Estimating F-statistics in a probabilistic PCA space

Divyaratan Popli, Benjamin M. Peter

September 20, 2023

Abstract

Principal component analysis (PCA) and F-statistics are routinely used in population genetic and archaeogenetic studies. However, these are closely related analyses and reveal the same biological signal. Here, we present a statistical framework to combine them into a joint analysis. In particular, we discuss the differences of probabilistic PCA, Latent Subspace Estimation and ordinary PCA, and show that F-statistics are more naturally interpreted in a probabilistic PCA framework. We also show that individual-based F-statistics can be accurately estimated from probabilistic PCA in the presence of large amounts of missing data. We compare estimates from probabilistic PCA-based framework to ADMIXTOOLS 2 using simulations and published data, and show that this joint estimation framework addresses limitations of estimating F-statistics and PCA independently.

1 Introduction

Admixture between previously isolated populations is common in nature, and affects the patterns of genetic diversity. From these patterns of genetic diversity, one can reconstruct the past events of admixture.

1.1 Overview of methods to study admixture

There are several methods available to make inferences about admixture events that explain the observed genetic diversity. These methods can be classified as local or global ancestry based. Local ancestry based methods infer ancestry at each locus and reveal recent history of each individual, but have low power to infer events in deep past Vi et al. (2023); Brisbin et al. (2012); Price et al. (2009); Sankararaman et al. (2008). Commonly used methods to infer global ancestry are Principal Component Analysis (PCA) McVean (2009), MDS Wang et al. (2009), STRUCTURE Pritchard et al. (2000) and ADMIXTURE Alexander et al. (2009).

These are powerful methods to infer admixture, but can be difficult to interpret since they do not provide model comparisons or formal tests of admixture.

F-statistics are a popular way to infer global ancestry. They provide an intuitive and powerful way to test hypothesis of admixture by measuring the genetic drift shared between two, three, or four populations Patterson et al. (2012); Peter (2016). In this framework, the null model is represented by a tree connecting the populations, with the branch lengths representing genetic drift. The null hypothesis can be rejected when the variation in observed data can not be explained by a tree-like history.

It is worth noting that estimating F-statistics from data is not trivial. For large sample-size this is a minor issue, but for small sample-size, the necessary bias correction term can be substantialPatterson et al. (2012); Peter (2016). We provide a detailed definition of F-statistics, and the correction terms for sampling bias in section 2.1.

1.2 Estimation of F-statistics

The issue with estimation of F-statistics is further compounded because many species with distributions over large ranges, are not easily subdivided into discrete populations. For humans, it has been a long-standing debate on whether populations as such exist, and to what degree they are the result of biased sampling designs Serre and Pääbo (2004); Rosenberg et al. (2005); Peter et al. (2020).

In the context of F-statistics, this causes a trade-off between the urge to group as many individuals together as possible (which improves statistical accuracy of statistics, and can help with missingness), but can lead to an overly simplistic view of human genetic structure. On the flip-side, treating each individual independently, will lead to a more accurate description of the overall, individual-level genetic structure, but will be much harder to interpret and the accuracy of statistics will be lower.

These issues are further compounded if the data is low-coverage and of heterogeneous quality, as is common for ancient DNA data sets, where the amount of DNA preserved is often a limiting factor that precludes the generation of high-quality/high-coverage data. In this case, missingness can make individual-based statistics even harder to estimate.

For example, a conservative approach to estimate individual-based F-statistics is to only retain sites where data is present from every single individual in the data set. However, even for moderately large data sets that quickly becomes prohibitive: As a toy example, consider a data set with 100 (haploid) individuals with 10% missing data, and 1,000,000 sites. Out of those, only 26 would be covered in every single individual and could be used for multivariate analyses.

However, grouping individuals into populations can dramatically increase the number of sites retained. A common practice is to retain sites if at least one individual in each population

has data. In the above example, if we grouped the 100 individuals into 10 populations of 10 individuals each, and retained sites where at least one individual in each population carried data, we would expect no missing sites in almost all cases. However, grouping individuals may not be justified when they do not form discrete clusters, or when there are very few samples whose population assignments are unknown.

These issues can be resolved by estimating individual-based F-statistics with a framework that does not require a priori assignment of individuals to discrete populations, and is not sensitive to randomly missing data.

1.3 PCA

PCA is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional representation while retaining as much of the original variance as possible. It achieves this by finding orthogonal axes, called principal components (PC's), that capture the maximum variance in the data. It is commonly used to understand structure and admixture between populations Patterson et al. (2006); Novembre et al. (2008); noa; McVean (2009); Brisbin et al. (2012).

PCA was introduced and popularized as a tool to study human genetic structure by Cavalli-Sforza, and using just a handful of genetic loci, he was able to use PCA to accurately describe patterns of human genetic variation, and to make inference about their possible causes, although the way these patterns were analyzed were typically quantitative Menozzi et al. (1978); Sforza and Sforza (1995); noa.

As is still the norm with F-statistics, Cavalli-Sforza aggregated individuals into populations before performing PCA. With the advent of genomic data, his methods became superseded by individual-based approaches, which addressed many of the issues about grouping individuals we discussed above Patterson et al. (2006); Novembre et al. (2008); Price et al. (2006).

1.4 Probabilistic PCA and Latent Subspace Estimation (LSE)

Probabilistic PCA (PPCA) is an extension of PCA that incorporates a probabilistic framework Tipping and Bishop. PPCA models the observed data as generated by a linear transformation of a lower-dimensional latent space W with added Gaussian noise Ψ. A similar approach to model the observed data is Latent Subspace Estimation (LSE) Cabreros and Storey (2019). Here, the noise is modelled as binomially distributed, and so is independent for each individual van Waaij et al. (2023); Cabreros and Storey (2019); Chen and Storey (2015). We give a detailed explanation of both the frameworks in sections 2.3 and 2.4. In practice, PCA, PPCA and LSE differ in the way that they model the noise in the observed data due to sampling (see Fig.1).

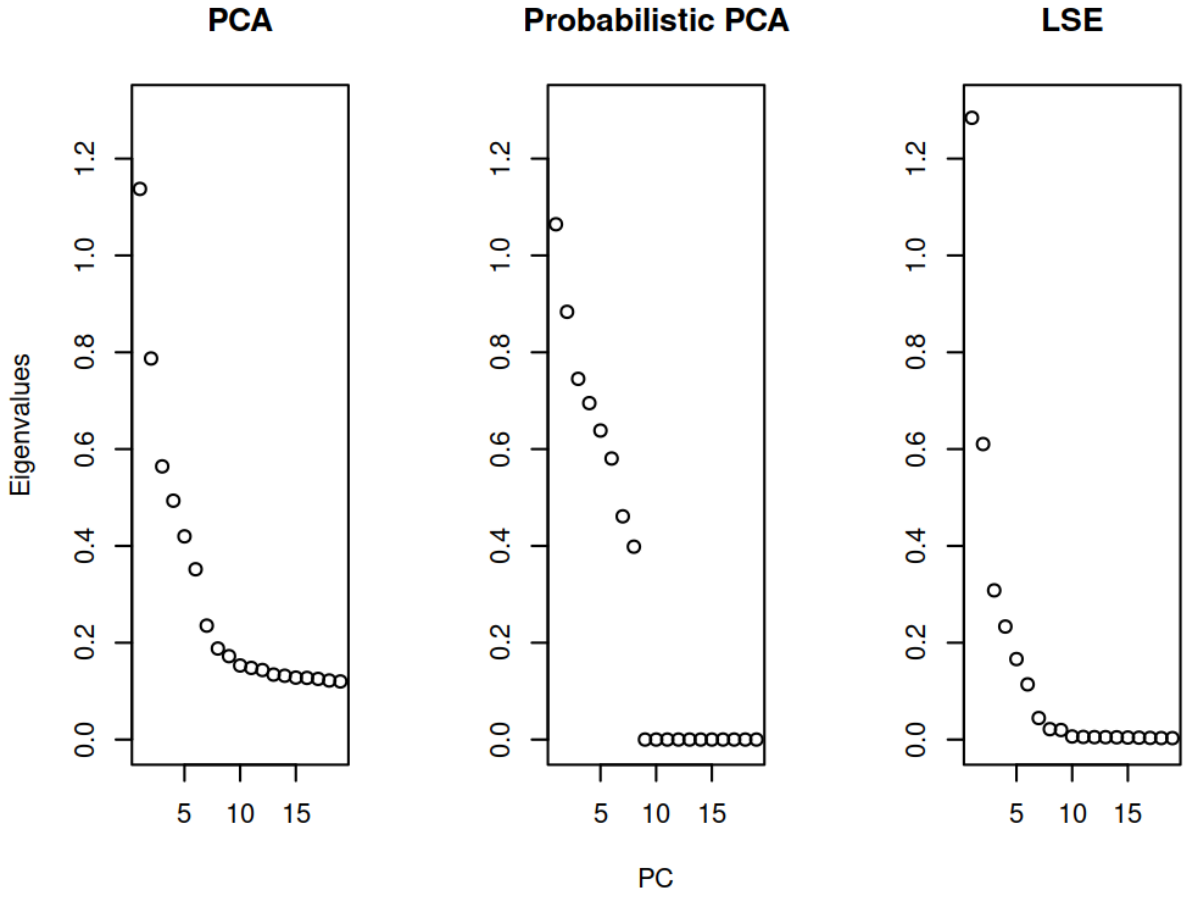

Figure 1: Comparison of PCA, PPCA and LSE: We simulated genotypes of 10 populations with 10 individuals each, and compared the top eigenvalues obtained from different PCA methods.

1.5 PCA and F-statistics

A common analysis paradigm in ancient DNA is to use PCA for exploratory and descriptive analyses, and then follow them up with methods based on F-statistic for a more formal treatment (typically in a third step, methods that synthesize many F-statistics are also applied, although we will not cover those here). Because they occur at different stages of the analyses, PCA and F-statistics use different data groupings and different normalizations, and are usually not directly compared.

Recently, we showed that the information contained in F-statistics and PCA is closely related, and that F-statistics can be interpreted geometrically in the context of PCA Peter (2022). For example, Fig. 2 illustrates that when allele frequencies are known, $F_2(X_1, X_2)$ can be thought of as the squared Eucledian distance between populations X_1 and X_2 on a PCA. Similarly, $F_3(X_1; X_3, X_4)$ is represented as the length of projection of the vector $X_1 - X_3$

on $X_1 - X_4$. Internal branch length $F_4(X_1, X_3; X_2, X_4)$ can be described as the length of projection of $X_1 - X_3$ on $X_2 - X_4$ on PCA, and the test of admixture $F_4(X_1, X_2, X_3, X_4)$ is equivalent to the length of projection of $X_1 - X_2$ on $X_3 - X_4$. We describe a formal relation between PCA and F-statistics (refer to Peter (2022) for a detailed derivation) in section 2.2, but here we point out that a joint framework to estimate PCA and F-statistics not only addresses the issue of population assignment in F-statistics, but also can potentially resolve some issues with the interpretations of PCA Novembre et al. (2008); noa; DeGiorgio and Rosenberg (2013); François et al. (2010).

However, while Peter (2022) develops a theoretical link between PCA and F-statistics, it does not develop a statistical framework to jointly estimate the two statistics from the data.

Here, we develop such a framework, and describe the impact of some of the choices of PCA-algorithm (classical PCA, PPCA, LSE). Using simulations, we show that using PPCA-based F-statistics can result in higher accuracy than using the naive estimators, especially when there is missing data. We also draw comparison between PPCA-based framework and ADMIXTOOLS 2 using published Neanderthal samples. At the end, we discuss that this approach not only solves issues related to F-statistics, but also is a step towards standardization and quantification of PCA methods.

2 Theory

2.1 F-statistics

We follow the original notation Patterson et al. (2012), and write the statistical estimates from empirical data as f_2, f_3 and f_4, and denote the theoretical values that depend only on the phylogenetic tree and the ascertainment scheme as F_2, F_3 and F_4. The three F-statistics are defined in terms of population allele frequencies as follows:

$$\begin{aligned} F_2(X_1, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(x_{1s} - x_{4s})^2 \\ F_3(X_1; X_3, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(x_{1s} - x_{3s})(x_{1s} - x_{4s}) \\ F_4(X_1, X_2; X_3, X_4) &= \frac{1}{S}\sum_{s=1}^{S}(x_{1s} - x_{2s})(x_{3s} - x_{4s}). \end{aligned} \tag{1}$$

Here, S is the total number of SNPs, and x_{is} is the (unobserved) population allele frequency

in population X_i at SNP s. Equivalently, we can also assign each individual to a separate population, and so these equations also hold for individuals.

Assuming a tree-like relationship between populations, $F_2(X_1, X_4)$ is interpreted as the branch length between populations X_1 and X_4 (Fig. 2 A) and it reflects the expected amount of drift that occurred between X_1 and X_4. $F_3(X_1; X_3, X_4)$ represents the amount of drift that occurred on the external branch connecting X_1 to the common ancestor node of X_3 and X_4 (Fig. 2 B). Under a tree-like model, F_3 will always be non-negative. However, in the case where X_1 is admixed between X_3 and X_4, $F_3(X_1; X_3, X_4)$ may be negative, and hence this is used as a test for admixturePeter (2016); Patterson et al. (2012).

$F_4(X_1, X_3; X_2, X_4)$ represents the covariance between shared drifts between X1,X2 and X3,X4. This would be represented by the internal branch between the common ancestor nodes of X_1, X_2 and X_3, X_4 (Fig. 2 C). F_4 statistic, with a different permutation of the populations can be used as test of admixture. $F_4(X_1, X_2; X_3, X_4)$ is expected to be 0 if X_1, X_2, X_3, X_4 are related to each other by a tree (Fig. 2 D). However, a large positive or negative value would suggest a departure from the null model of treeness due to gene flow between X1 and X3 or X2 and X3 respectively.

Since F_2 represents the branch length between a pair of populations, we can write F_3 and F_4 as linear combination of F_2's.

$$\begin{aligned} F_3(X_1; X_3, X_4) &= \frac{1}{2}[F_2(X_1, X_3) + F_2(X_1, X_4) - F_2(X_3, X_4)] \\ F_4(X_1, X_2; X_3, X_4) &= \frac{1}{2}[F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3)] \end{aligned} \tag{2}$$

Patterson et al. showed that the naive estimation of F_2 from *sample* allele frequency data will be biased, particularly when the sample size is small. They thus introduced the bias-corrected estimator Peter (2016); Patterson et al. (2012).

$$f_2(X_1, X_4) = \sum_{s=1}^{S}[(x_{1s} - x_{4s})^2 - \frac{x_{1s}(1 - x_{1s})}{n_{1s} - 1} - \frac{x_{4s}(1 - x_{4s})}{n_{2s} - 1}] \tag{3}$$

This sampling bias affects the estimates of F_3 as well but the correction terms cancel out for F_4.

$$f_3(X_1; X_3, X_4) = \sum_{s=1}^{S}(x_{1s} - x_{3s})(x_{1s} - x_{4s}) - \frac{x_{1s}(1 - x_{1s})}{n_{1s} - 1} \tag{4}$$

Assuming that the population allele frequencies are known, we can think of F-statistics in terms of Euclidean distances in an allele frequency space Oteo-García and Oteo (2021). In this framework, each population can be represented as a vector in a multi dimensional allele frequency space, and $F_2(X_1, X_4)$ can be calculated as squared Euclidean distance between the vectors (Fig. 2 E). $F_3(X_1; X_3, X_4)$ is then a dot product of the vectors $\vec{x1} - \vec{x3}$ and $\vec{x1} - \vec{x4}$ (Fig. 2 F). $F_4(X_1, X_3; X_2, X_4)$ is a dot product of vectors $\vec{x1} - \vec{x3}$ and $\vec{x2} - \vec{x4}$ (Fig. 2 G), and $F_4(X_1, X_2; X_3, X_4)$ is a dot product of vectors $\vec{x1} - \vec{x2}$ and $\vec{x3} - \vec{x4}$ (Fig. 2 H).

$$
\begin{aligned}
F_2(X_1, X_4) &= \frac{1}{S}||\vec{x_1} - \vec{x_4}||^2 \\
F_3(X_1; X_3, X_4) &= \frac{1}{S}||\vec{x_1} - \vec{x_3}) \cdot (\vec{x_1} - \vec{x_4})|| \\
F_4(X_1, X_2; X_3, X_4) &= \frac{1}{S}||(x_1 - x_2) \cdot (x_3 - x_4)||
\end{aligned}
\tag{5}
$$

2.2 PCA and F-statistics

This geometric framework provides an alternate way to understand the properties of F-statistics Oteo-García and Oteo (2021). However, many population genetic studies use a large number of SNPs (in the order of a million), and it is not possible to visualize population vectors in such a high dimensional space. Peter at al. showed that one can do dimentionality reduction on such datasets with PCA, and use the top PC's to estimate F-statistics Peter (2022).

Let us assume M populations and S SNPs, such that our dataset X has the dimension $[M \times S]$, and each entry of X is an allele frequency $\in$ [0,1]. PCA of mean-centered X allows us to project this $[M \times S]$ high dimensional data on a lower dimensional subspace $[M \times q]$. $q = M - 1$ represents the case where we retain all the PC's, and thus PCA only rotates X. However, in practice we often only need few PC's (q « M) to explain most variation in the populations Peter (2022). A common algorithm to estimate PC's is Singular Value Decomposition (SVD). For this approach we first mean-center X, and then decompose X into U, Σ, and V^T.

$$
Y = CX = (U\Sigma)V^T = WL,
$$

where C is the centering matrix such that $C = I_M - (1/M)J_M$. Here I_M is identity matrix of dimension M, and J_M is $M \times M$ matrix of all ones. We perform SVD to decompose Y into a product of $W = U\Sigma$ and $L = V^T$. In the context of PCA, $W_{M \times M}$ is a matrix of

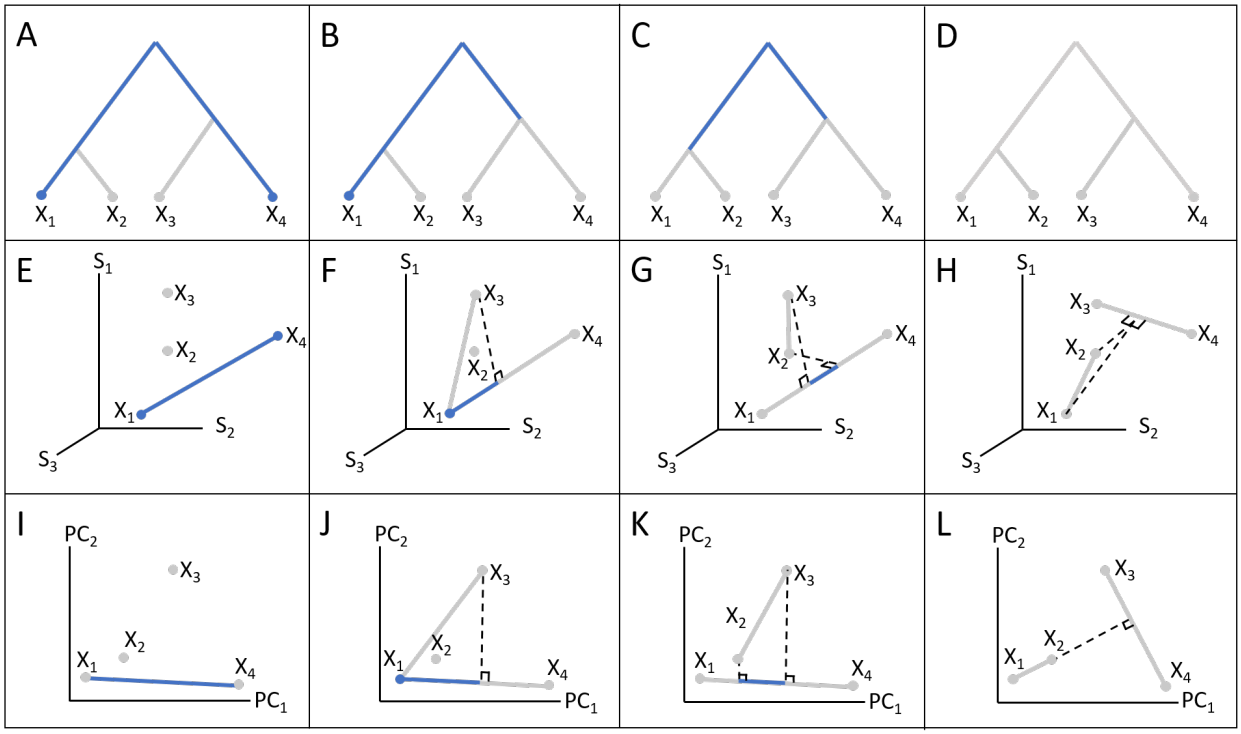

Figure 2: Schematics showing different interpretations of F-statistics. The columns represent $F_2(X_1, X_4)$, $F_3(X_1; X_3, X_4)$, $F4(X_1, X_3; X_2, X_4)$, $F4(X_1, X_2, X_3, X_4)$. The first row shows a tree interpretation of each statistic, the second row shows F-statistics in an allele-frequency space with three axes representing three SNPs, and the last row is the interpretation of F-statistics on a PCA. Blue lines represent the statistic, and the dotted lines represent orthogonal projections.

principal components and contains information about structure, while $L_{M\times S}$, also known as SNP loadings contains the contribution of each SNP to each PC, and can be used to identify outlier SNPs that may be potential candidates for selection Gower (1966).

In section 2.1 we showed that F-statistics can be written as the dot product of vectors in an allele frequency space. Dot product is invariant to PCA, which is just a rotation of the dataset to axes of largest variation. Therefore, we can calculate F-statistics from the PC's directly Peter (2022):

$$
\begin{aligned}
F_2(X_1, X_4) &= \sum_{s=1}^{S} (x_{1s} - x_{4s})^2 \\
&= \sum_{s=1}^{S} (x_{1s} - \mu_s)(x_{4s} - \mu_s) = F_2(Y_1, Y_4) \\
&= \sum_{s=1}^{q} (w_{1s} - w_{4s})^2 = F_2(W_1, W_4)
\end{aligned}
\tag{6}
$$

F_3 and F_4 can then be written in terms of F_2, and remain invariable to change in axes. For many applications, we only need few PC's with highest variation to approximate F-statistics Peter (2022).

2.3 PPCA

A difficulty in the practical application of this result is that the geometric considerations of Oteo-García and Oteo (2021) and Peter (2022) only hold for the (generally unobserved) population allele frequencies, but not for sample allele frequencies. In ancient DNA, PCA is most commonly run directly on individual-level genotype data Patterson et al. (2006), and hence on the biased sample allele frequencies. Thus, applying the PCA-based estimator (eq. 6) to calculate F-statistics would likewise result in biased estimate.

Oteo-García et al. resolved this by only calculating F_4 statistics, where the naive estimator is unbiased Oteo-García and Oteo (2021). In Peter (2022), unbiased estimates of the PCA reconstructions were obtained indirectly by first calculating all pairwise F_2-statistics, and then performing a multidimensional-scaling decomposition equivalent to PCA.

Here, we develop two related approaches that aim to calculate the bias-corrected estimates of F_2 from a PCA, by explicitly separating out the error in allele frequencies.

The first approach is based on probabilistic PCA (PPCA) Tipping and Bishop; Agrawal et al. (2020). PPCA is a dimensionality reduction technique that extends the classical PCA by introducing a probabilistic framework that allows for a homoskedastic error. Like PCA, it is a statistical model that assumes that the observed data is generated from a lower-dimensional latent space, but it adds a constant error term for each observation.

In classical PCA, the goal is to find a linear transformation (rotation) of the data that captures the maximum amount of variance in the original dataset. However, PCA does not provide a probabilistic interpretation of the data and does not explicitly model noise or uncertainty.

PPCA introduces a probabilistic generative model. It assumes that the observed data points are generated by a linear transformation of a lower-dimensional latent space, with an additional Gaussian noise term. The latent variables capture the underlying structure or patterns in the data, while the noise accounts for variability due to sampling error. Setting the Gaussian noise parameter to 0, converges PPCA to PCA.

PPCA models a latent structure of the form $X \sim N(WZ+\Psi I)$, where X is the observed data, W is a M x q matrix of linear mappings, Ψ is a Gaussian noise term, Z is a S-dimensional latent variable, and I is the identity matrix. Intuitively, WZ captures the covariance in the observed data, analogous to the F-statistics, and ΨI is analogous to the bias-correction term.

The goal of PPCA is to estimate the parameters of the model, namely W, μ, and Ψ, given the observed data. This is typically done through the standard expectation-maximization (EM) algorithm or maximum likelihood estimation (MLE).

2.4 LSE

LSE is a dimensionality reduction technique quite similar to PPCA, with the difference that LSE accounts for the heteroscedasticity in the data Chen and Storey (2015), and explicitly models the binomial error in genetic data. In this algorithm, we calculate the heterozygosity $d_{jj} = \frac{1}{S}\sum_i 2x_{ij}(1-x_{ij})$ from X. We define D as a diagonal matrix with j^{th} entry as δ_{jj}. We then calculate the eigenvalues of G $= \frac{1}{S}X^TX - D$. The eigenvectors of G then span the latent subspace of L , and the smallest $M-q$ eigenvalues converge to 0 for large M Cabreros and Storey (2019).

We show in the appendix that

$$\begin{aligned} f_2(X_1, X_4) &= \sum_s (x_{1s} - x_{2s})^2 - d_1 - d_2 \\ &= G_{11} + G_{44} - 2G_{14} \end{aligned} \tag{7}$$

It is worth noting that the sampling error terms in equation 2.1 are the same as the binomial error model of LSE, and so the distances based on LSE (using all components) are equivalent to those obtained from F-statistics (Fig. YY).

3 Results

We evaluate the performance of PCA, PPCA and LSE frameworks and compare to that of admixtools 2 using simulations, and a Neandertal dataset.

3.1 Evaluation on simulations

We simulated 10 populations with 10 individuals each using slendr Petr et al. (2022). We use a mutation rate of 10^{-8} per base per generation, a recombination rate of 10^{-8} per base per generation, and a generation time of 30 years. Fig. 3 shows the split times and migration events.

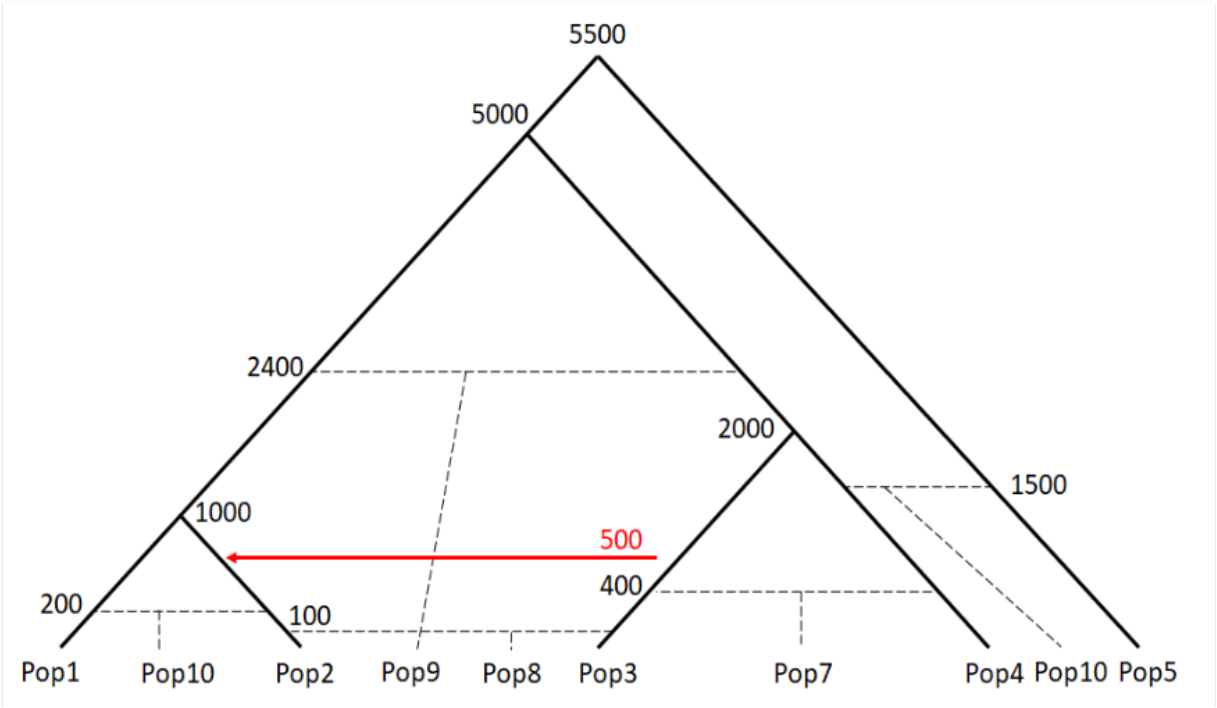

Figure 3: Simulated trees for testing our estimation of F-statistics. Split time for each node in generations and the admixture branches are shown in dashed lines. The red arrow represents a migration event from Pop3 to Pop2.

We use eq. 6 to calculate F_2 for a pair of populations using the squared Euclidean distance between them from the principal components (see Fig. S1) based on either PCA, PPCA and LSE, and compare F-statistics estimated from these methods to the true theoretical value of the statistic obtained from branch lengths in slendr Petr et al. (2022). We first do a comparison between f_2s estimated from PCA, PPCA and LSE with different number of PC's used (Fig. 4, Fig. S2). Here, we use 10 individuals in each population, and we see that all three methods are not sensitive to the number of PC's used, as long as the number of PC's is higher than 7. The estimate of F2 based on PCA is only slightly higher than that for PPCA and LSE, since sample sizes of 10 diploid individuals is large enough that we can ignore the sampling error correction. We next look at a comparison between the three methods using only one individual in each population (Fig. 4, Fig. S3). Here we observe that $F2$ estimates from PPCA and LSE are quite close to the true value at all cases with number of PCs greater than 7 and less than 75. However, estimates from PCA get increasingly higher if we add additional PC's. This is expected from theory (section 2), since PCA does not account for

sampling bias we end up with the biased estimator of F_2. We repeated the estimation of F_2s with PPCA and PCA in the presence of 50% missing data, and we see that in this case PPCA gives accurate results when the number of PCs used is higher than 7 and lower than 25, while PCA results are quite noisy. Implementation of LSE is not trivial when there is missingness in data, and is not included in this analysis.

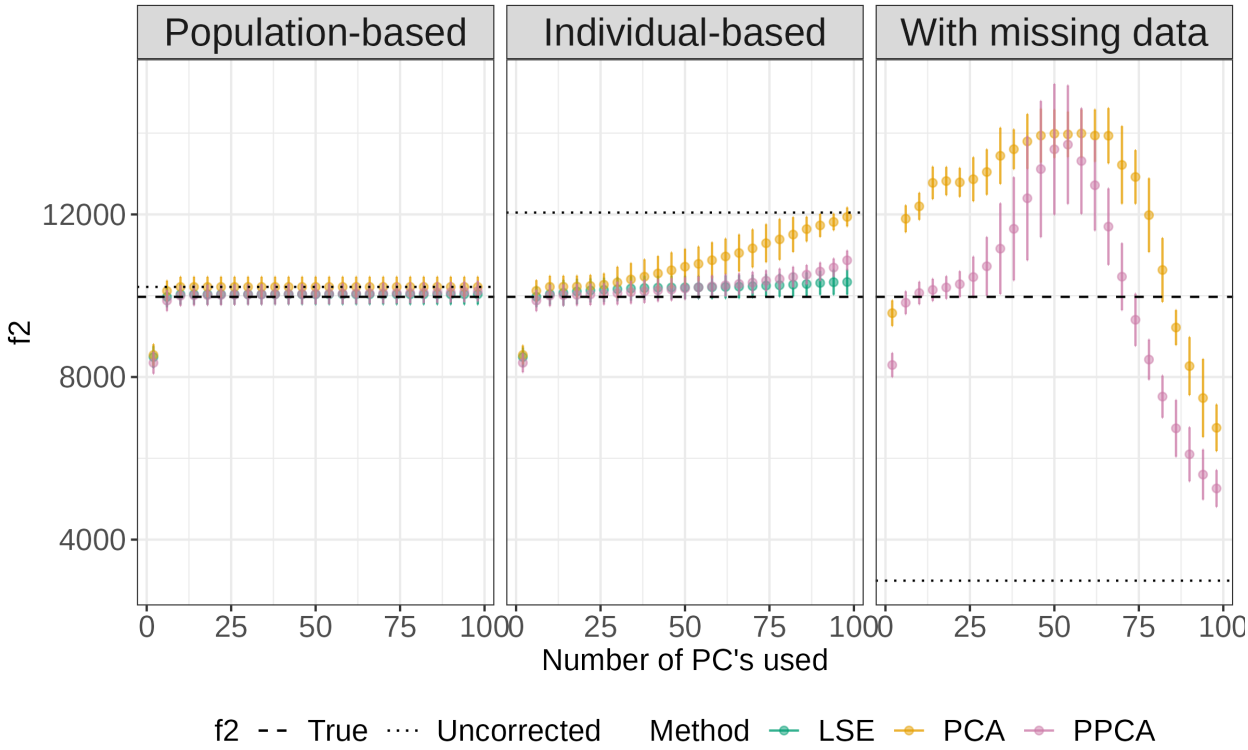

Figure 4: Comaprison of PCA approaches using $F_2(X_1, X_4)$ estimated using 10 individuals for each population (left), 1 individual for each populaltion (middle), and 1 individual for each population with 50% missing genotypes (right).

In the rest of the analyses, we exclude PCA, and compare PPCA and LSE with 8 and 12 PC's to ADMIXTOOLS 2 Maier et al. (2022), which is a recent reimplementation of ADMIXTOOLS Patterson et al. (2012). We chose 8 and 12 PCs because the number of PCs to use will not be known in most applications. We first compare f_2, f_3 and f_4 estimated by these methods in an ideal scenario, where each population has 10 individuals, and there is no missing data. In this case we find that the three methods perform well, and get F-statistics close to the truth (fig. S4).

We next address the issues listed in Introduction section 1.2. The first issue is about the estimation of F-statistics when population assignment is difficult, especially when few samples are available. We show that this can be resolved with individual-based F-statistics. In our simulations, we label each individual as a different population, and we sample one individual from each population to calculate F-statistics (fig. S5). We observe that in this case both PPCA and LSE- based frameworks perform atleast as well as ADMIXTOOLS 2. The mean

estimate from each method is close to the true value, however, the error bars for F_2 estimates are lower for PPCA compared to ADMIXTOOLS 2, specially for X1 and X2, which have low split-times. The improved accuracy of PCA-based tools versus ADMIXTOOLS 2 is explained because PCA incorporates a succinct summary of the full data of all the individuals, and thus the PCA-based estimates can "borrow" information from related individuals in the sample that are not used to calculate the statistic at hand. In contrast, ADMIXTOOLS 2 has only one individual from each population to assess structure / admixture, and while the estimates based on admixtools 2 are minimum-variance estimators for these subsets of the data (Patterson et al., 2012), PCA-based methods do better whenever we have data from additional individuals.

3.2 Missing data

Next, we address the issue of missing data and evaluate the estimation of these methods when there is random missing data. Our implementation of PPCA on missing data is inspired from EMU Meisner et al. (2021), and is described in Methods section 4.1. We use EMU software for inferring PC's for the calculation of F-statistics. We see that PPCA is not affected by missingness, while ADMIXTOOLS 2 and EMU results are inflated (Fig. 5).

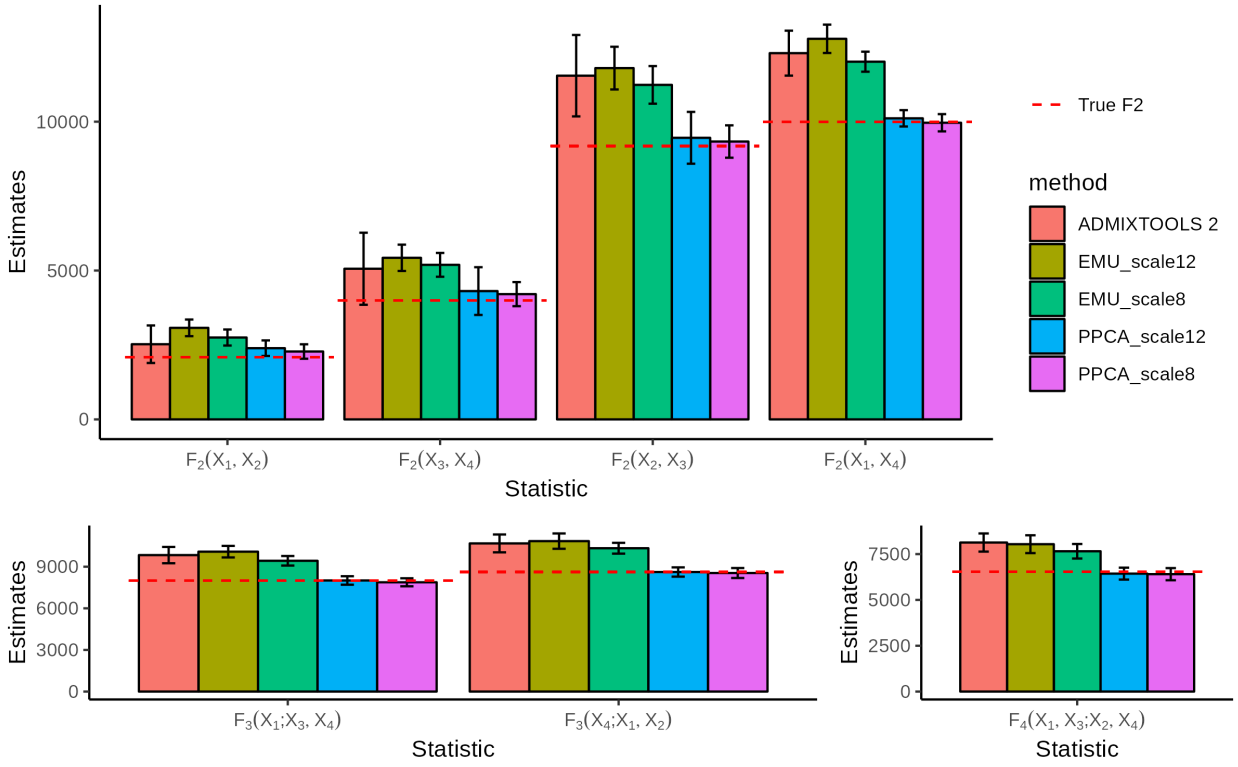

Figure 5: Comparison of PPCA and PCA to ADMIXTOOLS 2 in the presence of 50% random missingness, using population genotypes from one individual from each population.

3.3 Test of admixture

A major application of F-statistics are tests of admixture Orlando et al. (2021). We showed in the previous section that PPCA framework can be used to calculate the point estimates of F-statistics. In this section we show that we can also get standard errors for these estimates using block-jackknife Patterson, and use these to do hypothesis testing for admixture. We simulate a gene flow from X3 to X2 500 generations ago with the migration rate of $\mu\epsilon[0, 0.01, 0.05]$. We then compare PPCA framework to ADMIXTOOLS 2. We first test for admixture by checking if the estimate of $F_4(X1, X2, X3, X4)$ is significantly different from 0. We show that when there are 10 individuals in each population, both methods perform well (Fig. S6). In case of 0 migration rate, both methods estimate F_4 for all simulations to be close to 0, while at 5% migration rate, ADMIXTOOLS 2 and PPCA-framework have the power to detect admixture (with F_4 estimate 2 standard deviations below 0) in 90% and 70% simulations respectively. At migration rate of 1%, both the methods are unable to find admixture between X_2 and X_3, and instead incorrectly predict admixture between X_1 and X_3 (F_4 estimate is 2 standard deviations above the mean) for one simulation. Reducing the number of individuals to 1 from each population reduces the power for both the methods. With 50% missingness, both methods have no false positives in case of 0 migration rate, and at 5% migration rate ADMIXTOOLS 2 and PPCA framework detect admixture in 5% and 35% simulations respectively. At 1% migration rate, ADMIXTOOLS 2 infers admixture between X_2 and X_3 in 2 simulations and incorrect admixture between X_1 and X_3 in one simulation, while PPCA framework shows no prediction of admixture.

3.4 Evaluation on neandertal dataset

To test our framework on real data, we apply it to a published dataset of archaic humans from Eurasia **?**. This dataset consists of low-coverage late Neandertal specimens from Goyet (Goyet_L35MQ25), Spy (Spy_L35MQ25), Les Cottés (Les_Cottes_L35MQ25), Vindija (VindijaG1_L35MQ25), and Mezmaiskaya caves (Mezmaiskaya1_L35MQ25 and Mezmaiskaya2_L35MQ25) along with the high-coverage archaic specimens from Densiova (Altai, Denisova) and Vindija caves (Vindija33.19). We first estimate PC's for this dataset using PPCA, and show that our plot captures all the features of PCA from the authors (Fig. S7, S8). However, we demonstrate that with PPCA, the user can utilize all the specimens to estimate the PC's.

We analyze how close or distant the low-coverage late Neandertals are to high-coverage Vindija Neandertal using outgroup F3 statistic. F3(Altai, Vindija33.19, X) represents the branch length extending from Altai to the common ancestor of Vindija33.19 and X, where X is a low-coverage late Neandertal. Higher value of F3 denotes closeness of X to Vindija33.19 (Fig. 2). We compare the estimates from PPCA-based framework and ADMIXTOOLS 2, and show that they have a very similar pattern (Fig. 7). It is interesting to find that F_3(Altai, Vindija33.19, VindijaG1_L35MQ25) estimated by PPCA framework is higher than that by ADMIXTOOLS 2. We can write F_3 as a linear combination of F_2's:

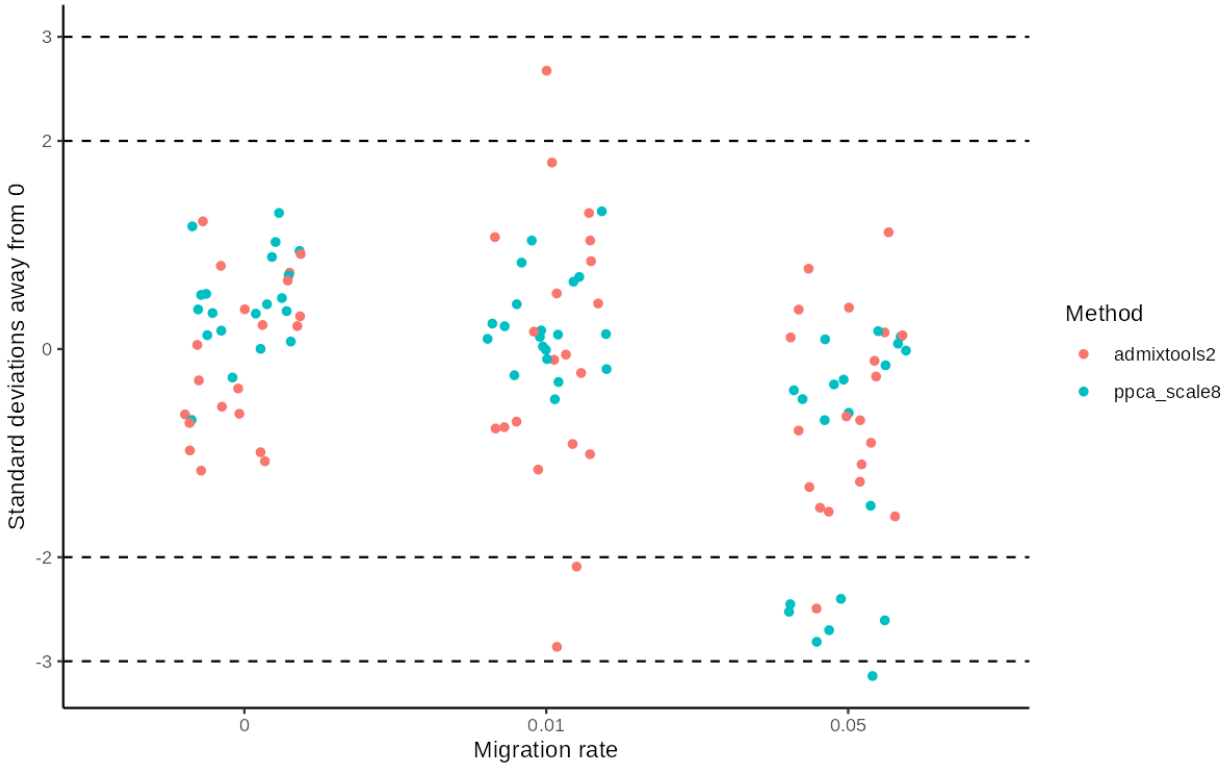

Figure 6: Test for admixture with individual-based F4 statistic. We compare ADMIXTOOLS 2 (orange)to PPCA-based-framework (blue) in the presence of 50% missingness in data.

$$\begin{aligned} &F_3(Altai, Vindija33.19, VindijaG1\_L35MQ25) \\ &= F_2(Altai, Vindija33.19) + F_2(Altai, VindijaG1\_L35MQ25) \\ &\quad - F_2(Vindija33.19, VindijaG1\_L35MQ25) \end{aligned} \tag{8}$$

We looked at the values of the three f_2-terms from ADMIXTOOLS 2 and PPCA framework to see why the two methods have different values. We found that f_2(Altai, Vindija33.19) and f_2(Altai, VindijaG1_L35MQ25) have values 0.072 and 0.135 from ADMIXTOOLS 2 respectively. Since both Vindija samples are from the same individual, the f_3 values should ideally be the same. PPCA framework outputs the values of f_2(Altai, Vindija33.19) and f_2(Altai, VindijaG1_L35MQ25) as 0.102 and 0.115, which are in line with the expectation. In addition, f_2(Vindija33.19, VindijaG1_L35MQ25) should ideally be 0. ADMIXTOOLS 2 shows a value of 0.0057, and PPCA-framework gives a value of 0.00094. It is interesting to note that Altai and Vindija33.19 are diploid genomes, and hence f_2(Altai, Vindija33.19) estimated with both the methods is similar for the two methods. In contrast, VindijaG1_L35MQ25 is apseudohaploid genome, and in this case we see that PPCA gives expected results and ADMIXTOOLS 2 does not. This is because the unbiased estimator in ADMIXTOOLS 2 for F_2 is undefined ($n_{1s} = 1$ in eq. 3).

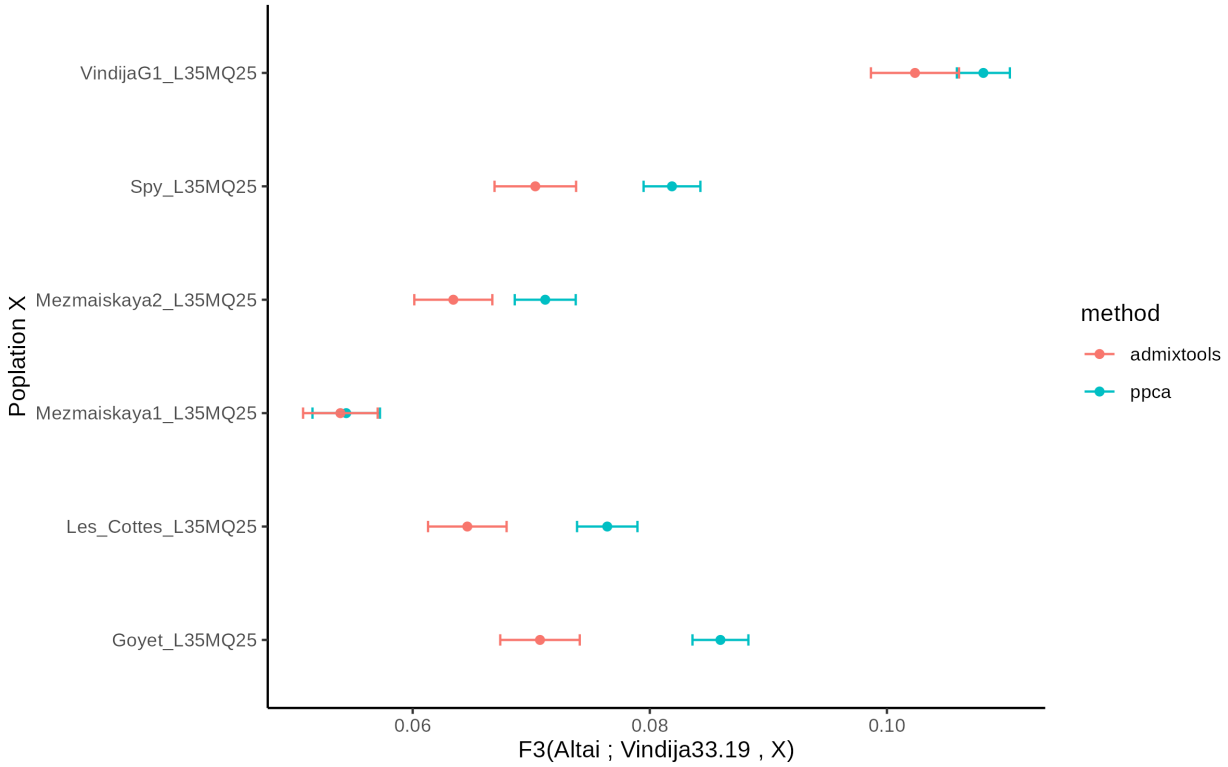

Figure 7: F_3(Altai, Vindija33.19, X) estimated with two methods. Larger value on x-axis represents more proximity to Vindija33.19. Bars show 2 standard errors.

4 Methods

4.1 PPCA implimentation

We implement PPCA using maximum-likelihood approach following Tipping and Bishop Tipping and Bishop, and modify this algorithm to work with missing data. Our approach to handle missingness is inspired from EMU Meisner et al. (2021). We describe our algorithm briefly:

1. Mean center data $Y = X - \mu$.
2. Set missing values to 0.
3. Perform SVD
4. Calculate the Gaussian noise parameter $\sigma^2 = \frac{1}{M-q}\sum_{k=M-q}^{M}\lambda_k^2$ as the sum of square of the $M - q$ smallest eigenvalues.
5. Obtain the MLE of the q^{th} eigenvalue as $\lambda_q^2 - \sigma^2$.
6. Calculate the linear mapping matrix $W = U_q(\lambda_q^2 - \sigma^2 I)$.

7. Reconstruct mean-centered data: $X_R = W(W^TW)^{-1}W^TY$.

8. Replace missing value with reconstructed values.

9. Repeat steps 2-7 until convergence.

4.2 Calculation of standard errors

We use a block-jackknife approach to calculate standard errors Maier et al. (2022). We divide the genome in 2 MB blocks, and estimate PC's and then F-statistics removing a block. Since the statistics obtained are not independent, we calculate variance using this equation Maier et al. (2022):

$$V = \frac{1}{g}\sum_{i=1}^{g}\frac{s_i}{S - s_i}(\hat{\theta} - \theta_i)^2 \tag{9}$$

Here, V is the variance of a statistic θ, g is the number of blocks, s is the number of sites in block i, and S is the total number of sites.

5 Discussion

In this study, we point out that the conventional method to estimate F-statistics suffers from two major problems. First, one needs to assign individuals to discrete populations which may not be justified for humans, and is difficult to do with ancient samples. And second, the estimates are inaccurate in presence of high amounts of missing genotypes, another characteristic of ancient DNA. We present a PCA-based framework to estimate F-statistics while taking in account these issues. We compare the statistics estimated with different PCA approaches, and show that all approaches work well on ideal simulations with enough samples. In case of individual-based F-statistics, both PPCA and LSE outperform classical PCA, and we provide a PPCA algorithm for the case of missing data. We compare PPCA framework to ADMIXTOOLS 2, and find that our framework outperforms ADMIXTOOLS 2 to estimate individual-based F-statistics with missing data.

PCA is widely used in population genetics to visualize clusters of individuals that may represent populations, and clines potentially representing historical admixture. Moreover PCA may reveal fine-scale structure in the population **?**. PCA's ability to condense complex genetic data into interpretable dimensions enhances our understanding of human evolution, migration, and admixture events, and throws light on the intricate mosaic of our species' history. However the visualization results from PCA may be influenced by the choice of PC's used, normalization, and the choice of populations used. Elhaik (2022). We provide a way of

quantifying the results of a PCA with F-statistics so that all PCA's can be comparable and can use the same normalizations. Such a quantification using all the top PC's also makes it more straight-forward to justify a visual result made with 2 PC's.

We used a PPCA-based framework with the Neandertal data to estimate PC's utilizing all the samples. This approach is more straight-forward than authors' approach to first estimate the PC's using high-coverage genomes, and then project the low-coverage genomes. We show that we can accurately estimate outgroup F_3 statistic from the PC's. We find that F_2 calculated using diploid genotype data with our framework is comparable to that of ADMIXTOOLS 2. However, ADMIXTOOLS 2 can not be used to estimate F_2 with pseudohaploid data since the unbiased estimator in ADMIXTOOLS 2 is undefined in this case. We show that our framework provides accurate estimates even with pseudohaploid samples.

One limitation of this framework is the need to perform multiple PPCAs to obtain standard errors, which can be computationally expensive. Further studies are needed to design a statistical framework that can estimate the errors using SNP loadings, and therefore can work fast with large datasets.

To summarize, we present a method to perform PCA and F-statistics jointly and show that this approach not only improves estimates of F-statistics, but also provides a solution to the standardization and quantification of PCA. Our framework is available on github as a snakemake pipeline: https://github.com/DivyaratanPopli/A-joint-framework-for-PCA-and-F-statistics.

6 Appendix

We provide the procedure to estimate LSE as laid out by Cabreros and Storey (2019). We define a symmetric matrix $\hat{\mathbf{H}}$:

$$\hat{\mathbf{H}} = \frac{1}{S}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{10}$$

Here, X is the (uncentered) genotype matrix with shape $m \times S$, $\hat{\mathbf{D}}$ is a matrix of diagonal entries "correcting" heterozygosities and S is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}(2 - x_{ik}), \tag{11}$$

for diploid data. The correction term given by Reich et al. (2009) and Patterson et al. (2012)

is different by a factor of 4:

$$d'_{ii} = \frac{1}{S}\sum_{k=1}^{S}\frac{x_{ik}}{2}(1-\frac{x_{ik}}{2}) = \frac{d_{ii}}{4}$$

The factor of 4 is due to the use of allele frequencies instead of genotypes, i.e. their approach would estimate $\hat{\mathbf{H}}$ (again for diploid data)

$$\hat{\mathbf{H}'} = \frac{1}{S}\frac{\mathbf{X}^T}{2}\frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}'}}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus,

$$h_{ii} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}^2 - d_i \tag{12a}$$

$$h_{ij} = \frac{1}{S}\sum_{k=1}^{S} x_{ik}x_{jk} \tag{12b}$$

Consider now

$$\begin{aligned} f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\ &= \frac{1}{S}\sum_{k=1}^{S} x_{ik}^2 - d_i + \frac{1}{S}\sum_{k=1}^{S} x_{jk}^2 - d_j - \frac{2}{S}\sum_{k=1}^{S} x_{ik}x_{jk} \\ &= \frac{1}{S}\sum_{k=1}^{m}(x_{ik}-x_{jk})^2 - d_i - d_j \\ &= F_2(i,j) \end{aligned} \tag{13}$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate unbiased F_2-statistics.

7 removed stuff

8 appendix old

$$\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{14}$$

Here, $\mathbf{X}$ is the (uncentered) genotype matrix and $\hat{\mathbf{D}}$ is a matrix of diagonal entries "correcting" heterozygosities and m is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = d_i = \frac{1}{m} \sum_{k=1}^{m} x_{ik}(2 - x_{ik}), \tag{15}$$

for diploid data. The correction term given by Reich 2009 and Patterson 2012 is different by a factor of four:

$$d'_{ii} = \frac{1}{m} \sum_{k=1}^{m} \frac{x_{ik}(2 - x_{ik})}{4} = \frac{d_{ii}}{4}$$

. The difference is explained that they use allele frequencies instead of genotypes, i.e. their approach would use (again for diploid data)

$$\hat{\mathbf{H}}' = \frac{1}{m} \frac{\mathbf{X}^T}{2} \frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}}'}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus, (TODO: double-check if rows/columns are aligned, sum should be over SNP)

$$h_{ii} = \frac{1}{m} \sum_{k=1}^{m} x_{ik}^2 - d_i \tag{16a}$$

$$h_{ij} = \frac{1}{m} \sum_{k=1}^{m} x_{ik} x_{jk} \tag{16b}$$

Consider now

$$\begin{aligned} f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\ &= \frac{1}{m} \sum_{k=1}^{m} x_{ik}^2 - d_i + \frac{1}{m} \sum_{k=1}^{m} x_{jk}^2 - d_j - \frac{2}{m} \sum_{k=1}^{m} x_{ik} x_{jk} \\ &= \frac{1}{m} \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 - d_i - d_j \\ &= F_2(i, j) \end{aligned} \tag{17}$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate F_2-statistics (possibly instead of PPCA?).

We describe two issues in accurate estimation of population allele frequencies:

1. Humans may not fit into well-differentiated discrete populations, except in cases where the populations have been isolated due to geographical features Novembre et al. (2008). The estimation of population allele frequencies depends on the assignment of individuals to discrete populations, and may be affected by miss-assignment especially when few samples are available.

2. Missing data in some individuals for certain sites can make it difficult to get accurate allele frequency estimates at those sites. One commonly used solution to this problem is to filter out all the sites with missing data. However, this may make the number of sites available for F-statistics quite small. E.g., for 100 individuals with 10% randomly missing sites, the available number of sites after filtering out positions with missing data would be ≈ 26 out of a total 1,000,000 sites.

Studies using PCA generally use specific PC's to visualize population structure or admixture, and the choice of the PC's used can be quite subjective Elhaik (2022). Estimation of F-statistics from PCA quantifies, using all the important PC's, what the researcher has visualized using seemingly arbitrary PC's.

However, a limitation with such a framework is that it is possible to define F-statistics in terms of PC's only when allele frequencies are known, and need not be estimated. This is due to the fact that PCA does not filter the noise in the data due to sampling. In addition, missing data can affect the computation of PC's, and subsequently F-statistics.

We would point out here that in ancient DNA studies PCA is sometimes used as quality-control step by constructing PC's using high-quality samples, and projecting the low-quality samples which may be from the same populations or even the same individuals as the high coverage samples. In the presence of contamination from present-day people, reference bias, ascertainment bias or batch effects, the projected sample may not overlap with an identical high-coverage sample. These biases and issues are not resolved with PPCA/LSE either, since PPCA only models sampling noise.

8.1 F-statistics with PPCA/LSE

In this study, we develop a statistical framework to estimate F-statistics between individuals in a PPCA framework. We show that PPCA explicitly models the error due to the sampling bias in allele frequencies. In addition, we demonstrate that PPCA based framework is not sensitive to random missing data, and so it can be used to visualize individuals in PCA-space without having to project lower quality samples.

We explain that PPCA provides a natural framework to estimate F-statistics with small sample-size and missing data. We show formal hypothesis tests for admixture and compare our results to admixtools2 Maier et al. (2022) on simulations. Finally we show the use of this framework on published datasets from neolithic Haak et al. and upper Paleolithic humans (Mateja).

Mathematically, the PPCA model can be represented as follows Tipping and Bishop:

Latent variable model: Latent variables: $Z \sim N(0, I)$, where Z is a S-dimensional latent variable, and I is the identity matrix.

Latent-to-observed mapping: $X = WZ + \mu + \Psi$, where X is the observed data, W is a M x q matrix of linear mappings, μ is the mean of the observed data, and Ψ is a Gaussian noise term.

Prior distributions:

Prior on the latent variables: $P(Z) = N(0, I)$

Prior on the noise term: $P(\Psi) = N(0, \sigma^2 I)$, where σ^2 is the variance of the noise.

Likelihood function: $p(X|Z, W, \mu, \Psi) = N(X|WZ + \mu, \sigma^2 I)$

9 old stuff

9.1 PCA

One way to estimate F-statistics

the estimation of allele frequencies can also be affected by large amounts of missing data. Since F-statistics are dependent on the ascertainment scheme, random missing data reduces the number of overlapping sites greatly. PCA is a powerful method to visualize population structure but it can be difficult to interpret (**???**). In particular, the population structure in PCA is a function of expected pairwise coalescence times (McVean, 2009), and thus is not explicitly tied to a particular scenario; different histories may yield similar or identical PCAs.

Thus, parameter estimation, model comparisions and formal tests of admixture are usually not carried out in PCA.

(add section here introducing the general idea of PCA, explain how it deals with error/noise and difference between probabilistic and regular PCA. Possibly also explain how people use PCA as a QC-step to detect batch effects, and how that compares with PCA used for population structure (i.e. Ainash' question from your talk))

F-statistics are a useful tool to quantify population structure, and provide tests for admixture. Hence, a common pipeline in many population genetic studies is to analyze PCA plots to look for visible patterns that could be due to past admixtures, followed by formal test with F-statistics Lazaridis et al. (2014, 2016). F-statistics use population allele frequencies to test for admixture, and hence the accuracy of these tests is limited by the accuracy of allele frequency estimates Peter (2016). Whereas F-statistics estimates are robust when there are enough high quality samples, we show that low number of samples and missing data decrease the accuracy of these tests.

A major issue when working with ancient DNA is low number of individuals and missing

data. This issue is exacerbated due to difficulty in assigning some of the individuals to discrete populations. Especially in the case of humans samples, we know that the genetic samples do not strictly belong to discrete populations, but form a continuous spectrum in the allele frequency space Oteo-García and Oteo (2021). Hence, it is a step in the right direction to think of a method to estimate F-statistics in a structure-aware framework. The easiest way to think about this is using PCA. PCA does not require assignment of individuals to discrete populations, and it has been shown that F-statistics can be estimated conveniently from distance between populations on PCA space Peter (2022). However, PCA distances are inaccurate when population sizes are small since PCA does not explicitly model sampling bias in the allele frequencies (see section XX). In addition, PCA is sensitive to missing data, and this makes it difficult to work with ancient DNA.

In this study, we develop a statistical framework to estimate F-statistics between many populations in a probabilistic PCA framework. that probabilistic PCA (PPCA) explicitly models the error due to the sampling bias in allele frequencies. In addition. we demonstrate that PPCA based framework is not sensitive to random missing data, and so it can be used to visualize individuals in PCA-space without having to project lower quality samples.

Finally, we show that PPCA provides a natural framework to estimate F-statistics with small sample-size and missing data. We show formal hypothesis tests for admixture and compare our results to admixtools2 citeadmixtools paper on simulations. Finally we show the use of this framework on published datasets from neolithic Haak et al. and upper Paleolithic humans (Mateja).

10 older stuff

There are several tools available to understand genetic diversity, and can be classified as the tools that make minimal assumptions to summarize population structure, and the tools that infer demographic parameters. Former set of tools includes f-statistics Patterson et al. (2012), PCA **?**, MDS Wang et al. (2009), Structure **?**, Admixture - There are different tools to study diversity: 1) tools that make minimum assumptions like fstatistics, PCA, MDS, Structure, Admixture, 2) tools that can be used to infer demographic parameters. - Many studies use PCA followed by f-statistics in their pipline. Peter et al., show that these analyses reveal the same biological signal, and can be done jointly. - This framework has limitation: population allele frequencies are not known. - Here we present an approach based on pPCA, and show that it is a more natural framework since it can take in account the errors associated with allele frequency estimation.

10.1 Theory

- f-statistics, and the sampling error terms in f2. - PCA and relation to f-statstics. - pPCA/PCA1 (Waaij et. al.) as a framework for dimentionality reduction taking in account the sampling error.

10.2 results

- Advantage in estimating PCs: Fig.1: PCA plot with standard errors

- Advantage in estimating f-statistics: We compare f-statistics from pPCA, PCA1, PCA, admixtools2 in terms of accuracy (and speed?). Fig.2: point estimates of f2's for slendr simulations where we have both ancient and modern populations. Fig.3: Examples of f4 test of treeness with different cases of migrations, we can also show a case of f3 test of admixture.

- Application to a published dataset (Fig.4)

10.3 Discussion

- One key advantage of this framework is that both point estimates and standard errors for PCA and fstatistics are estimated together in a consistent way. - This is a step towards solving the issue with the assumption of discrete populations. - This would be quite useful for cases where assigning individuals to populations can be difficult. - Future work: 1) A faster way of estimating standard errors. 1) to get uncertainty from snp loading. 2) Hypothesis testing

- This approach could also help with missingness as shown by Meisner et al. - pPCA makes it easy to analyse modern and ancient data together without having to project samples.

10.4 Methods

- Simulation method and parameters used

11 Abstract

Studies of genetic variation now routinely include data from thousands of individuals representing complex historical and temporal structure. Understanding and modeling patterns of genetic variation between large numbers of individuals and populations is thus a key

challenge in the usage of genetic data to answer questions about evolutionary history. Principle Component Analysis (PCA) and F-statistics sensu Patterson are both widely used for this purpose, but are usually analyzed dis-jointly. Here, we present a new framework based on probabilistic PCA to jointly estimate principal component and F-statistics from large panels of data. A key advantage of our approach is that we can calculate individual-based F-statistics efficiently, and so population assignments become a result rather than an a priori assumption. Furthermore, probabilistic PCA provides a natural framework for incorporating missing data, a common issue in ancient DNA analyses. Taken together, our results greatly simplify the analysis of large population genetic data sets, and allow for fast data exploration and statistical testing in a unified and consistent framework.

12 Background

12.1 Why study genetic diversity

The genetic diversity of human populations has been shaped by historical and environmental factors over hundreds of thousands of years. Therefore, a key objective of population genetics is to analyze the observable variations and patterns in order to understand and reconstruct the demographic and evolutionary history of our species.

12.2 The general pipeline

A general pipeline consists of a method to summarize the data with minimal assumptions. Examples are PCA, MDS, Structure, Admixture. These methods show a qualitative picture, but do not estimate a biologically meaningful parameter. And so, it is difficult to design statistical tests for the results of such methods. However, such qualitative results are generally followed by quantifiable methods based on f-statistics. F-statistics with 2,3, or 4 populations assumes a null-model as a tree-like structure and a deviation from the tree-like structure is represented the alternate model.

12.3 PCA and fstatistics

PCA is a method to rotate the dataset in a way that so that the axes that are analysed and plotted are aligned with the dimensions explaining the highest variation in the data. This is a way to do dimensionality reduction, and provides a way to do better data visualization. Ben[2020] showed that PCA and f-statistics are related, and

13 Introduction

13.1 Why combine pPCA and F-statistics?

In case of no missing data - Calculating f-statistics between populations already assumes clustering of individuals into populations. PCA shows this clustering, but it would be useful to quantify the distance between individuals to filter for individuals that cluster, spot outliers, and identify substructures.

-Faster calculation of f-statistics, when only few PCs are used. Calculation of pPCA may take some time, but afterwards f-statistics is less time taking. And since people anyway do both PCA and f-statistics, overall this would reduce time.

-admixture proportions from f-statistics using pca. We can check if it's more reliable.

In case of missing data - In case of admixturegraphs, missing data may reduce total snps (although, missing sites would be less if allele frequencies are calculated from available individuals in populations). pPCA may help in cases with e.g., few ancient individuals from each population with a lot of missing data.

- In case of ancient DNA, may help to include libraries with missing data instead of projecting them?

14 Practical application ideas

This is a list of ideas we could pursue. Goal would be to pick a few that would be easiest to accomplish.

1. Evaluate whether we use PPCA to calculate individual-based F-statistics accurately and fast?
    (a) in presence of missing data
    (b) for multivariate analyses including qpadm/qpgraph
    (c) include samples projected onto PCA
2. Grouping individuals in populations
3. Can we get confidence intervals on PCA?
    (a) Resampling SNPs
    (b) Resampling individuals
    (c) Calculate uncertainty based on SNP loadings
        i. For SVD $\mathbf{X} = (\mathbf{UD})\mathbf{V}^T = \mathbf{PV}^T$, we might be able to use the correlation in the entries of $\mathbf{V}$ to estimate the "effective" number of SNPs
4. Practical programming
    (a) Write software tool that jointly computes individidual-based F-stats and PCA
5. Data analysis – find a good data set or scenario to analyze
    (a) Standard Western Eurasian PCA
    (b) Indian data
    (c) Some application that Stephan / Wolfgang are working on
6. Looking at qpadm/qpwave
    (a) qpadm projects samples into a subspace made from a subset of samples. As these samples are not orthogonal, this subspace is likely highly non-orthogonal and therefore tricky to work with. Doing PCA before doing qpadm could be helpful.

15 Projecting onto prob PCA

For PCA, we can write $X_{[n\times p]} = USV^T$ and the PCs are given by SV^T. , and we would project by

$$X = USV^T \tag{18}$$
$$U^T X = SV^T \tag{19}$$
$$XVS^{-1} = USV^T VS^{-1} \tag{20}$$
$$= U \tag{21}$$
$$U^T Y = Y_{proj} \tag{22}$$
$$XVS^{-1} Y = Y_{proj} \tag{23}$$

and we can project using $U^T Y$ for new data Y.

16 F-tests using Wishart-log-likelihoods

Consider the F_4-statistic

$$F_4(A-B, C-D) = Cov(A-B, C-D) \tag{24}$$
$$= Cov(A,C) + Cov(B,D) - Cov(A,D) - Cov(B,D) \tag{25}$$
$$\tag{26}$$

16.1 PCA to Covariance matrix

If we have a matrix of PCs, $\mathbb{P}$, then $\mathbf{Y} = \mathbb{P}\mathbb{P}^T$ is an estimate of the covariance matrix. Consider the random variables $(A-B)$ and $(C-D)$. If we assume they are jointly normally distributed, then their joint distribution will again be normally distributed with mean zero and covariance matrix

$$\mathbf{X} = \begin{pmatrix} y_{11} + y_{22} - 2y_{12} & y_{13} + y_{24} - y_{14} - y_{23} \\ y_{13} + y_{24} - y_{14} - y_{23} & y_{33} + y_{44} - 2y_{34} \end{pmatrix} \tag{27}$$

where the y are the entries of the covariance matrix obtained from the PCA. Practically, we can either first calculate Y (if we want all F-stats), or first subset $\mathbb{P}$ to the four pops involved.

The off-diagonal elements of $\mathbf{X}$ are precisely the F-statistics we aim to calculate. And thus we set up a test to see whether they are zero.

16.2 Derivation of the test statistic

The sampling distribution of a covariance follows a Wishart distribution. This is a $p \times p$ -matrix-valued probability distribution that is parametrized by a degree-of-freedom parameter n and a covariance matrix $\mathbf{S}$, also of dimension $p \times p$. The simplest way to generate Wishart random variates is,

$$\mathbf{X} = \sum_{i=1}^{n} Y_i Y_i^T \tag{28}$$

where $Y_i \sim N(0, \mathbf{S})$. We also have

$$E[\mathbf{X}] = n\mathbf{S} \tag{29}$$
$$mode(\mathbf{X}) = (n-p-1)\mathbf{S} \tag{30}$$

The log-likelihood of a Wishart Distribution is

$$\log P(\mathbf{X}|\mathbf{S},n) = -\frac{np}{2}\log(2) - \frac{n}{2}\log|\mathbf{S}| - \Gamma_p\left(\frac{n}{2}\right) + \frac{n-p-1}{2}|\mathbf{X}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}^{-1}\mathbf{X}) \tag{31}$$

$$\propto -\frac{n}{2}\log|\mathbf{S}| - \frac{1}{2}\operatorname{tr}(\mathbf{S}^{-1}\mathbf{X})$$

$$= -\frac{n}{2}\log\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right) - \frac{1}{2}\frac{\sigma_{22}x_{11} + \sigma_{11}x_{22} - 2\sigma_{12}x_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \tag{32}$$

where the last step assumes a 2×2 matrix. Under H0, we have $\sigma_{12} = 0$, therefore

$$\log P(\mathbf{X}|\mathbf{S}_0,n) = -\frac{n}{2}\log\left(\sigma_{11}\sigma_{22}\right) - \frac{1}{2}\left[\frac{x_{11}}{\sigma_{11}} + \frac{x_{22}}{\sigma_{22}}\right] \tag{33}$$

This likelihood is easily separatable and we can estimate σ_{11} from x_{11} and σ_{22} from x_{22} directly.

The log-likelihood-ratio statistic can then be calculated as

$$\begin{aligned} R &= -2\log\left(\frac{P(\mathbf{X}|\mathbf{S}_0,n)}{P(\mathbf{X}|\mathbf{S},n)}\right) \\ &= 2[\log P(\mathbf{X}|\mathbf{S},n) - \log P(\mathbf{X}|\mathbf{S}_0,n)] \\ &= n\log\left(\sigma_{11}\sigma_{22}\right) - n\log\left(\sigma_{11}\sigma_{22} - \sigma_{12}^2\right) \\ &+ n\left[\frac{x_{11}}{\sigma_{11}} + \frac{x_{22}}{\sigma_{22}}\right] - n\frac{\sigma_{22}x_{11} + \sigma_{11}x_{22} - 2\sigma_{12}x_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \end{aligned} \tag{34}$$

using the estimates $\sigma_{ij} = \frac{x_{ij}}{n}$ we get

$$\begin{aligned} R &\approx n\log\left(x_{11}x_{22}\right) - n\log\left(x_{11}x_{22} - x_{12}^2\right) + 2n - 2n \\ &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) \end{aligned} \tag{35}$$

A $\log(n^2)$ term in each of the log-terms cancels. The statistic R is asymptotically χ^2 distributed with one degree of freedom.

If we instead use the mode $\sigma_{ij} \approx \frac{x_{ij}}{n-3}$:

$$\begin{aligned} R &\approx n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) + 2(n-3)\frac{x_{11}x_{22} - x_{12}^2}{x_{11}x_{22} - x_{12}^2} - 2(n-3) \\ &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) \end{aligned} \tag{36}$$

16.3 Comparison to Cavalli-Sforza & Piazza, 1975

The authors propose

$$R = \frac{|\mathbf{X}|}{|\mathbf{S}|} \tag{37}$$

where $|\cdot|$ is the determinant. For 2×2 matrices,

$$|\mathbf{X}| = x_{11}x_{22} - x_{12}^2 \tag{38}$$
$$|\mathbf{S}| = x_{11}x_{22} \tag{39}$$
$$R = \frac{x_{11}x_{22} - x_{12}^2}{x_{11}x_{22}} = 1 - \frac{x_{12}^2}{x_{11}x_{22}} \tag{40}$$
$$T = -2n\log(R) = 2n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) \tag{41}$$

The factor of 2 might be wrong..

16.4 other statistics

for x_{12} small, we may further approximate

$$\begin{aligned} R &= n\log\left(\frac{x_{11}x_{22}}{x_{11}x_{22} - x_{12}^2}\right) \\ &= n\log(x_{11}x_{22}) - n\log(x_{11}x_{22} - x_{12}^2) \\ &= n\log(x_{11}x_{22}) - n\log\left[x_{11}x_{22}\left(1 - \frac{x_{12}^2}{x_{11}x_{22}}\right)\right] \\ &= n\log\left(1 - \frac{x_{12}^2}{x_{11}x_{22}}\right) \\ &\approx n\frac{x_{12}^2}{x_{11}x_{22}} \end{aligned} \tag{42}$$

This is the coefficient of determination, which is the square of the correlation coefficient

$$r = \sqrt{R/n} = \frac{x_{12}}{\sqrt{x_{11}x_{22}}} \tag{43}$$

for which we have a t-distributed null

$$r\sqrt{\frac{n-2}{1-r^2}} \sim t(n) \tag{44}$$

The Fisher-Transform then yields

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = \arctan(r) \sim N(0,1) \tag{45}$$

which simplifies to

$$\arctan(r) = \frac{1}{2}\log\left(\frac{\sqrt{x_{11}x_{22}}+x_{12}}{\sqrt{x_{11}x_{22}}-x_{12}}\right) \tag{46}$$

This statistic is normally distributed under the null-hypothesis.

in tests with a simple bivariate normal, all of them behaved equally well.

17 Calibrating the standard errors of F-stats

Traditionally, the standard errors of F-stats are estimated using a block-Jackknife approach. However, the block size is usually hard to estimate, and may impact the resulting values.

17.1 What do the standard errors measure?

There are two types of uncertainty:

- **sampling uncertainty**, that stems from the fact that we only have a small sample from each population. Because it depends on the sample, we expect these uncertainties to be independent for each sampled population
- **evolutionary uncertainty** there is also uncertainty due to the randomness in evolution. In particular, the realized mean allele frequencies in populations will be different from those expected under some model.

17.2 Covariance of F_2-statistics

$$\begin{aligned} K &= Cov((X_1 - X_2)^2, (X_3 - X_4)^2) \\ &= Cov(X_1^2, X_3^2) + Cov(X_1^2, X_4^2) + Cov(X_2^2, X_3^2) + Cov(X_2^2, X_4^2) \\ &- 2\left[Cov(X_1^2, X_3X_4) + Cov(X_2^2, X_3X_4) + Cov(X_3^2, X_1X_2) + Cov(X_4^2, X_1X_2))\right] \\ &+ 4\left[Cov(X_1X_2, X_3X_4)\right] \end{aligned} \tag{47}$$

At the same time, we have

$$
\begin{aligned}
F_4^2 &= Cov(X_1 - X_2, X_3 - X_4)^2 \\
&= \left[Cov(X_1, X_3) + Cov(X_2, X_4) - Cov(X_1, X_4) - Cov(X_2, X_3)\right]^2 \\
&= Cov(X_1, X_3)^2 + Cov(X_2, X_4)^2 + Cov(X_1, X_4)^2 + Cov(X_2, X_3)^2 \\
&+ 2[Cov(X_1, X_3)Cov(X_2, X_4) + Cov(X_1, X_4)Cov(X_2, X_3)] \\
&- 2[Cov(X_1, X_3)Cov(X_1, X_4) + Cov(X_1, X_3)Cov(X_2, X_3)] \\
&- 2[Cov(X_2, X_4)Cov(X_1, X_4) + Cov(X_2, X_4)Cov(X_2, X_3)] \qquad (48) \\
&= (E[X_1^2] + E[X_2^2])(E[X_3^2] + E[X_4^2]) \\
&+ 2[E[X_1X_3]E[X_2X_4] + E[X_1X_4]E[X_2X_3]] \\
&- 2[E[X_1X_3]E[X_1X_4] + E[X_1X_3]E[X_2X_3]] \\
&- 2[EX_2X_4]E[X_2X_3] + E[X_2X_4]E[X_2X_3]] \qquad (49) \\
&= (E[X_1^2] + E[X_2^2])(E[X_3^2] + E[X_4^2]) \\
&+ 2[E[X_1X_3][E[X_2X_4] - E[X_1X_4]] + E[X_2X_3][E[X_1X_4] - E[X_1X_3]]] \\
&- 2[EX_2X_4]E[X_2X_3] + E[X_2X_4]E[X_2X_3]] \\
& \qquad (50)
\end{aligned}
$$

But we have $Cov(A, B) = E[(A - E[A])(B - E[B])] = E[AB] - E[A]E[B]$ and

$$
\begin{aligned}
Cov(A, B)Cov(C, D) &= E[(A - E[A])(B - E[B])]E[(C - E[C])(D - E[D])] \\
&= E[AB]E[CD] = E[ABCD] - Cov(AB, CD) \qquad (51)
\end{aligned}
$$

18 Notes on PCA1 of van Waaij et al

van Waaij et al. suggest at PCA on a matrix of the following form for PCA https://doi.org/10.48550/arXiv.2302.04596 (their PCA1). This approach is further motivated by eq 7 in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707457/, which is a very technical (but probably useful) reference.

$$\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}, \tag{52}$$

where $\mathbf{X}$ is the (uncentered) genotype matrix and $\hat{\mathbf{D}}$ is a matrix of diagonal entries "correcting" heterozygosities and m is the number of SNPs.

In particular $d_{ij} = 0$ for $i \neq j$ and,

$$d_{ii} = d_i = \frac{1}{m}\sum_{k=1}^{m} x_{ik}(2 - x_{ik}), \tag{53}$$

for diploid data. The correction term given by Reich 2009 and Patterson 2012 is different by a factor of four:

$$d'_{ii} = \frac{1}{m}\sum_{k=1}^{m} \frac{x_{ik}(2 - x_{ik})}{4} = \frac{d_{ii}}{4}$$

. The difference is explained that they use allele frequencies instead of genotypes, i.e. their approach would use (again for diploid data)

$$\hat{\mathbf{H}'} = \frac{1}{m}\frac{\mathbf{X}^T}{2}\frac{\mathbf{X}}{2} - \frac{\hat{\mathbf{D}'}}{4} = \frac{\hat{\mathbf{H}}}{4},$$

and so the parametrizations are equivalent, but will differ by a factor of four.

Thus, (TODO: double-check if rows/columns are aligned, sum should be over SNP)

$$h_{ii} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i \tag{54a}$$

$$h_{ij} = \frac{1}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \tag{54b}$$

Consider now

$$\begin{aligned} f_{ij} &= h_{ii} + h_{jj} - 2h_{ij} \\ &= \frac{1}{m}\sum_{k=1}^{m} x_{ik}^2 - d_i + \frac{1}{m}\sum_{k=1}^{m} x_{jk}^2 - d_j - \frac{2}{m}\sum_{k=1}^{m} x_{ik}x_{jk} \\ &= \frac{1}{m}\sum_{k=1}^{m}(x_{ik} - x_{jk})^2 - d_i - d_j \\ &= F_2(i,j) \end{aligned} \tag{55}$$

Hence the matrix $\hat{\mathbf{H}}$ can be used to estimate F_2-statistics (possibly instead of PPCA?). The detailed justification of this can be found in this statsexchange post that will need to be adapted.

Thus, this might be a useful alternative to PPCA to calculate F-statistics:

1. Calculate $\hat{\mathbf{H}} = \frac{1}{m}\mathbf{X}^T\mathbf{X} - \hat{\mathbf{D}}$
2. Double-Center $\hat{\mathbf{H}}$: $\mathbf{H}_c = \mathbf{C}\hat{\mathbf{H}}\mathbf{C}$, where $\mathbf{C}$ is a centering matrix
3. Obtain PCs using an eigendecomposition of $\mathbf{H}_c$: $\mathbf{P}\mathbf{P}^T = \mathbf{H}_c$
4. Calculate F_2 from the smaller space $\mathbf{P}$

19 Standard errors and effective number of SNPs

An issue in calculating standard errors for F-statistics is that SNP are usually correlated, and so standard variance and standard error calculations will fail.

let us assume n populations can be represented by some population structure model that is parameterized by some covariance matrix $\mathbf{X}$. We do not observe SNPs, but rather we have a noisy sample $\mathbf{G}_{[S\times n]}$ at S loci.

Let, as above, denote the data matrix as $\mathbf{G}$, which is a noisy version of an allele frequency matrix $\mathbf{X}$, and we assume we can do a PCA on some estimate of $\hat{\mathbf{X}}$ as $\mathbf{L}\mathbf{Z} = \hat{\mathbf{X}}$ where $\mathbf{L}$ are the orthonormal SNP-loadings and $\mathbf{Z}$ are the PCs. For example, we could do that using probabilistic PCA or the Cabreros-Storey-PCA

We are interested in statistics of the form

$$F_{ij} = (X_i - X_j)^2 \tag{56}$$

which can be estimated from $\mathbf{G}$ using the unbiased estimator of Patterson et al.

$$f_{ij} = \frac{1}{S}\sum_{s=1}^{n}(g_{si} - g_{sj})^2 - H_i - H_j. \tag{57}$$

alternatively, we can also obtain an estimate from the decomposition of $\hat{\mathbf{X}}$ as

$$p_{ij} = \frac{1}{S}\sum_{p=1}^{n}(z_{si} - z_{sj})^2, \tag{58}$$

this sum is over PCs. This estimator has thus the advantage that it can be computed a lot faster since typically $n \ll S$, and if X is low-rank we can even truncate the sum

An issue is the calculation of the standard error of f_{ij} and p_{ij}. A simple estimator is

$$\sigma_p = \sqrt{\frac{\mathrm{Var}(p_{ij})}{S}} \tag{59}$$

a problem with σ_p is that SNPs are not independent, and so the variance estimates are underestimated. For this purpose, a block-jackknife can be used.

We block data using a vector b, s.t $b_i = j$ means that the i-th SNP is in block j. For each block, we then have the pseudovalue

$$\tilde{f}^{(j)} = \frac{1}{m_j}\sum_{s} I[b_s = j](x_{si} - x_{sj})^2$$

where $I[\cdot]$ is an indicator and m_j is the number of entries in block j.

then

$$\sigma'_f = \sqrt{\frac{1}{g}\sum_{j}\left[\frac{m_j}{n - m_j}(\tilde{f}^{(j)} - f)^2\right]} \tag{60}$$

This motivates the effective number of SNPs,

$$S_e = S\left(\frac{\sigma_f}{\sigma'_f}\right)^2 \tag{61}$$

which gives the number of pseudo-independent observations.

simplifying block-JK Writing

$$f = \frac{1}{n}\sum_{s}(x_{si} - x_{sj})^2$$

References

Cavalli-Sforza: The history and geography of human genes - Google Scholar. URL https://scholar.google.com/scholar_lookup?title=The%20History%20and%20Geography%20of%20Human%20Genes&publication_year=1994&author=Cavalli-Sforza%2CLL&author=Menozzi%2CP&author=Piazza%2CA.

Agrawal, A., A. M. Chiu, M. Le, E. Halperin, and S. Sankararaman. 2020. Scalable probabilistic PCA for large-scale genetic variation data. PLOS Genetics 16:e1008773. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008773. Publisher: Public Library of Science.

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

Brisbin, A., K. Bryc, J. Byrnes, et al. 2012. PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. Human biology 84:343–364. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740525/.

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

Chen, X., and J. D. Storey. 2015. Consistent Estimation of Low-Dimensional Latent Structure in High-Dimensional Data. URL http://arxiv.org/abs/1510.03497. ArXiv:1510.03497 [stat].

DeGiorgio, M., and N. A. Rosenberg. 2013. Geographic Sampling Scheme as a Determinant of the Major Axis of Genetic Variation in Principal Components Analysis. Molecular Biology and Evolution 30:480–488. URL https://doi.org/10.1093/molbev/mss233.

Elhaik, E. 2022. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Scientific Reports 12:14683.

François, O., M. Currat, N. Ray, et al. 2010. Principal component analysis under population genetic models of range expansion and admixture. Molecular Biology and Evolution 27:1257–1268.

Gower, J. C. 1966. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika 53:325–338. URL https://www.jstor.org/stable/2333639. Publisher: [Oxford University Press, Biometrika Trust].

Lazaridis, I., D. Nadel, G. Rollefson, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536:419–424. URL https://www.nature.com/articles/nature19310. Number: 7617 Publisher: Nature Publishing Group.

Lazaridis, I., N. Patterson, A. Mittnik, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413. URL https://

www.nature.com/articles/nature13673. Number: 7518 Publisher: Nature Publishing Group.

Maier, R., P. Flegontov, O. Flegontova, P. Changmai, and D. Reich. 2022. On the limits of fitting complex models of population history to genetic data. URL https://www.biorxiv.org/content/10.1101/2022.05.08.491072v2. Pages: 2022.05.08.491072 Section: New Results.

McVean, G. 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genetics 5:e1000686. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757795/.

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

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. Science 201:786–792. URL https://www.science.org/doi/10.1126/science.356262. Publisher: American Association for the Advancement of Science.

Novembre, J., T. Johnson, K. Bryc, et al. 2008. Genes mirror geography within Europe. Nature 456:98–101. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/.

Orlando, L., R. Allaby, P. Skoglund, et al. 2021. Ancient DNA analysis. Nature Reviews Methods Primers 1:1–26. URL https://www.nature.com/articles/s43586-020-00011-0. Number: 1 Publisher: Nature Publishing Group.

Oteo-García, G., and J.-A. Oteo. 2021. A Geometrical Framework for f-Statistics. Bulletin of Mathematical Biology 83:14. URL https://doi.org/10.1007/s11538-020-00850-8.

Patterson, N. A modification to the jackknife to deal with adjacent blocks .

Patterson, N., P. Moorjani, Y. Luo, et al. 2012. Ancient Admixture in Human History. Genetics 192:1065–1093. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522152/.

Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. PLOS Genetics 2:e190. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190. Publisher: Public Library of Science.

Peter, B. M. 2016. Admixture, Population Structure, and F-Statistics. Genetics 202:1485–1501. URL https://www.genetics.org/content/202/4/1485.

Peter, B. M. 2022. A geometric relationship of F2, F3 and F4-statistics with principal component analysis. Philosophical Transactions of the Royal Society B: Biological Sciences 377:20200413. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2020.0413. Publisher: Royal Society.

Peter, B. M., D. Petkova, and J. Novembre. 2020. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. Molecular Biology and Evolution 37:943–951. URL https://doi.org/10.1093/molbev/msz280.

Petr, M., B. C. Haller, P. L. Ralph, and F. Racimo. 2022. slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. URL https://www.biorxiv.org/content/10.1101/2022.03.20.485041v2. Pages: 2022.03.20.485041 Section: New Results.

Price, A. L., N. J. Patterson, R. M. Plenge, et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904–909. URL https://www.nature.com/articles/ng1847. Number: 8 Publisher: Nature Publishing Group.

Price, A. L., A. Tandon, N. Patterson, et al. 2009. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLOS Genetics 5:e1000519. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000519. Publisher: Public Library of Science.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–959. URL https://academic.oup.com/genetics/article/155/2/945/6048111.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. Reconstructing Indian Population History. Nature 461:489–494. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842210/.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, et al. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLOS Genetics 1:e70. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0010070. Publisher: Public Library of Science.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin. 2008. Estimating local ancestry in admixed populations. American Journal of Human Genetics 82:290–303.

Serre, D., and S. Pääbo. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Research 14:1679–1685.

Sforza, L. L. C., and F. C. Sforza. 1995. The Great Human Diasporas: A History Of Diversity And Evolution. Basic Books. Google-Books-ID: JWV2tAEACAAJ.

Tipping, M. E., and C. M. Bishop. Probabilistic Principal Component Analysis .

Vi, T., Y. Vigouroux, P. Cubry, et al. 2023. Genome-Wide Admixture Mapping Identifies Wild Ancestry-of-Origin Segments in Cultivated Robusta Coffee. Genome Biology and Evolution 15:evad065.

van Waaij, J., S. Li, G. Garcia-Erill, A. Albrechtsen, and C. Wiuf. 2023. Evaluation of population structure inferred by principal component analysis or the admixture model. URL http://arxiv.org/abs/2302.04596. ArXiv:2302.04596 [stat].

Wang, D., Y. Sun, P. Stang, et al. 2009. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. BMC Proceedings 3:S109. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2795880/.