an estimate of $\bar{\delta}_{M,ii}$. And by Theorem 1, let $\hat{\mathbf{D}}_M = \mathbf{diag}\left\{\hat{\delta}_{M,1}, \ldots, \hat{\delta}_{M,N}\right\}$ be the estimate of $\mathbf{D}_M$. To get a consistent estimator $\widehat{\mathbf{Q}}$ of the latent space $\mathbf{Q}$ that spans $\mathbf{\Pi}$, one can consider the following eigenvalue decomposition when M is large enough,

$$\frac{1}{M}\mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{D}}_M = \widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^T, \tag{6}$$

where $\widetilde{\boldsymbol{\Sigma}}$ is a diagonal matrix whose diagonal elements are eigenvalues, and $\widetilde{\boldsymbol{V}} = \left(\tilde{v}_{ij}\right)_{N\times N}$ is an orthonormal matrix composed of corresponding eigenvectors. Denote

$$\widehat{\mathbf{Q}} = \widetilde{\boldsymbol{V}}_{1:K} = \left(\tilde{v}_{ij}\right)_{K\times N} \tag{7}$$

as the submatrix of eigenvectors corresponding to the top K positive eigenvalues. Using least square estimation $\hat{\mathbf{F}} = \underset{\mathbf{F}}{\operatorname{argmin}}\left\|\mathbf{Y} - 2\mathbf{F}\widehat{\mathbf{Q}}\right\| = \mathbf{Y}\widehat{\mathbf{Q}}^T\left(\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T\right)^{-1}/2$, we obtain an estimator of $\mathbf{\Pi}$,

$$\widehat{\mathbf{\Pi}} = \frac{1}{2}\mathbf{Y}\widehat{\mathbf{Q}}^T\left(\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T\right)^{-1}\widehat{\mathbf{Q}}. \tag{8}$$

Note that when $K < N$, (8) is reduced to

$$\widehat{\mathbf{\Pi}} = \frac{1}{2}\mathbf{Y}\widehat{\mathbf{Q}}^T\widehat{\mathbf{Q}}. \tag{9}$$

It can be seen from (7)-(9) that the allele frequency estimation is based on the combination of the top K principal components selected, which is regarded as a PCA method. We estimate allele frequencies of studied individuals based on the above PCA method, and different individuals have corresponding data sets used in PCA. We design the following procedure to divide the data set from the total sample.

Firstly, we calculate $\widehat{\kappa}_{ij}$ for all individual pairs in the sample according to the KING-robust estimator formula (Manichaikul et al., 2010),

$$\widehat{\kappa}_{ij} = \frac{\sum_{s=1}^{M}\left[y_i^s\left(1 - y_i^s\right) + y_j^s\left(1 - y_j^s\right) + y_i^s y_j^s\right]}{\sum_{s=1}^{M}\left[y_i^s\left(2 - y_i^s\right) + y_j^s\left(2 - y_j^s\right)\right]}. \tag{10}$$

Assume that individuals i and j are expected to be unrelated if $\widehat{\kappa}_{ij} \in [-0.025, 0.025]$ (Conomos et al., 2015). By determining $\widehat{\kappa}_{ij}$ for all pairs of individuals in the sample, all individuals which are unrelated to each other and do not include the two individuals under study are collected into a set called the unrelated individual set. In the unrelated individual set, we then reserve one sample that is related to the two individuals under study and all samples which are not related to the two individuals. Next, the two studied individuals are added to the unrelated individual set respectively to form two new data sets, which are used to estimate the allele frequency of the corresponding individual by the PCA method introduced above.

2.3 Estimation of Recent Kinship Coefficient

We show that the allele frequency of each individual in a pedigree also satisfies (5) form. The founder l_i belonging to individual i have

$$\pi_{l_i}^{(i)s} = \sum_{k=1}^{K} f_k^s q_{kl_i}^{(i)}, \tag{11}$$

where $q_{kl_i}^{(i)} \in [0,1]$ is the admixture proportion of founder l_i derived from the ancestral population k and $\sum_{k=1}^{K} q_{kl_i}^{(i)} = 1$. And then we substitute (11) into (1)

$$\begin{aligned} \mathbf{Pr}\left(y_{i,a}^s = 1\right) &= \sum_{l_i=1}^{L_i} w_{il_i} \sum_{k=1}^{K} f_k^s q_{kl_i}^{(i)}, \\ &:= \sum_{k=1}^{K} f_k^s \overline{q}_{ki}, \end{aligned} \tag{12}$$

where $\overline{q}_{ki} = \sum_{l_i=1}^{L_i} w_{il_i} q_{kl_i}^{(i)}$ and $\sum_{k=1}^{K} \overline{q}_{ki} = 1$. According to (2), we take the equal weights and rewrite $\pi^{(i)s}$ as π_i^s, then

$$\pi_i^s = \sum_{k=1}^{K} f_k^s q_{ki} \quad \left(q_{ki} = \frac{1}{L_i} \sum_{l_i=1}^{L_i} q_{kl_i}^{(i)}\right)$$

is individual-specific allele frequency which can be estimated by the above PCA method . We denote $\widehat{\pi}_i^s$ as the estimated value. Taking **Full-siblings** case in Figure 1, we come up with an estimation of the kinship coefficient according to equation (4),

$$\widehat{\varphi}_{ij}^F = \frac{\sum_{s=1}^{M} \left(y_i^s - 2\widehat{\pi}_i^s\right)\left(y_j^s - 2\widehat{\pi}_j^s\right)}{\sum_{s=1}^{M} \left[\left(y_i^s - 2\widehat{\pi}_i^s\right)^2 + \left(y_j^s - 2\widehat{\pi}_j^s\right)^2\right]}. \tag{13}$$

To get equation (13), we consider the following analysis

$$\begin{aligned}\mathbf{Var}(y_i^s) &= \mathbf{Var}\left(y^s_{c_1\to(i,1)}\right) + \mathbf{Var}\left(y^s_{c_2\to(i,2)}\right)\\ &= \pi^s_{c_1}\left(1-\pi^s_{c_1}\right) + \pi^s_{c_2}\left(1-\pi^s_{c_2}\right),\end{aligned} \tag{14}$$

where, $y^s_{c\to(i,a)}$ means that the first or second allele at SNP s of individual i comes from the parent c and parent c_1, c_2 genotypes are independent of each other. As a result, in **Full-siblings** case we update equation (4)

$$\begin{aligned}\varphi_{ij} &= \frac{1}{4M}\cdot\sum_{s=1}^{M}\frac{\mathbb{E}\left[\left(y_i^s-2\pi_i^s\right)\left(y_j^s-2\pi_j^s\right)\right]}{\frac{1}{2}\left[\pi^s_{c_1}\left(1-\pi^s_{c_1}\right)+\pi^s_{c_2}\left(1-\pi^s_{c_2}\right)\right]}\\ &= \frac{1}{4M}\cdot\sum_{s=1}^{M}\frac{\mathbb{E}\left[\left(y_i^s-2\pi_i^s\right)\left(y_j^s-2\pi_j^s\right)\right]}{\frac{1}{4}\left[\mathbf{Var}(y_i^s)+\mathbf{Var}(y_j^s)\right]}\\ &= \frac{1}{M}\cdot\sum_{s=1}^{M}\frac{\mathbb{E}\left[\left(y_i^s-2\pi_i^s\right)\left(y_j^s-2\pi_j^s\right)\right]}{\mathbb{E}\left[\left(y_i^s-2\pi_i^s\right)^2+\left(y_j^s-2\pi_j^s\right)^2\right]}.\end{aligned} \tag{15}$$

The last item in (15) can be viewed as an average of the ratios. When genotype and allele frequency are replaced by the observed and estimated values, respectively, ignoring the expectation, if each ratio is biased then their average across locus will be also biased, even as $M\to\infty$ (Ochoa and Storey, 2021). If genetic linkage is not considered or genetic linkage exists but the effective number of independent SNPs is large enough, adjusting to the ratio of the two averages will make the estimated value perform better. According to Theorem 1, the expectation in two averages can be omitted, so (13) is the estimation form we propose.

The PC-Relate kinship estimator's consistency can not be shown in all kinds of population structure scenarios for related pairs of individuals, and only in discrete population substructure, this bias is small enough to ensure a consistent estimate (see **Appendix A** in Conomos et al. (2016)). With the advantage of the KING-robust estimator that only depends on the data, a set of unrelated individuals can be obtained in advance according to $\widehat{\kappa}_{ij}\in[-0.025, 0.025]$. Since PCA method is sensitive to the presence of related individuals, the PC-Relate program uses $\widehat{\kappa}_{ij}$ to extract a set of unrelated individuals from the sample when estimating allele frequencies. In the following section, we present simulations in a series of scenarios to verify the validity of the formula and estimation. For the convenience of comparing different estimation methods, the results of $\widehat{\varphi}^F_{ij}$, the KING-robust estimator and the PC-Relate kinship estimator in **Full-siblings** case are labeled as "phiF", "phiK" and "phiPC", respectively.

3 Simulation Studies

3.1 Simulated Data Setting

In pedigree, the founders are set as unrelated individuals and the following binomial distribution describes the genotype of the founder l,

$$y_l^s \mid \pi_l^s \sim \mathbf{Bi}\left(2, \pi_l^s\right), \tag{16}$$

where, $\pi_l^s = \sum_{k=1}^{K} f_k^s q_{kl}$. Each founder has an admixture history between K subpopulations and these subpopulations all descend from a common ancestral population. Then for the descendants in pedigrees, we assign the alleles from the corresponding founders to them as genotypes according to Mendelian laws of inheritance, and assume that different SNPs are independent. Specifically, we consider that the allele frequency p^s of the common ancestral population is derived from a uniform distribution $\mathbf{U}[0.1, 0.9]$, and the number of subpopulations that diverge from the ancestral population is $K = 3$. The population-specific allele frequency f_k^s is derived from $\mathbf{Beta}(\alpha_k^s, \beta_k^s)$ with $\alpha_k^s = p^s(1-\gamma_k)/\gamma_k$ and $\beta_k^s = (1-p^s)(1-\gamma_k)/\gamma_k$, where γ_k is a constant that refers specifically to the degree of population divergences (Balding and Nichols, 1995). Here, we set $\boldsymbol{\gamma} := (\gamma_1, \gamma_2, \gamma_3) = (0.05, 0.15, 0.25)$. Denote the admixture proportion vector of founder l as $\mathbf{q}_l := (q_{1l}, q_{2l}, q_{3l}) \sim \mathbf{Dirichlet}(\boldsymbol{\lambda}_l)$, where $\boldsymbol{\lambda}_l := (\lambda_{1l}, \lambda_{2l}, \lambda_{3l})$. The population structure depends on $\mathbf{q}_l$.

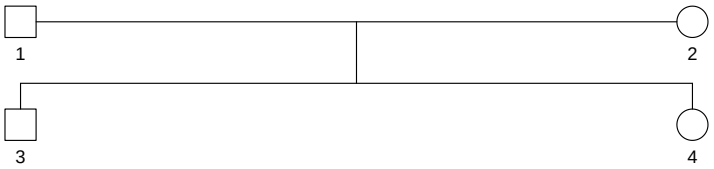


Figure 3: Simple pedigree configuration for the simulation studies

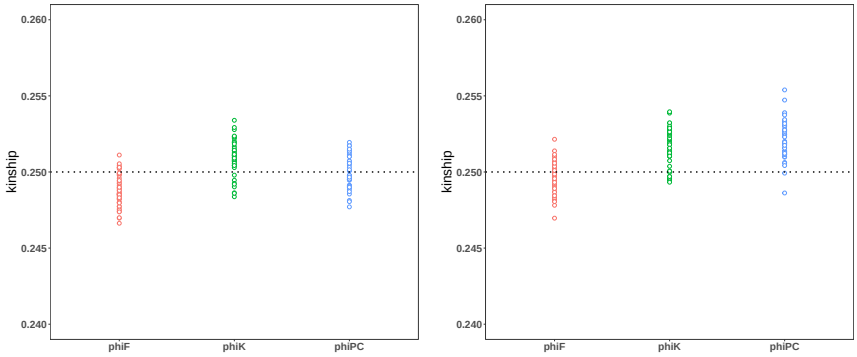
Figure 3 is the simple pedigree, which we use as the basic unit for building a pair of siblings. The following is the specific process of forming the sample set through simulation.

- Firstly, we simulate the genotypes of 80 unrelated individuals as the founder set according to model (16).
- Secondly, two of the 80 individuals are randomly selected as fixed parent 1 and 2 in Figure 3. We can get the genotypes of 50 siblings.
- Thirdly, one of the 50 siblings is randomly selected as fixed sibling 3 in Figure 3, and 49 cases of full-siblings are formed by sibling 3 and each of the remaining siblings who could be regarded as sibling 4 in Figure 3. The sibling 3 and 4 are combined with the founder set to form the sample set.

The two individual-specific allele frequencies are estimated by applying the method described in section 2.2 to the sample set and the proposed kinship coefficient estimator in **Full-siblings** case can be calculated from equation (13). With pedigree and $\mathbf{q}_l$, we give some scenarios. In **scenario 1**, we set $\boldsymbol{\lambda}_l = (6, 2, 0.25)$ to mean that, on average, ancestral contribution proportions of subpopulations 1, 2 and 3 to founder l are 0.73, 0.24 and 0.03, respectively. In **scenario 2**, we set $\boldsymbol{\lambda}_l = (1, 1, 1)$ to mean that, on average, each subpopulation has an equal ancestral contribution to founder l. For each scenarios, we set $M = 100,000$ independent SNPs for each individual. In principle, it is expected that the larger M is, the more accurate the estimate is.

3.2 Results and Discussion

We obtain the kinship coefficient estimation results from our proposed method, KING-robust and PC-Relate for full-siblings under scenario 1 and 2, which are shown in Figure 4 and Table 1.



(a) Scenario 1 (b) Scenario 2

Figure 4: The kinship coefficient estimation in full-siblings case.

Table 1: Comparison of kinship coefficient estimators in full-siblings case

Scenario	**Expected**	**phiF**[a]	**phiK**	**phiPC**
1	0.2500	0.2488(0.0010)[b]	0.2510(0.0011)	0.2500(0.0011)
2	0.2500	0.2496(0.0010)	0.2517(0.0012)	0.2521(0.0012)

[a] $\widehat{\varphi}_{ij}^{F}$, the KING-robust estimator and the PC-Relate kinship estimator are labeled as "phiF", "phiK" and "phiPC", respectively.

[b] The values presented in the table for each of the estimators are mean (standard deviation) of the estimated kinship coefficients.

PC-Relate has the smallest bias in the simulated scenario 1, while the estimator $\widehat{\varphi}_{ij}^{F}$ proposed by us has the smallest bias in the simulated scenario 2. In both scenarios, $\widehat{\varphi}_{ij}^{F}$ has the smallest variability. The effect of population structure on estimation can be observed from the setting of $\boldsymbol{\lambda}_l$. In scenario 1, it is more likely that the first two populations contribute alleles to the offspring, and the kinship coefficient estimated by PC-Relate is more accurate. In scenario 2, when all populations had equal contributions, the PC-Relate estimator shows a large bias, indicating that the PC-Relate estimator is sensitive to the population structure. In terms of sensitivity to the two population structures, both $\widehat{\varphi}_{ij}^{F}$ and the KING-robust are less than PC-Relate. In scenario 2, $\widehat{\varphi}_{ij}^{F}$ performs better than the KING-robust. $\widehat{\varphi}_{ij}^{F}$ tends to have a negative bias, and the other two tend to have a positive bias.

The results above inspire us to extend the approach to other relationship types, such as parent-offspring and half-siblings in Figure 1. Comparison of the differences between the proposed method and existing methods for different population structure backgrounds will be one of future work. The validity of the method should also be reflected in more simulated scenarios and real human data.

References

Doug Speed and David J. Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015. doi: 10.1038/nrg3821.

Amy D Anderson and Bruce S Weir. A Maximum-Likelihood Method for the Estimation of Pairwise Relatedness in Structured Populations. *Genetics*, 176(1):421–440, 05 2007. doi: 10.1534/genetics.106.063149.

Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010. doi: 10.1093/bioinformatics/btq559.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 06 2000. doi: 10.1093/genetics/155.2.945.

D. H. Alexander, J. Novembre, and K Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655—-1664, 2009. doi: 10.1101/gr.094052.109.

Timothy Thornton, Hua Tang, Thomas J. Hoffmann, Heather M. Ochs-Balcom, Bette J. Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012. ISSN 0002-9297. doi: 10.1016/j.ajhg.2012.05.024.

W Hao, M Song, and JD. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, 2016. doi: 10.1093/bioinformatics/btv641.

Matthew P. Conomos, Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1):127–148, 2016. ISSN 0002-9297. doi: 10.1016/j.ajhg.2015.11.022.

X. Chen and J.D. Storey. Consistent estimation of low-dimensional latent structure in high-dimensional data, 2015.

Matthew P. Conomos, Michael B. Miller, and Timothy A. Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, 2015. doi: 10.1002/gepi.21896.

Alejandro Ochoa and John D. Storey. Estimating FST and kinship for arbitrary population structures. *PLOS Genetics*, 17(1):e1009241, jan 2021. doi: 10.1371/journal.pgen.1009241.

David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1):3–12, 1995.

H. Walk. Strong laws of large numbers by elementary tauberian arguments. *Mh Math*, 144 (4):329–346, 2005. doi: 10.1007/s00605-004-0284-x.

Appendix A Kinship coefficient based on pedigree and genotype

A.1 Kinship coefficient calculation at one locus

In Figure 2, clearly that $\varphi_{12}^s = 2C/(2L_1 \cdot 2L_2) = C/(2L_1L_2)$ and we get

$$\mathbf{Pr}\left(y_{1,a}^s = 1\right) = \frac{1}{L_1}\sum_{l_1=1}^{L_1}\pi_{l_1}^{(1)s} := \pi^{(1)s},$$

$$\mathbf{Pr}\left(y_{2,a}^s = 1\right) = \frac{1}{L_2}\sum_{l_2=1}^{L_2}\pi_{l_2}^{(2)s} := \pi^{(2)s}, \quad a = 1, 2.$$

Then the following probability should be calculated to

$$\begin{aligned}
&\mathbf{Pr}\left(y_{1,a}^s = y_{2,a}^s = 1\right) \\
&= \varphi_{12}^s \frac{1}{2C}\left(2\sum_{c=1}^{C}\pi_c^{(1,2)s}\right) \\
&\quad + (1-\varphi_{12}^s)\frac{1}{(2L_1)\cdot(2L_2)-2C}\left[\left(2\sum_{l_1=1}^{L_1}\pi_{l_1}^{(1)s} - 2\sum_{c=1}^{C}\pi_c^{(1,2)s}\right)\left(2\sum_{l_2=1}^{L_2}\pi_{l_2}^{(2)s}\right)\right. \\
&\quad \left. +2\sum_{c=1}^{C}\pi_c^{(1,2)s}\left(2\sum_{l_2=1}^{L_2}\pi_{l_2}^{(2)s} - \pi_c^{(1,2)s}\right)\right] \\
&= \frac{\varphi_{12}^s}{C}\sum_{c=1}^{C}\pi_c^{(1,2)s} + \frac{1-\varphi_{12}^s}{2L_1L_2-C}\left[2\sum_{l_1=1}^{L_1}\pi_{l_1}^{(1)s}\sum_{l_2=1}^{L_2}\pi_{l_2}^{(2)s} - \sum_{c=1}^{C}\left(\pi_c^{(1,2)s}\right)^2\right] \\
&\overset{\varphi_{12}^s = C/(2L_1L_2)}{=\!=\!=} \frac{1}{L_1L_2}\sum_{l_1=1}^{L_1}\pi_{l_1}^{(1)s}\sum_{l_2=1}^{L_2}\pi_{l_2}^{(2)s} + \frac{1}{2L_1L_2}\sum_{c=1}^{C}\pi_c^{(1,2)s}\left(1-\pi_c^{(1,2)s}\right).
\end{aligned} \tag{17}$$

Clearly,

$$\begin{aligned}
\mathbb{E}\left(y_i^s\right) &= \mathbb{E}\left(y_{i,1}^s + y_{i,2}^s\right) \\
&= 2\mathbb{E}\left(y_{i,a}^s\right) \\
&= 2\mathbf{Pr}\left(y_{i,a}^s = 1\right), \quad i = 1, 2, \\
\mathbb{E}\left(y_1^s y_2^s\right) &= \mathbb{E}\left[\left(y_{1,1}^s + y_{1,2}^s\right)\left(y_{2,1}^s + y_{2,2}^s\right)\right] \\
&= 4\mathbb{E}\left(y_{1,a}^s y_{2,a}^s\right) \\
&= 4\mathbf{Pr}\left(y_{1,a}^s = y_{2,a}^s = 1\right).
\end{aligned}$$

And the covariance is obtained,

$$\begin{aligned}
\mathbf{Cov}\left(y_1^s, y_2^s\right) &= \mathbb{E}\left(y_1^s y_2^s\right) - \left(\mathbb{E}y_1^s\right)\left(\mathbb{E}y_2^s\right) \\
&= 4\left[\mathbf{Pr}\left(y_{1,a}^s = y_{2,a}^s = 1\right) - \mathbf{Pr}\left(y_{1,a}^s = 1\right)\mathbf{Pr}\left(y_{2,a}^s = 1\right)\right] \\
&= \frac{2}{L_1L_2}\sum_{c=1}^{C}\pi_c^{(1,2)s}\left(1-\pi_c^{(1,2)s}\right) \\
&\overset{(L_1L_2)^{-1} = 2\varphi_{12}^s/C}{=\!=\!=} \frac{4\varphi_{12}^s}{C}\sum_{c=1}^{C}\pi_c^{(1,2)s}\left(1-\pi_c^{(1,2)s}\right).
\end{aligned}$$

The kinship formula of individuals 1 and 2 at site s is obtained directly,

$$\varphi_{12}^{s} = \frac{1}{4} \cdot \frac{\mathbf{Cov}\left(y_1^s, y_2^s\right)}{\frac{1}{C}\sum_{c=1}^{C} \pi_c^{(1,2)s}\left(1-\pi_c^{(1,2)s}\right)}.$$

A.2 Kinship coefficient estimation

Given $s \in \{1, 2, \ldots, M\}$ and the fact that the kinship coefficient is constant over SNPs, for individuals i and j an average form of the kinship coefficient is the following,

$$\varphi_{ij} = \frac{1}{4M} \cdot \sum_{s=1}^{M} \frac{\mathbb{E}\left[\left(y_i^s - 2\pi^{(i)s}\right)\left(y_j^s - 2\pi^{(j)s}\right)\right]}{\frac{1}{C}\sum_{c=1}^{C} \pi_c^{(i,j)s}\left(1-\pi_c^{(i,j)s}\right)}. \tag{18}$$

Here, founder c is updated as the shared founder of individuals i and j. By Theorem 1 and convergence (Ochoa and Storey, 2021), an estimation form of the above kinship across M SNPs is the following,

$$\widehat{\varphi}_{ij} = \frac{1}{4} \cdot \frac{\sum_{s=1}^{M}\left(y_i^s - 2\widehat{\pi}^{(i)s}\right)\left(y_j^s - 2\widehat{\pi}^{(j)s}\right)}{\sum_{s=1}^{M}\frac{1}{C}\sum_{c=1}^{C} \widehat{\pi}_c^{(i,j)s}\left(1-\widehat{\pi}_c^{(i,j)s}\right)}, \tag{19}$$

where $\widehat{\pi}^{(i)s}, \widehat{\pi}^{(j)s}$ and $\widehat{\pi}_c^{(i,j)s}$ are the corresponding estimators.

Appendix B Proof of Theorem 1

In this appendix section, we set out to prove the theorem stated in the main paper.

Proof. Consider the $\mathbf{D}_M$, whose (i,j)th entry $d_{M,ij}$ can be written as

$$d_{M,ij} = M^{-1}\sum_{s=1}^{M} \tilde{d}_{s,ij},$$

where $\tilde{d}_{s,ij} = \left(y_i^s - \theta_i^s\right)\left(y_j^s - \theta_j^s\right)$. Denote $\tilde{y}_i^s = y_i^s - \theta_i^s$, $\tilde{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot \mid \mathbf{Q}]$ and $\tilde{\mathbb{V}}[\cdot] = \mathbb{V}[\cdot \mid \mathbf{Q}]$, where $\mathbb{E}$ and $\mathbb{V}$ are the expectation and variance operator, respectively. Clearly,

$$\tilde{\mathbb{E}}\left[\tilde{d}_{s,ij}\right] = \mathbf{Cov}\left[y_i^s, y_j^s \mid \mathbf{Q}\right].$$

By the independence of genotypes at different SNPs and uniformly bounded 4th conditional moments of $\tilde{y}_i^s$, we have

$$\begin{aligned}
\sum_{M\geq 1}\frac{1}{M}\tilde{\mathbb{V}}\left[d_{M,ij}\right] &= \sum_{M\geq 1}\frac{1}{M^3}\sum_{s=1}^{M}\tilde{\mathbb{V}}\left[\tilde{d}_{s,ij}\right] \\
&\leq \sum_{M\geq 1}\frac{1}{M^3}\sum_{s=1}^{M}\tilde{\mathbb{E}}\left[\tilde{d}_{s,ij}^2\right] \\
&\leq \sum_{M\geq 1}\frac{1}{M^3}\sum_{s=1}^{M}\left(\tilde{\mathbb{E}}\left[(\tilde{y}_i^s)^4\right]\right)^{1/2}\left(\tilde{\mathbb{E}}\left[(\tilde{y}_j^s)^4\right]\right)^{1/2} \\
&\leq C\sum_{M\geq 1}\frac{1}{M^2} = \frac{C\pi^2}{6},
\end{aligned} \tag{20}$$

where, C is a bounded constant. The second inequality in (20) is obtained by the Hölder's inequality. Denote

$$\bar{\delta}_{M,ij} := \frac{1}{M}\sum_{s=1}^{M}\tilde{\mathbb{E}}\left[\tilde{d}_{s,ij}\right] = \frac{1}{M}\sum_{s=1}^{M}\mathbf{Cov}\left[y_i^s, y_j^s \mid \mathbf{Q}\right].$$

Therefore, Theorem 1 of Walk (2005) implies that

$$\lim_{M\to\infty}\left|d_{M,ij} - \bar{\delta}_{M,ij}\right| = 0 \text{ a.s. }.$$

□