Classification of human populations into races from five dimensional PCA data

Jorma Jormakka

I got this question from a commenter Justwondering. The question was what in my opinion would be the best way to classify populations to races assuming that the starting data is a five dimensional PCA-data. That is, for each population we have five real numbers PCA1 to PCA5 and the name of the population.

There are excellent ways of classifying populations assuming that we have admixture data, or full genome data from a few individuals, or even Y-DNA and mtDNA data, but in this theoretical problem we have only the five dimensional PCA data.

Even two dimensional PCA data divides (PCA1-PCA2 plot) the populations quite well into continental races, but that is not the answer in this theoretical problem: you should use the five dimensional data.

I would approach like this. First we find the five dimensional PCA data for the ancestral populations starting from Europe. There is a strong belief that Europeans (excluding Finnic populations) derive from three ancestral populations: West European Hunter-Gatherers (WHG), Early European Farmers (EEF) and the steppe people (Yamnaya). We find a small number of European populations that can reasonably be derived from these ancestral populations and invert a nine dimensional matrix (as I will soon show, it is not difficult today to invert 9-dimensional matrices with a computer) and get the ancestral PCA data. Then we can add other European populations and see if they can be expressed as linear combinations of these three ancestors. If not, there may have been other admixture, or evolution (like a genetic bottleneck, the PCA data mostly is not from alleles that are selected).

Justwondering was interested in Jews. Going to the Middle East, we do the same process of finding three ancestral populations. This is because there is a strong belief that Middle Eastern populations derive from original Middle East Neolithic population and admixture with European (Greeks and Romans) and African (slave trade) populations. If we find these populations, we can see if the European population in the Middle East is close to Greeks and Romans (Mediterranean), or maybe e.g. Slavic from white slave trade. Especially, if we are interested in Ashkenazi Jews (which I am not), we can check if they are close to Italians in this PCA data: we express Italians with the three European ancestral populations and Ashkenazi with three European and a Middle East population. This may give something new. I have not done this analysis, but I did calculate how to get the ancestral populations from five dimensional PCA data.

The data Justwondering was referring to is in the file k15-pca.xls (an excel file). It can be saved as text and it looks like this:

```
      PC 1  PC 2  PC 3  PC 4  PC 5
Abkhasian   -8.8086     -18.28      1.1011      30.412      -5.8134
Adygei      -8.5323     -13.669     3.4862      20.7  -5.7693
Afghan_Hazara     19.266      -4.7453     7.1278      3.9377      -3.6038
Afghan_Pashtun    2.0853      -28.011     4.2171      2.3799      -4.4215
Afghan_Tadjik     7.8421      -17.83      4.3782      6.7512      -4.8517
Afghan_Turkmen    13.462      -0.43163    6.9654      5.2387      -10.051
```

There are some two hundred or so populations. I found the data form the site https://antropologia-fizyczna.pl/wielocechowe-analizy-statystyczne/genetyka-populacyjna/globalne/analizy-autosomalnego-dna/k15-eurogenes-wielocechowa-analiza-audna-dla-populacji-z-calego-swiata.

You click the link under the text:

Wykres Principal Component Analysis (PCA) w 3D. Można go obracać. 1. Wykres gdzie PC1 to oś x, PC2 to oś y, PC3 to os z, PC4 to wielkość kul.

It may help if you read Polish, but it is not necessary. You get the file, save it as text and take your Linux programming environments and write a program to do this what I write here. Before writing the program you might look at the plots in this antropologia webpage: they already did lots of analysis. There are separate plots for each PCA1-PCA5 and you can see already from there that populations do fall into races and you can just simply fix a threshold on each PCAj and make a race predictor. It will work nicely.

But I would not do such a thing. When I noticed that there is a mathematical way that requires inverting matrices and could easily give nonlinear equations in 15 unknowns (originally so, but it went nicely to 9 linear equations), I was quite taken. So, I had to do this a bit more complicated method. There has to be some use to all that mathematics I once studied.

So, let us start. For each population j we have a five-vector

(1) $$Y_j = \begin{bmatrix} y_{j,1} & y_{j,2} & y_{j,3} & y_{j,4} & y_{j,5} \end{bmatrix}^T \qquad j = 1,2,3,4,...$$

that is the PCA data, five real numbers in a vector. We want to find three ancestral populations

(2) $$X_m = \begin{bmatrix} x_{m,1} & x_{m,2} & x_{m,3} & x_{m,4} & x_{m,5} \end{bmatrix}^T \qquad m = 1,2,3$$

such that

(3) $$Y_j = \sum_{m=1}^{3} c_{j,m} X_m$$

for some real numbers $c_{j,m}$. Let us write

(4) $$C_j = \begin{bmatrix} c_{j,1} & c_{j,2} & c_{j,3} \end{bmatrix}^T \qquad Y_{j,A} = \begin{bmatrix} y_{j,1} & y_{j,2} & y_{j,3} \end{bmatrix}^T \qquad Y_{j,B} = \begin{bmatrix} y_{j,4} & y_{j,5} \end{bmatrix}^T$$

$$A = \begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} \\ x_{1,2} & x_{2,2} & x_{3,2} \\ x_{1.3} & x_{2,3} & x_{3,3} \end{bmatrix} \qquad D = A^{-1} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix} \qquad B = \begin{bmatrix} x_{1,4} & x_{2,4} & x_{3,4} \\ x_{1,5} & x_{2,5} & x_{3,5} \end{bmatrix}.$$

Then

(5) $$Y_{j,A} = AC_j \quad \text{and} \quad Y_{j,B} = BC_j \quad \text{thus } Y_{j,B} = BA^{-1}Y_{j,A} = BDY_{j,A}.$$

For $k \in \{4,5\}$ we have

(6) $$y_{j,k} = \sum_{m=1}^{3} \left(\sum_{i=1}^{3} x_{i,k} d_{in} \right) y_{j,n} = \sum_{i=1}^{3} \left(\sum_{n=1}^{3} d_{in} y_{j,n} \right) x_{i,k}.$$

Write

(7) $$D_{j_1,j_2,j_3} = \begin{bmatrix} \sum_{n=1}^{3} d_{1n} y_{j_1,n} & \sum_{n=1}^{3} d_{2n} y_{j_1,n} & \sum_{n=1}^{3} d_{3n} y_{j_1,n} \\ \sum_{n=1}^{3} d_{1n} y_{j_2,n} & \sum_{n=1}^{3} d_{2n} y_{j_2,n} & \sum_{n=1}^{3} d_{3n} y_{j_2,n} \\ \sum_{n=1}^{3} d_{1n} y_{j_3,n} & \sum_{n=1}^{3} d_{2n} y_{j_3,n} & \sum_{n=1}^{3} d_{3n} y_{j_3,n} \end{bmatrix}.$$

Then for $k \in \{4,5\}$

(8) $$\begin{bmatrix} y_{j_1,k} \\ y_{j_2,k} \\ y_{j_3,k} \end{bmatrix} = D_{j_1,j_2,j_3} \begin{bmatrix} x_{1,k} \\ x_{2,k} \\ x_{3,k} \end{bmatrix}.$$

For any j_1, j_2, j_3 holds

(9) $$\begin{bmatrix} x_{1,k} \\ x_{2,k} \\ x_{3,k} \end{bmatrix} = \left(D_{j_1,j_2,j_3}\right)^{-1} \begin{bmatrix} y_{j_1,k} \\ y_{j_2,k} \\ y_{j_3,k} \end{bmatrix}.$$

Thus, the following equation holds

(10) $$\left(D_{j_1,j_2,j_3}\right)^{-1} \begin{bmatrix} y_{j_1,k} \\ y_{j_2,k} \\ y_{j_3,k} \end{bmatrix} = \left(D_{j_4,j_5,j_6}\right)^{-1} \begin{bmatrix} y_{j_4,k} \\ y_{j_5,k} \\ y_{j_6,k} \end{bmatrix} \qquad \text{i.e.,}$$

$$D_{j_4,j_5,j_6} \begin{bmatrix} y_{j_1,k} \\ y_{j_2,k} \\ y_{j_3,k} \end{bmatrix} - D_{j_1,j_2,j_3} \begin{bmatrix} y_{j_4,k} \\ y_{j_5,k} \\ y_{j_6,k} \end{bmatrix} = 0.$$

Expanding the first row of this matrix equation gives a linear equation for d_{in}

(11) $$\sum_{i=1}^{3}\sum_{n=1}^{3}\left(y_{j_4,k}y_{j_1,k} - y_{j_1,k}y_{j_4,k}\right)d_{in} = 0.$$

Renamig j_4 to j_2 as the first row did not need the original j_2 anywhere, we have one linear equation for the nine unknowns d_{in}

(12) $$f_{j_1,j_2,k}(d_{11},d_{12},\ldots,d_{33}) = \sum_{i=1}^{3}\sum_{n=1}^{3}\left(y_{j_2k}y_{j_1,k} - y_{j_1,k}y_{j_2,k}\right)d_{in} = 0.$$

For solving nine unknowns from linear equations we need nine equations. The first six can be obtained from three populations Y_j

(13) $f_{1,2,4} = 0 \quad f_{1,2,5} = 0 \quad f_{1,3,4} = 0 \quad f_{1,3,5} = 0 \quad f_{2,3,4} = 0 \quad f_{2,3,5} = 0$

but we need four populations to get nine equations

(14) $f_{1,4,4} = 0 \quad f_{1,4,5} = 0 \quad f_{2,4,4} = 0$.

That is enough, but with four populations we get three additional equations

$f_{1,4,5} = 0 \quad f_{3,4,4} = 0 \quad f_{3,4,5} = 0$.

Thus, the elements d_{in} are overdetermined. Assuming that the four populations do derive from three ancestral populations, these additional equations should be roughly fulfilled. We cannot expect a perfect match since populations do not only admix with each others, they also evolve by random drift, especially genetic bottlenecks. (The PCA data is typically taken form SNPs that are not acted on by selection.)

When we have the elements d_{ij} we can invert the matrix D and get A, the first three PCA values for the three ancestral populations. For the last two PCA values, we calculate $D_{1,2,3}$ from (7) and insert it to (9).

This method requires inverting the nine dimensional matrix created by (13) and (14) and two inversions of a 3x3 matrix. That is not so bad. Yes, I think I would proceed like this. Then getting the ancestral populations to Europeans, I would notice that Finns are not quite explained as a linear combination of the three populations and I would add another population and a genetic bottleneck. Something like that, hope this is enough for Justwondering. I originally thought of taking the k15-pca data and calculating numeric values, but programming a matrix inversion (I do not have it in my own program that I use to do everything) is some work. I will not do it right now.