Modelling complex population structure using F-statistics and Principal Component Analysis

Benjamin M Peter

October 29, 2021

Abstract

Human genetic diversity is shaped by our complex history. Population genetic tools to understand this variation can broadly be classified into data-driven methods such as Principal Component Analysis (PCA), and model-based approaches such as F-statistics. Here, I show that these two perspectives are closely related, and I derive explicit connections between the two approaches. I show that F-statistics have a simple geometrical interpretation in the context of PCA, and that orthogonal projections are the key concept to establish this link. I illustrate my results on two examples, one of local, and one of global human diversity. In both examples, I find that population structure is sparse, and only a few components contribute to most statistics. Based on these results, I develop novel visualizations that allow for investigating specific hypotheses, checking the assumptions of more sophisticated models. My results extend F-statistics to non-discrete populations, moving towards more complete and less biased descriptions of human genetic variation.d

1 Introduction

As in most species, the genetic diversity of human populations has been influenced by our history and environment over the last several hundred thousand years (e.g Cavalli-Sforza *et al.*, 1994, **?**, Reich, 2018, **?**). In turn, an important goal of population genetics is to use observed patterns of variation to investigate and reconstruct the demographic and evolutionary history of our species (Schraiber and Akey, 2015, **?**).

The complicated genetic structure observed in present-day human populations (The 1000 Genomes Project Consortium, 2015, **?**) is caused by the interplay of demographic and evolutionary processes with both discrete and continuous components (Pritchard *et al.*, 2000, Rosenberg *et al.*, 2002, Serre and Pääbo, 2004, Rosenberg *et al.*, 2005, Bradburd *et al.*, 2018, Reich, 2018, Peter *et al.*, 2020). In particular, populations are expected to slowly differentiate if they are isolated from each other (Wahlund, 1928, Cavalli-Sforza and Piazza, 1975). In humans, this may be caused because continental-scale geographic distances limit migration, causing a pattern known as isolation-by-distance (SLATKIN, 1985). However, these patterns are usually not uniform, but shaped by geography, particularly barriers to migration such as mountain ranges, oceans or deserts (Cavalli-Sforza *et al.*, 1994, **?**, Rosenberg *et al.*, 2005, Bradburd *et al.*, 2013, Peter *et al.*, 2020). In addition, major historical population movements such as the out-of-Africa, Austronesian or Bantu expansions lead to more gradual patterns of genetic diversity over space (Cavalli-Sforza *et al.*, 1994, Ramachandran *et al.*, 2005, Novembre *et al.*, 2008, Stoneking, 2016, Racimo *et al.*, 2020). Local migration between neighboring populations will reduce differentiation, and long-distance migrations (Alves *et al.*, 2016), and secondary contact between diverged populations, such as Neandertals and modern humans (Green *et al.*, 2010) may lead to locally increased diversity (**?**).

Particularly for large and heterogeneous data sets, disentangling all these processes is challenging, and we cannot expect to devise a single model catching both broad strokes and minute details of human history. A commonly used analysis paradigm is thus to integrate tools based on different sets of assumptions. each emphasizing particular aspects of the data.

A typical analysis starts with data-driven, exploratory methods that summarize data making minimal assumptions (e.g. Schraiber and Akey, 2015). Examples are population trees (Cavalli-Sforza and Edwards, 1967, Felsenstein, 1973, Cavalli-Sforza and Piazza, 1975), Principal Component Analysis (PCA, Cavalli-Sforza *et al.*, 1994, Patterson *et al.*, 2006)) structure-like models (Pritchard *et al.*, 2000, Alexander *et al.*, 2009) or multidimensional scaling (MDS **?**)). However, these methods are not designed to answer specific research questions, and are limited in their ability to estimate biologically meaningful parameters. For this purpose, methods based on explicit demographic models are often used that aim to fit a specified or estimated model of divergence, migration and genetic drift to the data (Gutenkunst *et al.*, 2009, Excoffier *et al.*, 2013, Kamm *et al.*, 2015). The drawback of these methods is that, to make inference mathematically feasible, we need to introduce strong modeling assumptions such as that populations are discrete, randomly mating, or at equilibrium. While in most cases these assumptions are violated to some extent and cannot be verified, we hope that the resulting model fits provide sufficiently accurate answers to specific research questions.

F-statistics However, when the number of populations exceeds a few dozen, even codifying reasonable population models can be prohibitively difficult. One approach is to pick a small set of "representative" samples, and restrict modeling to this subset (e.g. Gravel *et al.*, 2011, Harney *et al.*, 2021). However, this has the drawback that a large proportion of the available data remains unused. An increasingly popular alternative approach, particularly in the analysis of human ancient DNA, is therefore to build up complex models from smaller building blocks based on the relationship between two, three or four populations.

The framework is based on a set of parameters called F-statistics *sensu* Patterson (Reich *et al.*, 2009, Patterson *et al.*, 2012, Peter, 2016). Formal definition will be given in the Theory section; but an informal motivation starts with the null model that populations are related as a tree, in which each F-statistic measures the length of a particular set of branches. (Figure 1; Semple and Steel, 2003, Peter, 2016).

In most applications, F-statistics are estimated from data, and then used as tests of treeness. In particular, under the assumption of a tree, F_3 is restricted to be non-negative, and many F_4-statistics will be zero (Semple and Steel, 2003, Patterson *et al.*, 2012), and data that violates these constraints is incompatible with a tree-like relationship between populations. The canonical alternative model is an admixture graph (or phylogenetic network) (Patterson *et al.*, 2012, Huson *et al.*, 2010), which is a tree which allows for additional edges reflecting gene flow (Figure 2A). However, admixture graphs are not the only plausible alternative model, and expected F-statistics can be calculated for a wide range of population genetic demographic models (Peter, 2016).

F-statistics and PCA The practical issue addressed in this study is how F-statistics can be reconciled with PCA, one of the most widely used data-driven modeling techniques. One way PCA can be motivated is as generating a low-dimensional representation of the data, with each dimension (called principal component, PC) retaining a maximum of the variance present in the data. To understand population structure, the use of PCA has been pioneered by Cavalli-Sforza *et al.* (1964), who used allele-frequency data at a population level to visualize genetic diversity (Cavalli-Sforza *et al.*, 1994). Currently, PCA is most commonly performed on individual-level genotype data (e.g. Patterson *et al.*, 2006, Novembre *et al.*, 2008), making use of the hundreds of thousands of loci available in most genome-scale data sets. The PCA-decomposition has been studied for a number of explicit population genetic models including trees (Cavalli-Sforza and Piazza, 1975), spatially continuous structure (Novembre and Stephens, 2008), the coalescent (McVean, 2009) and discrete population models (?). Here, in order to link PCA to F-statistics, I interpret both of them geometrically in *allele frequency space*, i.e. as functions of a high-dimensional Euclidean space. For F-statistics, this interpretation was recently developed by Oteo-Garcia and Oteo (2021), and for PCA it follows naturally from the interpretation of approximating a high-dimensional space with a low-dimensional one.

In the next section, I will formally derive the connection between F-statistics and PCA, and show how F-statistics can be interpreted geometrically, with a particular emphasis on two-dimensional PCA plots. In the Results section, I will then discuss how some of the most common applications of F-statistics manifest themselves on a PCA, and illustrate them on two example data sets, before ending with a discussion.

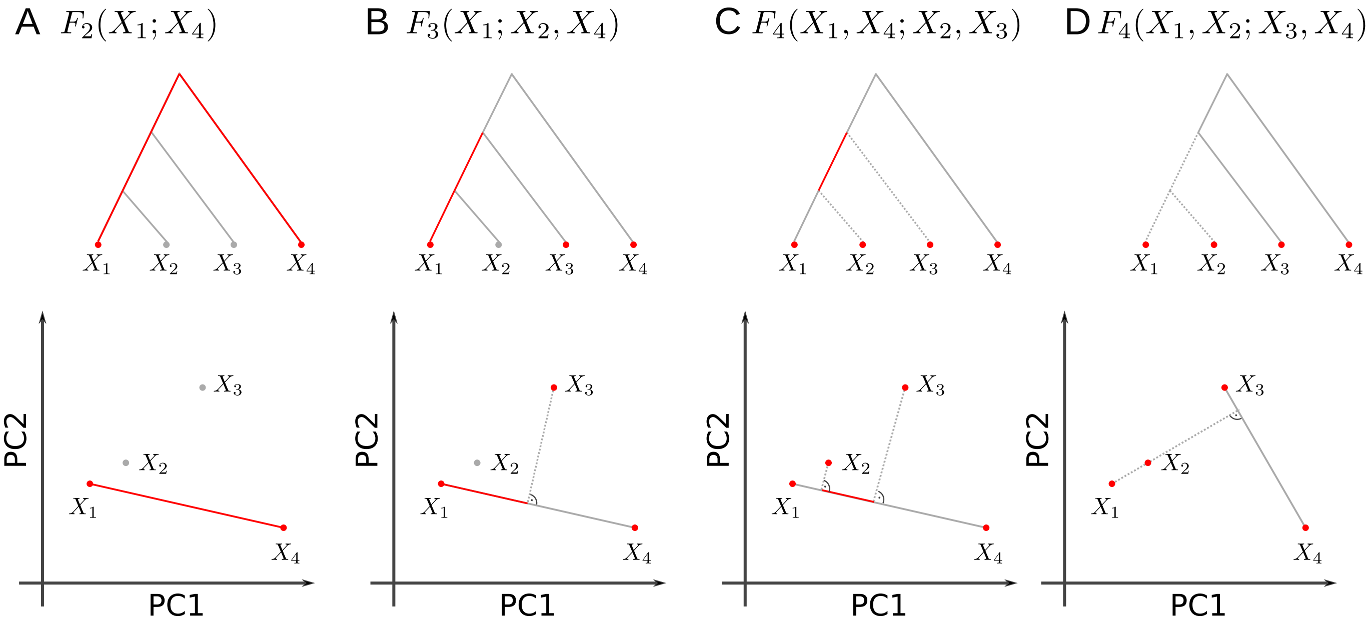

Figure 1: **Representation of F-statistics on trees and 2D-PCA-plots.** The schematics show four populations and their representation using a tree (top row) or a 2D-PCA plot (bottom row). A: F_2 represents the (squared) Euclidean distance between two tree leafs, and in PC-space. B: $F_3(X_1;X_3,X_4)$ corresponds to the external branch from X_1 to the internal node joining the populations, and is proportional to the orthogonal projection of $X_1 - X_3$ onto X_1-X_4. C: $F_4(X_1,X_4;X_2,X_3)$ corresponds to the internal branch in the tree, or the orthogonal projection of $X_2 - X_3$ on $X_1 - X4$. D: $F_4(X_1,X_2;X_3,X_4)$ The two paths from X_1 to X_2 and X_3 and X_4 are non-overlapping in the tree, which corresponds to orthogonal vectors in PCA-space.

2 Theory

In this section, I will introduce the mathematics and notations for F-statistics and PCA. A comprehensive treatise on PCA is given by e.g. Jolliffe (2013), a useful primer on the mathematics is Pachter (2014), and a helpful guide to interpretation is Cavalli-Sforza *et al.* (1994). Readers unfamiliar with F-statistics may find Patterson *et al.* (2012), Peter (2016) or Oteo-Garcia and Oteo (2021) helpful.

2.1 Formal Definition of F-statistics

Let us assume we have a set of populations for which we have SNP allele frequency data from S loci. Let x_{il} denote the frequency at the l-th SNP in the i-th population; and let $X_i = (x_{i1}, x_{i2}, \dots x_{iS})$ be a vector collecting all allele frequencies for population i. As X_i will be the only data summary considered here for population i, I make no distinction between the population and the allele frequency vector used to represent it.

The three F-statistics are defined as

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 \tag{1a}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) \tag{1b}$$

$$F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}), \tag{1c}$$

The normalization by the number of SNPs S is assumed to be the same for all calculations and is thus omitted subsequently. Both F_3 and F_4 can be written as sums of F_2-statistics:

$$2F_3(X_1; X_2, X_3) = F_2(X_1, X_2) + F_2(X_1, X_3) - F_2(X_2, X_3) \tag{2a}$$

$$2F_4(X_1, X_2; X_3, X_4) = F_2(X_1, X_3) + F_2(X_2, X_4) - F_2(X_1, X_4) - F_2(X_2, X_3) \tag{2b}$$

F-statistics have been primarily motivated in the context of trees and admixture graphs (Patterson *et al.*, 2012). In a tree, the squared Euclidean distance $F_2(X_1, X_2)$ measures the length of the path between populations X_1 and X_2 (Figure 1A); F_3 represents the length of an external branch (Figure 1B) and F_4 the length of an internal branch, respectively (Figure 1C). Crucially, for branches that do not exist in the tree (as in Figure 1D), F_4 will be zero. The length of each branch can be thought of in units of genetic drift, and is non-negative (Patterson *et al.*, 2012).

Thinking of F-statistics as branch lengths is useful for a number of applications, including building multi-population models (Patterson *et al.*, 2012, Lipson *et al.*, 2013), estimating admixture proportions (Petr *et al.*, 2019, Harney *et al.*, 2021) and finding the population most closely related to an unknown sample ("Outgroup"-F_3-statistic).

Most commonly however, F_3 and F_4 are used as tests of treeness (Patterson *et al.*, 2012): Negative F_3-values correspond to a branch with negative genetic drift, which is not allowed under the null assumption of a tree-like population relationship. Similarly if four populations are related as a tree, then at least one of the F_4 statistics between the populations will be zero (Buneman, 1974, Patterson *et al.*, 2012).

The most widely considered alternative model is an admixture graph (Patterson *et al.*, 2012), an example is given in Figure 2A. Here, the *typically unobserved) population X_y is generated by a mixture of individuals from the ancestors of X_2 and X_3. Over time, genetic drift will change X_y to X_x, which is the admixed population we observe. This will result in F_4-statistics that are non-zero, and, in some cases, in negative F_3-statistics (exact conditions can be found in Peter, 2016).

2.1.1 Geometric interpretation of F-statistics

An implicit assumption in the development of F-statistics is that population lineages are mostly discrete, and that gene flow is rare. Recently, Oteo-Garcia and Oteo (2021) showed re-deriving F-statistics in a geometric framework, showing that these assumptions are not necessary. Specifically, they interpret the populations X_i as points or vectors in the S-dimensional *allele frequency space* $\mathbb{R}^S$. In this case, the F-statistics can be thought of as inner (or dot) products, and they showed that all properties and tests related to treeness can be derived in this larger space. In particular the

F-statistics can be written as

$$F_2(X_1, X_2) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})^2 = \frac{1}{S}\langle X_1 - X_2, X_1 - X_2\rangle = \frac{1}{S}\left\|X_1 - X_2\right\|^2 \tag{3a}$$

$$F_3(X_1; X_2, X_3) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{1l} - x_{3l}) = \frac{1}{S}\langle X_1 - X_2, X_1 - X_3\rangle \tag{3b}$$

$$F_4(X_1, X_2; X_3, X_4) = \frac{1}{S}\sum_{l=1}^{S}(x_{1l} - x_{2l})(x_{3l} - x_{4l}) = \frac{1}{S}\langle X_1 - X_2, X_3 - X_4\rangle, \tag{3c}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\langle\cdot,\cdot\rangle$ denotes the dot product. Some elementary properties of the dot product between vectors a, b, c that I will use later are

$$\langle a, b\rangle = \sum_i a_i b_i \tag{4a}$$

$$\langle a, b\rangle = \|a\|\,\|b\|\cos(\phi) \tag{4b}$$

$$\langle a, a\rangle = \|a\|^2 \tag{4c}$$

$$\langle a + c, b\rangle = \langle a, b\rangle + \langle b, c\rangle, \tag{4d}$$

where ϕ is the angle between a and b. The inner product is closely related to vector projections

$$proj_b a = \frac{\langle a, b\rangle}{\|b\|^2} b, \tag{5}$$

which is a vector colinear to b whose length measures how much vector a points in the direction of b. Thinking of F-statistics as projections also holds on trees: In e.g. a $F_4(X_1, X_4; X_2, X_3)$-statistic (Figure 2C), the internal branch is precisely the intersection of the paths from X_1 to X_4 and from X_2 and X_3. The external branches are independent populations, and so they are expected to drift orthogonally to each other.

The drawback of the geometric approach of Oteo-Garcia and Oteo (2021) is that we have to deal with an very high-dimensional space, as the number of SNPs is frequently in the millions. However, it has been commonly observed that population structure is quite low-dimensional, and that the first few PCs provide a good approximation of the covariance structure in the data (Patterson *et al.*, 2006). Therefore, we may hope that PCA could yield a reasonable approximation of the allele frequency space, and that F-statistics as measures of population structure may likewise be well-approximated by the first few PCs.

2.2 Formal Definition of PCA

PCA is a common way of summarizing genetic data, and so a large number of variations of PCA exist, e.g. in how SNPs are standardized, how missing data is treated or whether we use individuals or populations as units of analysis. The version of PCA I use here is set up such that the similarities to F-statistics are maximized, and does *not* reflect how PCA is most commonly applied to genome-scale human genetic variation data sets. In particular, I assume that a PCA is performed on unscaled, estimated population allele frequencies, whereas many applications of PCA are based on individual-level sample allele frequency, scaled by the estimated standard deviation of each SNP (Patterson *et al.*, 2006). The differences this causes will be addressed in the discussion.

Let us again assume we have allele frequency data as above, but let us now assume we aggregate the allele frequency vectors X_i of many populations in a matrix $\mathbf{X}$ whose entry x_{il} reflects the allele frequency of the i-th population at the l-th genotype. If we have S SNPs and n populations, $\mathbf{X}$ will have dimension $n \times S$. Since the allele frequencies are between zero and one, we can interpret each Population X_i of $\mathbf{X}$ as a point in $[0,1]^S$, the allele frequency or *data space*, which is a subset of $\mathbb{R}^S$.

One way PCA can be motivated is that it aims to find a K-dimensional subspace of the data space that retains most variation in the data. K is at most $n-1$, in which case the data is simply rotated. However, the historical processes that generated genetic variation often result in *low-rank* data (Engelhardt and Stephens, 2010), so that $K \ll n$ explains a substantial portion of the variation; for visualization $K = 2$ is frequently used.

There are several algorithms that are used to perform PCAs, the most common one is based on singular value decomposition (e.g. Jolliffe, 2013). In this approach, we first mean-center $\mathbf{X}$, obtaining a centered matrix $\mathbf{Y}$

$$y_{il} = x_{il} - \mu_l$$

where μ_l is the mean allele frequency at the l-th locus.

PCA can then be written as

$$\mathbf{Y} = \mathbf{C}\mathbf{X} = (\mathbf{U}\mathbf{\Sigma})\mathbf{V}^T = \mathbf{P}\mathbf{L}, \tag{6}$$

where $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}$ is a centering matrix that subtracts row means, with $\mathbf{I}, \mathbf{1}$ the identity matrix and a matrix of ones, respectively. For any matrix $\mathbf{Y}$, we can perform a singular value decomposition $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ which, in the context of PCA, is interpreted as follows: The matrix of principal components $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}$ has size $n \times n$ and contains information about population structure. The SNP loadings $\mathbf{L} = \mathbf{V}^T$ form an orthonormal basis of size $n \times S$, its rows give the contribution of each SNP to each PC. It is often used to look for outliers, which might be indicative of selection (e.g **?**). Alternatively, the PCs can also be obtained from an eigendecomposition of the covariance matrix $\mathbf{Y}\mathbf{Y}^T$. This can be motivated from (6):

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{P}\mathbf{L}\mathbf{L}^T\mathbf{P}^T = \mathbf{P}\mathbf{P}^T, \tag{7}$$

since $\mathbf{L}\mathbf{L}^T = \mathbf{I}$.

2.3 Connection between PCA and F-statsitics

2.3.1 Principal components from F-statistics

PCA, as defined above, and F-statistics are closely related. In fact, the principal components can be directly calculated from F-statistics using multidimensional scaling, which, for squared Euclidean (F_2)-distances, leads to an identical decomposition to PCA (Gower, 1966). Suppose we calculate the pairwise $F_2(X_i, X_j)$ between all n populations, and collect them in a matrix $\mathbf{F}_2$. We can obtain the principal components from this matrix by double-centering it, so that its row and column means are zero, and perform an eigendecomposition of the resulting matrix:

$$\mathbf{P}\mathbf{P}^T = -\frac{1}{2}\mathbf{C}\mathbf{F}_2\mathbf{C}. \tag{8}$$

2.3.2 F-statistics in PCA-space

By performing a PCA, we rotate our data to reveal the axes of highest variation. However, the dot product is invariant under rotation, and F-statistics can be thought of as dot products (Oteo-Garcia and Oteo, 2021). What this means is that we are free to calculate F_2 either on the uncentered data $\mathbf{X}$, the centered data $\mathbf{Y}$ or any other orthogonal basis such as the principal components $\mathbf{P}$. Formally,

$$
\begin{aligned}
F_2(X_i, X_j) = \quad & \sum_{l=1}^{L} \left(x_{il} - x_{jl}\right)^2 \\
= \quad & \sum_{l=1}^{L} \left((x_{il} - \mu_l) - (x_{jl} - \mu_l)\right)^2 = \quad F_2(Y_i, Y_j) \\
= \quad & \sum_{k=1}^{n} (p_{ik} - p_{jk})^2 = \quad F_2(P_i, P_j),
\end{aligned}
\tag{9}
$$

A derivation of this change-of-basis is given in Appendix A, Equation A1. As F_3 and F_4 can be written as sums of F_2-terms (Eqs. 2a, 2b), analogous relations apply.

In most applications, we do not use all PCs, but instead truncate to the first K PCs, which explain most of the between-population genetic variation. Thus,

$$
\begin{aligned}
F_2(P_i, P_j) = \quad & \sum_{k=1}^{K} (p_{ik} - p_{jk})^2 + \quad \sum_{k=K+1}^{n} (p_{ik} - p_{jk})^2 \\
= \quad & \hat{F_2}^{(K)}(P_i, P_j) + \quad \epsilon^{(K)}(P_i, P_j) \quad .
\end{aligned}
\tag{10}
$$

In this notation, $\hat{F_2}^{(K)}$ is the approximation of F_2 with only the first K PCs considered, and $\epsilon^{(K)}$ is the corresponding approximation error. I will omit the superscript of $\hat{F_2}$ when the exact number of PCs is not relevant. If we sum up the squared approximation errors over all pairs of populations in our sample, we obtain

$$
\sum_{i,j} \epsilon^{(K)}(P_i, P_j)^2 = \sum_{i,j} \left(\hat{F_2}^{(K)}(P_i, P_j) - F_2^{(K)}(P_i, P_j)\right)^2 = \left\|\mathbf{F}_2 - \hat{\mathbf{F}_2}\right\|_F^2, \tag{11}
$$

where the Frobenius-norm $\|\cdot\|_F^2$ of a matrix is defined as the square root of the sum-of-squares of all its elements. This is precisely the function that is minimized in MDS (Jolliffe, 2013). In that sense, $\hat{\mathbf{F}_2}^{(K)}$ is the optimal low-rank approximation of $\mathbf{F}_2$ for any K in that it minimizes the sum of approximation errors of all F_2-statistics.

2.3.3 F-statistics and samples projected onto PCA

One of the easiest ways of dealing with missing data in PCA is to calculate the principal components (equation 6) only on a subset of the data with no missingness, and then to *project* the lower quality samples with high missingness onto this PCA. The simplest way to do this is to note that

$$
\mathbf{Y}\mathbf{L}^T = \mathbf{P}\mathbf{L}\mathbf{L}^T = \mathbf{P},
$$

and so a new (centered) population Y_{new} can be projected onto an existing PCA simply by post-multiplying it with $\mathbf{L}^T$:

$$
P_{\text{proj}} = Y_{\text{new}}\mathbf{L}^T;
$$

the k-th entry of P_{proj} gives the coordinates of the new sample on the k-th PC. However, it is likely that Y_{new} lies outside the variation of the original samples. In this case, there is a projection error

$$
\|Y_{\text{new}} - P_{\text{proj}}\mathbf{L}\|^2 = F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}).
$$

If we project with missing data, a similar projection can be used where we remove the rows from Y_{new} and $\mathbf{L}$ where data in Y_{new} is missing, and add a scaling factor (Patterson *et al.*, 2006).

Thus, if we compare the F-statistic of a projected sample, we have

$$\begin{aligned} F_2(X_i, X_{\text{new}}) &= F_2(Y_i, Y_{\text{new}}) \\ &= F_2(P_i, P_{\text{proj}}) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}) \\ &= \hat{F_2}(P_i, P_j) + \epsilon(P_i, P_j) + F_2(P_{\text{proj}}\mathbf{L}, Y_{\text{new}}). \end{aligned} \tag{12}$$

The second row follows because the projection error and projection are orthogonal to each other. The main implication of equation 12 is that both truncation and projection introduce some error, and that $\hat{F_2}(P_i, P_j)$ will be a good approximation to $F_2(P_i, P_j)$ only if both errors are small.

3 Material & Methods

The theory outlined in the previous section suggests that F-statistics have a geometric interpretation in PCA-space, which can be approximated on PCA plots. In the next section I explore this connection in detail, and illustrate it on two sample data sets that I briefly introduce here. Both are based on the analyses by Lazaridis *et al.* (2014). The data is from the Reich lab compendium data set (v44.3), downloaded from `https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-` using data on the "Human Origins"-SNP set (597,573 SNPs). SNPs with missing data in any population are excluded. The code used to create all figures and analyses will be available on `https://github.com/BenjaminPeter/fstats_pca`.

"World" data set This data set is a subset of the "World Foci" data set of Lazaridis *et al.* (2014), where I removed samples which are not permitted for free reuse. These populations span the globe and roughly represents global human genetic variation (638 individuals from 33 population) As adjacent sampling locations are often thousands of kilometers apart, I speculate that gene flow between these populations may not be particularly common; and their structure may therefore be well-approximated by an admixture graph. A file with all individuals used and their assigned population is given in **Supplementary File 1.**

Western Eurasian data set This data set of 1,119 individuals from 62 populations contains present-day individuals from the Eastern Mediterranean, Caucasus and Europe. It is motivated by the analysis of Lazaridis *et al.* (2014), who used it as a basis of comparison for ancient genetic analyses of Western Eurasian individuals, and PCAs based on similar sets of samples have been used in many other ancient DNA studies (e.g. **?**Haak *et al.*, 2015). Genetic differentiation in this region is low and closely mirrors geography (Novembre *et al.*, 2008). I thus speculate that gene flow between these populations is common (Ralph and Coop, 2013), and a discrete model such as a tree or an admixture graph might be a rather poor reflection of this data. A file with all individuals used and their assigned population is given in **Supplementary File 2.**

Computing F-statistics and PCA All computations are performed in `R`. I use `admixtools 2.0.0` (`https://github.com/uqrmaie1/admixtools`) to compute F-statistics. To obtain a PC-decomposition, I first calculate all pairwise F_2-statistics, and then use equation 8 and the `eigen` function to obtain the PCs. The right-hand side matrix of equation 8 is supposed to have non-negative eigenvalues (i.e. $-\mathbf{C}\mathbf{F}_2\mathbf{C}$ is positive-semidefinite). However as F_2-statistics are estimates, some eigenvalues might be slightly negative, which would lead to imaginary PCs. I avoid this by using the `nearPD`-function in R that ensures all eigenvalues have the correct sign.

4 Results

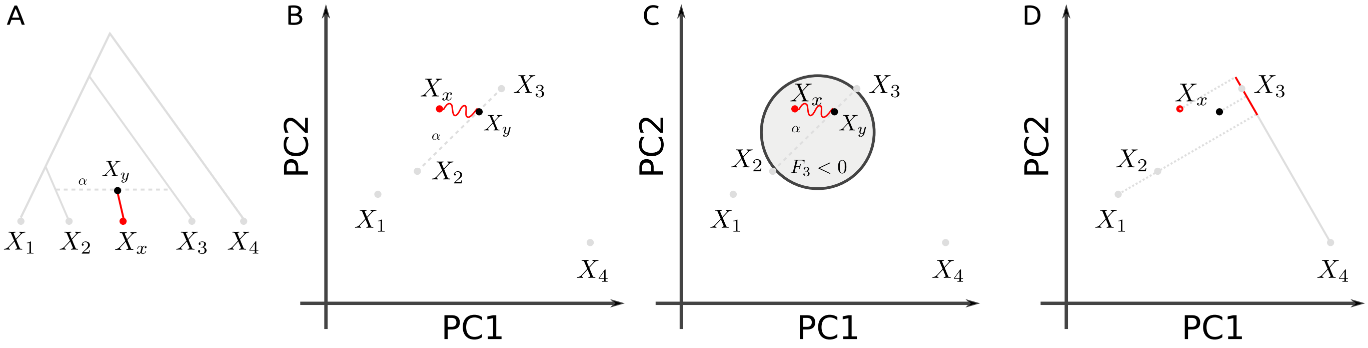

Figure 2: **Admixture representation on 2D-PCA-plot.** The schematics show four populations and their representation using an admixture graph (A) or a 2D-PCA plot. A: Admixture graph, with population X_y originating as an admixture of X_2 and X_3, with X_2 contributing proportion α. Subsequent drift (red branch) will change allele frequency to sampled admixture population X_x. B: PCA representation of the scenario in A. X_y originates on the segment connecting X_2 and X_3, and subsequent drift may move it in a random direction. C: $F_3(X_x; X_2, X_3)$ and negative region (light gray circle). $F_4(X_1, X_x; X_3, X_4)$ will no longer be zero (compare to Figure 1D).

The transformation from the previous section allows us to consider the geometry of F-statistics in PCA-space. The relationships we will discuss formally only hold if we use all PCs. However, the appeal of PCA is that frequently, only a very small number $K \ll n$ of PCS contain most information that is relevant for population structure, in which case the geometric interpretations become very simple. Thus, throughout the schematic figures, I assume that two PCs are sufficient to characterize population structure. In the data applications I evaluate how deviations of this assumption my manifest themselves in PCA plots.

4.1 F_2 in PC-space

The F_2-statistic is an estimate of the squared allele-frequency distance between two populations. On a tree (Figure 1A) this corresponds to the branch between two populations. In allele-frequency space, it corresponds to the squared Euclidean distance, and thus reflects the intuition that closely related populations will fall close to each other on a PCA-plot, and have low pairwise F_2-statistics. However, since F_2 can be written as a sum of squared (non-negative) terms for each PC (eq. 9), the distance on a PCA-plot will always be an underestimate of the full F_2-distance. Thus, PCA might project two populations with high F_2-distance very close to each other, which would indicate that these particular PCs are not suitable to understand and visualize the relationship between these particular populations. In converse, populations that are distant on a PCA-plot are guaranteed to also have a large F_2-distance.

4.2 When are admixture-F_3 statistics negative?

Consider again the admixture scenario in Figure 2A, where population X_y is the result of a mixture of X_2 and X_3, and subsequent drift changes the allele frequencies of the admixed population from X_y to X_x. How is such a scenario displayed on a PCA? Since the allele frequencies of X_y are a linear combination of X_2 and X_3, it will lie on the line segment connecting these two populations (Figure 2B), at a location predicted by the admixture proportions. Subsequent drift will change the allele frequency of X_x, and so in general it might fall on a different point on a PCA-plot. An exception occurs when X_x (and no other populations related to X_x) are not part of the construction of the PCA, so that $X_x - X_y$ is orthogonal to all PCs, i.e.

$$\langle X_x - X_y, X_i - X_j \rangle = \langle X_x - X_y, P_i \rangle = 0$$

for all populations $i, j \leq n$. In this case, X_x and X_y project to the same point, and the location on the PCA can directly be used to predict the admixture proportions (McVean, 2009, Brisbin *et al.*, 2012, Oteo-Garcia and Oteo, 2021). However, if either X_x, is included in the construction of the PCA, or if some gene flow occurred between X_x and any of the populations used to construct the PCA, X_x and X_y may project on different spots (Figure 2B).

Thus, a natural question to ask is given two source populations X_2, X_3, can we use PCA to predict which populations might be considered admixed between them? One way to address this question is to consider the space for which F_3 is negative, i.e.

$$\begin{aligned} 2F_3(X_x; X_2, X_3) &= 2\langle X_x - X_2, X_x - X_3 \rangle \\ &= \|X_x - X_2\|^2 + \|X_x - X_3\|^2 - \|X_2 - X_3\|^2 < 0. \end{aligned} \tag{13}$$

By the Pythagorean theorem, $F_3 = 0$ if and only if X_2, X_3 and X_x form a right-angled triangle. The associated region where $F_3 = 0$ is a n-sphere (or a circle in two dimensions) with diameter $\overline{X_2X_3}$ (The overline denotes a line segment). F_3 is negative when the triangle is obtuse, i.e. X_x could be considered admixed if it lies inside the n-ball with diameter $\overline{X_2X_3}$ (Figure 1B, Equation A2).

F_3 on a 2D PCA-plot. If we project this n-ball on a two-dimensional plot, $\overline{X_2X_3}$ will usually not align with the PCs; thus the ball may be somewhat larger than it appears on the plot. This geometry is perhaps easiest visualized on a globe. If we look at the globe from a view point parallel to the equator, both the north and south poles are visible at the very edge of the circle. But if we look at it from above the north pole, the north- and south-poles will be at the very same point.

Thus if $\hat{F}_3 \ll F_3$, the "true" circle will be bigger than what would be predicted from a 2D-plot, and populations that appear inside the circle on a PCA-plot may, in fact, have positive F_3-statistics. This is because they are outside the n-ball in higher dimensions. The converse interpretation is more strict: if a population lies outside the circle on *any* 2D-projection, F_3 is guaranteed to be bigger than 0 (see Equation A4 in the Appendix).

Example As an example, I visualize the admixture statistic $F_3(X; \text{Sardinian}, \text{Finnish})$, on the first two PCs of the Western Eurasian data set (Figure 3A). In this case, the projected n-ball (light gray) and circle based on two dimensions (dark gray) have similar sizes. However, several populations that appear inside the circles (e.g. Basque, Canary Islandersh) have, in fact, positive F_3-values, so they lie outside the n-ball. This reveals that the first two PCs do not capture all the genetic variation relevant for European population structure. Consequently, approximating F_3 by the first two or even ten PCs (Figure 3B) only gives a coarse approximation of F_3, and from Figure 3C we see that many higher PCs contribute to F_3 statistics.

However, many populations, particularly from Western Asia and the Caucasus, on the right-hand side of the plot, fall outside the circle. This allows us to immediately conclude that their F_3-statistics must be positive; and we should not consider them as a mixture between Sardinians and Fins.

4.3 Outgroup-F_3-statistics as projections

A common application of F_3-statistics is, given an unknown sample X_U, to find the most closely related population among a reference panel (X_i) (Raghavan *et al.*, 2014). This is done using an *outgroup-F_3*-statistic $F_3(X_O; X_U, X_i)$, where X_O is an outgroup. The reason an outgroup is introduced is to account for differences in sample times and additional drift in the reference populations (4A) The

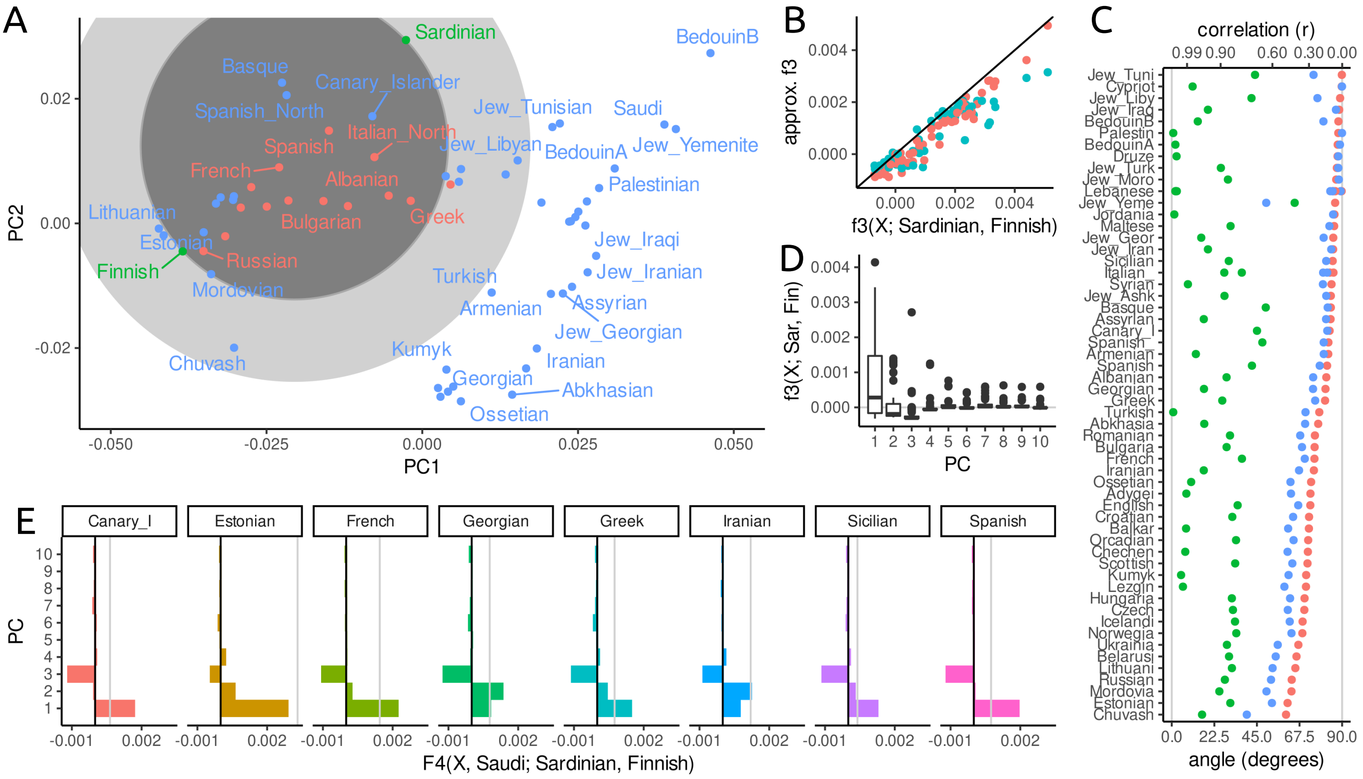

Figure 3: **PCA and F-statistics for the Western Eurasian data set** A: PCA-biplot; the light grey circle denotes the region for which $F_3(X; \text{Sardinian}, \text{Finnish})$ may be negative, the dark circle is based on just the first two PCs. Populations for which F_3 is negative are colored in red. B: F_3 approximated with two (blue) and ten (red) PCs versus the full spectrum. C: Boxplot of contributions of PCs 1-10 to each F_3-statistic. D: Projection angle and correlation interpretation of $F_4(X, \text{Saudi}; \text{Sardinian}, \text{Finnish})$ based on two PCs (green), three PCs (blue) or full data (red). E: Contribution of the first ten PCs to select F_4-statistics, with the first three PCs containing the majority of contributions.

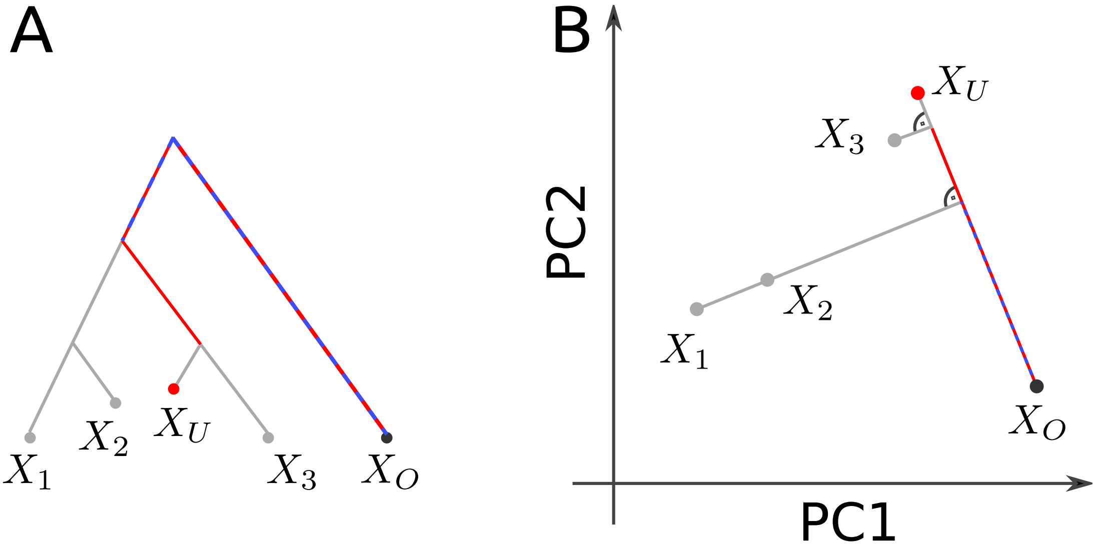

Figure 4: **Outgroup-F_3-statistics**. Interpretation of outgroup F_3-statistic on a tree (A) and PCA-plot (B). The red segment represents $F_3(X_O; X_U, X_3)$ and the dashed blue segment reflects $F_3(X_O; X_U, X_1)$ and $F_3(X_O; X_U, X_2)$, which have the same value.

outgroup-F_3-statistic $F_3(X_O; X_U, X_3)$ represents the branch length from X_O to the common node between the three samples in the statistic, and the closer this node is to X_U, the longer the branch and hence the larger the F_3-statistic.

To make sense of outgroup-F_3-statistics in the PCA context, I use the association of F_3-statistics to projections (Equation 5): On a PCA-plot, we can visualize this F_3-statistic as the projection of the vector $X_i - X_O$ onto $X_U - X_O$:

$$proj_{X_U - X_O} X_i - X_O = F_3(X_O; X_U, X_i) \frac{X_U - X_O}{F_2(X_O; X_U)}.$$

On the right-hand-side terms only the F_3 term depends on the X_i. The fraction can be thought of as a normalizing constant, which justifies the argument that the F_3-statistic and length of the projected vector are proportional to each other, and can thus be interpreted similarly. Thus, the outgroup-F_3-statistic is largest for whichever X_i projects furthest along the axis from the outgroup to the unknown population; in the example in Figure 4B this is X_3.

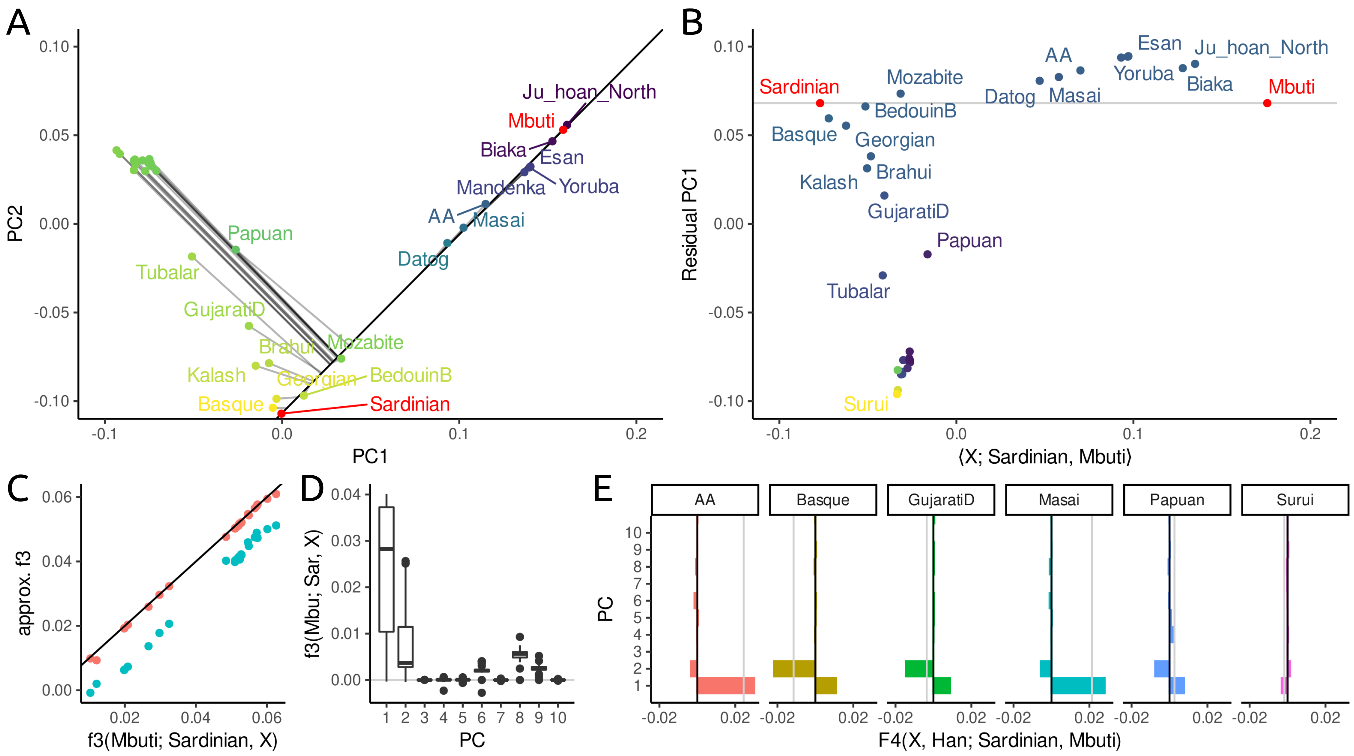

Figure 5: **PCA and F-statistics for the World data set** A: Visualization of Outgroup-F_3-statistic $F_3(\text{Mbuti}; \text{Sardinian}, X)$ on a PCA-biplot. The color of points correspond to the value of the F_3 statistic, with brighter yellows indicating higher values, i.e. higher similarity to Sardinians. The F_3-projection axis is given by a black line, the projection of populations onto this axis by thin gray lines. In the full data space, these projection are orthogonal to the axis. B: Projection along the axis Sardinian-Mbuti (X-axis), and PCA on residual of this projection (PC1 on Y-axis, PC2 as coloring). C: Approximation of $F_3(\text{Mbuti}; \text{Sardinian}, X)$ using the first two (blue) and first ten (red) PCs, respectively. D: Contributions of first ten PCs to all statistics of the form $F_3(\text{Mbuti}; \text{Sardinian}, X)$. E: Contributions of the first ten PCs to select F_4-statistics.

Example In Figure 5A, I use the World data set to visualize the outgroup-F_3-statistic $F_3(\text{Mbuti}; \text{Sardinian}, X$ in i.e. a statistic that aims to find the population most closely related to Sardinian (a Mediterranean Island), assuming the Mbuti are an outgroup to all populations in the data set. On a PCA, we can interpret this F_3 statistic as the projection of the line segment from Mbuti to population X_i onto the line through Mbuti and Sardinians (black line). For each population, the projection is indicated with a grey line. In the full data space, this line is always orthogonal to the segment Mbuti-Sardinian, but on the plot (i.e. the subspace spanned by the first two PCs), this is not necessarily the case. The coloring is based on the F_3-statistic calculated from all the data, with brighter values indicating higher F_3-statistics. In this case, the first two PCs approximate the F_3-statistic very well: Particularly the samples from East Asia, Siberia and the Americas (cluster in the top left of the plot) project very close to orthogonally, suggesting that most of the genetic variation relevant for this analysis is captured by these first two PCs. We can quantify this and find that the first two PCs slightly underestimate the absolute value of F_3 (Figure **??**C), but keep the relative ordering. I also find that many PCs, e.g. PCs 3-5, 7 and 10 have almost zero contribution to all F_3-statistics (Figure 3D), and PCs 6, 8 and 9 having a similar non-zero contribution for almost all statistics, likely because these PCs explain within-African variation.

4.4 F_4-statistics as angles

One interpretation of F_4 on PCA plots is similar to that of F_3; as a projection of one vector onto another, with the difference that now all four points may be distinct. F_4-statistics that correspond to a branch in a tree (as in Figure 1C), can be interpreted as being proportional to the length of a projected segment on a PCA plot (Figure 1G), again with the caveat that we need to scale it by a constant. If the F_4-statistic corresponds to a branch that does not exist in the tree, i.e it is a test statistic (Figure 1D), then, from the tree-interpretation, we expect $F_4(X_1, X_2; X_3, X_4) = 0$ implies that the vectors $X_1 - X_2$ and $X_3 - X_4$ are orthogonal to each other, i.e. that X_1 and X_2 map to the same point on the projection axis $\overline{X_3X_4}$ (Figure 1H). In the case of an admixture graph, this is no longer the case: Both population X_y and X_x in Figure 2D do *not* map to the same point as X_1 or X_2 do, implying that statistics of the form $F_4(X_1, X_x; X_3, X_4) \neq 0$.

Since F_4 is a covariance, its magnitude lacks an interpretation. Therefore, commonly correlation coefficients are used, as there, zero means independence and one means maximum correlation. For F_4, we can write

$$\mathrm{Cor}(X_1 - X_2, X_3 - X_4) = \frac{F_4(X_1, X_2; X_3, X_4)}{\|X_1 - X_2\| \, \|X_3 - X_3\|} = \cos(\phi), \tag{14}$$

where ϕ is the angle between $X_1 - X_2$ and $X_3 - X_4$. Thus, independent drift events lead to $\cos(\phi) = 0$, so that the angle is 90 degrees, whereas an angle close to zero means $\cos(\phi) \approx 1$, which means most of the genetic drift on this branch is shared.

Example To illustrate the angle interpretation I return to the Western Eurasian data. The PCA-biplot shows two roughly parallel clines (Figure 3A), a European gradient (from Sardinian to Finnish and Chuvash), and a Asian cline from Arab populations (top right) to the Caucasus (bottom right). This is quantified in Figure 3D, where I plot the angle corresponding to $F_4(X, \mathrm{Saudi}; \mathrm{Sardinian}, \mathrm{Finnish})$. For most Asian populations, using two PCs (green points) gives an angle close to zero, corresponding to a correlation coefficient between the two clines of $r > 0.9$. Just adding a third PC (blue), however, shows that the clines are not, in fact, parallel, and the correlation for most populations is low. The finding that three PCs are necessary to explain this data can also be seen from the spectrum of these F_4-statistics (Figure 3E), which have high contributions only from the first three PCs.

4.5 Other projections

So far, I used eq. 9 to interpret F-statistics on a PCA-plot, but the argument holds for *any* orthonormal projection of the data space. This is useful in particular for estimates of admixture proportions, which are often done in a small reference space (Patterson *et al.*, 2012, Petr *et al.*, 2019, Harney *et al.*, 2021, Oteo-Garcia and Oteo, 2021).

The simplest approach is the F_4-ratio to infer the admixture sources of population X as

$$\alpha = \frac{F_4(R_1, R_2; X, A)}{F_4(R_1, R_2; B, A)} = \frac{proj_{R_1-R_2}X - A}{proj_{R_1-R_2}B - A}, \tag{15}$$

which can be interpreted as projecting $X - A$ and $B - A$ onto $R_1 - R_2$ and measuring their relative proportions (Oteo-Garcia and Oteo, 2021). One important assumption is that there is no gene flow between X and the reference populations after gene flow (Petr *et al.*, 2019).

qpAdm extends this approach to a higher-dimensional reference space and multiple potential source populations. One open practical question in many applications is which reference and putative source populations to use (Harney *et al.*, 2021).

The theory developed here suggests some possible visualizations that may address this issue.

4.5.1 Example

In the PCA on the world overview data set, I found a gradient from Africans to Europeans (Figure 3D). I focus on this cline using an alternative projection by using F-statistics of the form $F_4(X, Y; \text{Sardinian}, \text{Yoruba}))$, which might e.g. be used in an F_4-ratio. These F_4-statistics are very well-approximated by the first two PCs, with a 99.2% correlation between F_4 and its approximation using the first two PCs (Figure 5C).

In Figure 5D, I show the projection $\langle X; \text{Sardinian}, \text{Yoruba}\rangle$ on the X-axis, which means that the horizontal difference between any pair of population is proportional to their F_4-statistic relative to Sardinians and Yorubans. We can also ask what variation is not represented by performinc a PCA on the residual of this projection, the first two residual PCs are given on the Y-axis and in the coloring. This visualization reveals that variation within Africans (with Mbuti, Biaka and Ju|'hoansi, top right) and the variation in East Asians and Americans are largely orthogonal to this projection axis, and so Sardinians and Yoruba would be poor references if we were interested in studying East Asian genetic variation.

The percentage of between-population variance explained by the Sardinia-Yoruba axis (24%) is much lower than that of the first PC (40%, Figure 5E). However, the cumulative variance explained by the first two axes is similar, with (52%) explained when adding residual PC1 to the projection, compared to 55% for the first two PCs. The advantage of specifying one axis is that it displays the orthogonal components more explicitly, reveals distinct structure in Africans and non-Africans and thus can be used to test assumptions of more complex models.

5 Discussion

Particularly for the analysis of human genetic variation, F-statistics are a powerful tool to describe population genetic diversity. Here, I show that the geometry of F-statistics (Oteo-Garcia and Oteo, 2021) leads to a number of simple interpretations of F-statistics on a PCA-plot. This allows for direct and quantitative comparisons between F-statistic-based results and PCA biplots. As PCA is often ran in an early step in data analysis, this also aids in generation of hypotheses that can be more directly evaluated using generative models, e.g using a lower number of populations. It also allows reconciling apparent contradictions between F-statistics and PCA-plots; differences between the two data summaries are either due to variation on higher PCs, or due to differences in assumptions about normalizations or population groupings. Previous interpretation of PCA in the context of population genetic models have focused on simple models such as trees (Cavalli-Sforza and Piazza, 1975), homogeneous spatial models (Novembre and Stephens, 2008) and discrete-population models (?). My interpretation here is different in that it puts more emphasis on the geometry itself, rather than directly interpreting the PCs. One consequence is that the results here are not impacted by sample ascertainment, sample sizes or number of principal components analyzed, which are common concerns in the interpretation of PCA. However, a very skewed sampling distribution will increases the likelihood that more or different PCs will have to be included in the analysis. From this perspective, one could envision a framework where F-statistics are used to decide which samples should be included to obtain a low-dimensional PCA-plot "representative" of the data.

As F-statistics are motivated by trees, they assume that populations are discrete, related as a graph, and that gene flow between populations is rare (Patterson *et al.*, 2012, Harney *et al.*, 2021). However, in many regions, all humans populations are admixed to some degree (Pickrell and Reich, 2014), and in regions such as Europe, genetic diversity is distributed continuously (Novembre *et al.*, 2008, Novembre and Stephens, 2008). This provides a challenge for interpretation; as many F_3 and F_4 statistics may indicate gene flow. In my example (Figure 3A), most Southern European populations are "admixed" between Basques and Turkish, but a more accurate model might be one of continuous variation where Basque and Turkish lie on one of multiple gradients; which is more directly visualized with PCA. There are a number of tools that have been developed that use multiple F-statistics to build complex models, such as `qpGraph` (Lazaridis *et al.*, 2014) and `qpAdm` (Harney *et al.*, 2021). One issue with these approaches is that they are usually restricted to at most a few dozen populations. As ancient DNA data sets now commonly include thousands of individuals, analysts are faced with the challenge of which data to include. A common approach is to sample a large number of distinct models, and retain the ones that are compatible with the data. However, as both `qpGraph` and `qpAdm` assume that gene flow is rare and discrete, selecting sets of populations that did experience little gene flow will provide good fits. One example of this is the world foci data set used here, which contains only 33 populations from across the world, and which is well-approximated by two PCs. However, this ascertainment misses a large amount of variation; a more dense sampling would show that in many places human genetic diversity is very gradual and multi-layered (Lazaridis *et al.*, 2014, Peter *et al.*, 2020). The PCA-based interpretation offers an alternative that trades interpretability for robustness. Particularly interpreting a (normalized) F_4-statistic as a correlation coefficient translates to generalized models of gene flow. Separating F-statistics in a sum of model and residuals, and performing a PCA on the latter (such as in Figure 5D) is another way how we can visualize F-statistics and evaluate the model fit.

The version of PCA I used for my analyses was chosen such that the similarities to F-statistics are maximized. In particular, I assume here that i) we have no missing data, ii) SNPs are equally weighted and iii) that individuals can be grouped into populations and iv) we use estimated allele frequencies. In contrast, most data analyses have to grapple with missing data, SNP are often weighted according to their allele frequency and observed, individual-level genotypes are used as the basis of PCA.

The liability of PCA to missing data is a well-studied problem and a number of algorithms for imputing have been proposed (e.g. Hastie *et al.*, 2015, **?**, Meisner *et al.*, 2021). Alternatively, samples with large amounts of missing data are projected onto PCAs computed without missing data (Patterson *et al.*, 2006). In contrast, missing data in F-statistics is handled by estimating a standard error using resampling across the genome (Patterson *et al.*, 2012), which does not distinguish between biological and sampling variation. These strategies are distinct, but not unique to the relative approaches and a PCA-like decomposition from F-statistics is commonly applied using MDS (e.g. Fu *et al.*, 2016).

The normalization of SNPs is similarly a matter of convention. The F-statistic framework assumes that each SNP is an identically-distributed (but not independent) random variable; and the same would hold if SNPs were weighted. The drawback for individual F-statistics is that this adds a dependency on additional samples (through the mean allele frequency) that may be unwanted for individual F-statistics, but could be advantageous for tools that aim to do joint inference from many F-statistics (Patterson *et al.*, 2012, Harney *et al.*, 2021).

The third issue of individual-based vs population-based analysis is similarly a matter of interpretation. For n samples, the number of possible F-statistics is on the order of n^4, and so the number of statistics is kept low by grouping individuals into populations. This is, however, not necessary, and F-statistics are often applied to individuals (e.g. Green *et al.*, 2010, Massilani *et al.*, 2020, **?**). Here, individual-based PCA can provide some guidance for grouping samples: Since F-statistics assume individuals are randomly drawn from a population, they should form tight clusters on a PCA-plot, otherwise population substructure becomes a possible alternative model for negative F_3-statistics and non-zero F_4-statistics (Peter, 2016).

The final difference between the F-statistic-based PCA used here and individual-based PCA is on the usage of estimated allele frequencies versus individual-based genotypes. The fact that PCA does not distinguish between sample-based errors and the underlying structure is a well-known drawback of standard PCA, and applying the theory presented here to individual-based PCA would result in F-statistics that incorporate some sampling noise. Probabilistic PCA is one class of approaches that aim to separate the population structure from individual-level noise (Agrawal *et al.*, 2020), and it seems likely that probabilistic PCA would yield a representation of the data that corresponds more closely aligned with F-statistics than regular PCA.

Thus, while the version of PCA used here differs from that proposed by Patterson *et al.* (2012), the differences are largely due to conventions that are partially arbitrary, and partially explained by the focus of PCA on exploratory data analysis. Particularly for studies where the description of population structure is a major focus, results might be easier to interpret if PCA and F-statistics are used in a way such that the results are comparable to each other.

References

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.

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

Alves, I., M. Arenas, M. Currat, A. Sramkova Hanulova, V. C. Sousa *et al.*, 2016 Long-distance dispersal shaped patterns of human genetic diversity in Eurasia. Molecular biology and evolution **33**: 946–958.

Bradburd, G. S., G. M. Coop, and P. L. Ralph, 2018 Inferring continuous and discrete population genetic structure across space. Genetics **210**: 33–52.

Bradburd, G. S., P. L. Ralph, and G. M. Coop, 2013 Disentangling the Effects of Geographic and Ecological Isolation on Genetic Differentiation. Evolution **67**: 3258–3273.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg *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.

Buneman, P., 1974 A note on the metric properties of trees. Journal of Combinatorial Theory, Series B **17**: 48–50.

Cavalli-Sforza, L. L., I. Barrai, and A. W. F. Edwards, 1964 Analysis of Human Evolution Under Random Genetic Drift. Cold Spring Harbor Symposia on Quantitative Biology **29**: 9–20.

Cavalli-Sforza, L. L., and A. W. F. Edwards, 1967 Phylogenetic Analysis: Models and Estimation Procedures. Evolution **21**: 550–570.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza, 1994 *The history and geography of human genes*. Princeton university press.

Cavalli-Sforza, L. L., and A. Piazza, 1975 Analysis of evolution: Evolutionary rates, independence and treeness. Theoretical Population Biology **8**: 127–165.

Engelhardt, B. E., and M. Stephens, 2010 Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. PLoS Genet **6**: e1001117.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLOS Genetics **9**: e1003905.

Felsenstein, J., 1973 Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics **25**: 471–492.

Fu, Q., C. Posth, M. Hajdinjak, M. Petr, S. Mallick *et al.*, 2016 The genetic history of Ice Age Europe. Nature **534**: 200–205.

Gower, J. C., 1966 Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**: 325–338.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences : 201019276.

Green, R., J. Krause, A. Briggs, T. Maricic, U. Stenzel *et al.*, 2010 A draft sequence of the Neandertal genome. science **328**: 710.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet **5**: e1000695.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick *et al.*, 2015 Massive migration from the steppe was a source for Indo-European languages in Europe. Nature **522**: 207–211.

Harney, E., N. Patterson, D. Reich, and J. Wakeley, 2021 Assessing the performance of qpAdm: a statistical tool for studying population admixture. Genetics **217**.

Hastie, T., R. Mazumder, J. D. Lee, and R. Zadeh, 2015 Matrix completion and low-rank SVD via fast alternating least squares. The Journal of Machine Learning Research **16**: 3367–3402.

Huson, D. H., R. Rupp, and C. Scornavacca, 2010 *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.

Jolliffe, I. T., 2013 *Principal Component Analysis*. Springer Science & Business Media.

Kamm, J. A., J. Terhorst, and Y. S. Song, 2015 Efficient computation of the joint sample frequency spectra for multiple populations. arXiv:1503.01133 [math, q-bio] .

Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick *et al.*, 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature **513**: 409–413.

Lipson, M., P.-R. Loh, A. Levin, D. Reich, N. Patterson *et al.*, 2013 Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. Molecular Biology and Evolution **30**: 1788–1802.

Massilani, D., L. Skov, M. Hajdinjak, B. Gunchinsuren, D. Tseveendorj *et al.*, 2020 Denisovan ancestry and population history of early East Asians. Science **370**: 579–583.

McVean, G., 2009 A genealogical interpretation of principal components analysis. PLoS genetics **5**: e1000686.

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) : btab027.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature **456**: 98–101.

Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. Nature genetics **40**: 646–649.

Oteo-Garcia, G., and J.-A. Oteo, 2021 A geometrical framework for f-statistics. Bulletin of Mathematical Biology **83**: 1–22.

Pachter, L., 2014 What is principal component analysis?

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich, 2006 Genetic evidence for complex speciation of humans and chimpanzees. Nature **441**: 1103–1108.

Patterson, N. J., P. Moorjani, Y. Luo, S. Mallick, N. Rohland *et al.*, 2012 Ancient Admixture in Human History. Genetics : genetics.112.145037.

Peter, B. M., 2016 Admixture, Population Structure and F-Statistics. Genetics : genetics.115.183913.

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.

Petr, M., S. Pääbo, J. Kelso, and B. Vernot, 2019 Limits of long-term selection against Neandertal introgression. Proceedings of the National Academy of Sciences **116**: 1639–1644.

Pickrell, J. K., and D. Reich, 2014 Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics **30**: 377–389.

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

Racimo, F., J. Woodbridge, R. M. Fyfe, M. Sikora, K.-G. Sjögren *et al.*, 2020 The spatiotemporal spread of human migrations during the European Holocene. Proceedings of the National Academy of Sciences **117**: 8989–9000.

Raghavan, M., P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen *et al.*, 2014 Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature **505**: 87–91.

Ralph, P., and G. Coop, 2013 The Geography of Recent Genetic Ancestry across Europe. PLoS Biol **11**: e1001555.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America **102**: 15942–15947.

Reich, D., 2018 *Who We Are and How We Got Here: Alte DNA und die neue Wissenschaft der menschlichen Vergangenheit.*. Pantheon, New York, illustrated edition edition.

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

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard *et al.*, 2005 Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet **1**: e70.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science (New York, N.Y.) **298**: 2381–2385.

Schraiber, J. G., and J. M. Akey, 2015 Methods and models for unravelling human evolutionary history. Nature Reviews Genetics .

Semple, C., and M. A. Steel, 2003 *Phylogenetics*. Oxford University Press.

Serre, D., and S. Pääbo, 2004 Evidence for Gradients of Human Genetic Diversity Within and Among Continents. Genome Research **14**: 1679–1685.

SLATKIN, M., 1985 GENE FLOW IN NATURAL-POPULATIONS. Annual Review of Ecology and Systematics **16**: 393–430.

Stoneking, M., 2016 *An Introduction to Molecular Anthropology*. John Wiley & Sons.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature **526**: 68–74.

Wahlund, S., 1928 Zusammensetzung Von Populationen Und Korrelationserscheinungen Vom Standpunkt Der Vererbungslehre Aus Betrachtet. Hereditas **11**: 65–106.

A Derivations

Depending on a readers' background in linear algebra, these results may appear elementary; I include them here for reference and because they were not obvious to me at the onset of this project.

F-statistics are invariant under a change-of-basis

$$\begin{aligned}
F_2(X_i, X_j) &= \sum_{l=1}^{S} \left((x_{il}-\mu_l)-(x_{jl}-\mu_l)\right)^2 = F_2(Y_i, Y_j) \\
&= \sum_{l=1}^{S} \big(\sum_k L_{kl}P_{ik} - \sum_k L_{kl}P_{jk}\big)^2 \\
&= \sum_{l=1}^{S} \left(\sum_k L_{kl}(P_{ik}-P_{jk})\right)^2 \\
&= \sum_{l=1}^{S} \left(\sum_k L_{kl}^2(P_{ik}-P_{jk})^2 + 2\sum_{k\neq k'} L_{kl}L_{k'l}(P_{ik}-P_{jk'})^2\right) \\
&= \sum_k \underbrace{\left(\sum_{l=1}^{L} L_{kl}^2\right)}_{1}(P_{ik}-P_{jk})^2 + 2\sum_{k\neq k'} \underbrace{\left(\sum_{l=1}^{S} L_{kl}L_{k'l}\right)}_{0}(P_{ik}-P_{jk'})^2 \\
&= \sum_k (P_{ik}-P_{jk})^2 \qquad \text{(A1)}
\end{aligned}$$

In summary, the first row shows that F_2 on the centered data will give the same results (as distances are invariant to translations), in the second row we apply the PC-decomposition. The third row is obtained from factoring out L_{lk}. Row four is obtained by multiplying out the sum inside the square term for a particular l. We have k terms when for $\binom{k}{2}$ terms for different k's. Row five is obtained by expanding the outer sum and grouping terms by k.The final line is obtained by recognizing that $\mathbf{L}$ is an orthonormal basis; where dot products of different vectors have lengths zero.

Note that if we estimate F_2, unbiased estimators are obtained by subtracting the population-heterozygosities H_i, H_j from the statistic. As these are scalars, they do not change above calculation.

The region of negative F_3-statistics is a n-ball Without loss of generality, assume that $X_1 = (r, 0, 0, \dots)$ and $X_2 = (-r, 0, 0, \dots)$, and let us assume that X_x has coordinates $(x_1, x_2, \dots, x_S)$ Assuming $F_3(X_x; X_1, X_2) = 0$, equation 13 becomes

$$\begin{aligned}
2F_3(X_x; X_1, X_2) &= \|X_x - X_1\|^2 + \|X_x - X_2\|^2 - \|X_1 - X_2\|^2 = 0 \\
&= \left[(x_1 - r)^2 + \sum_{i=2}^{S} x_i^2\right] + \left[(x_1 + r)^2 + \sum_{i=2}^{S} x_i^2\right] - 4r^2 \\
&= 2\left[\sum_{i=1}^{S} x_i^2 + r^2 + x_1 r - x_1 r\right] - 4r^2 \\
F_3(X_x; X_1, X_2) &= -r^2 + \sum_{i=1}^{S} x_i^2 = -r^2 + \|X_x\|^2 = 0, \qquad \text{(A2)}
\end{aligned}$$

which is the equation of a n-sphere with radius r and center at the origin, as assumed from the placing of X_1 and X_2. Now, assume that F_3 is negative, i.e. $F_3(X_x; X_1, X_2) = -k < 0$. Moving r^2 to the left we obtain

$$r^2 - k = \|X_x\|^2, \qquad \text{(A3)}$$

which is another n-sphere with a smaller radius, showing that all points inside the n-sphere will have negative F_3-values.

If a population lies outside the circle of this n-Sphere in any 2D-projection, F_3 is positive Assume the center of the n-sphere $C = \frac{X_1+X_2}{2} = (c_1, c_2, \dots c_S)$, and $X_x = (x_1, x_2, \dots x_S)$. Then,

$$\begin{aligned} F_3(X_x; X_1, X_2) &= \|X_x - C\|^2 - r^2 \\ &= \underbrace{(x_1 - c_1)^2 + (x_2 - c_2)^2}_{>r^2} + \underbrace{\sum_{i=3}^{S}(x_i - c_i)^2}_{\geq 0} - r^2 \\ &> 0. \end{aligned} \tag{A4}$$

The condition $(x_1 - c_1)^2 + (x_2 - c_2)^2 > r^2$ is satisfied whenever X_x is outside the circle obtained from projecting the n-sphere on the first two dimensions. An analogous argument applies for any low-dimensional representation.