Principal Components (Chapter 12)

The idea of principal components is to find linear combinations of variables that explain variation in the data.

Typically, we have a single sample and many variables, all of which are considered random.

Principal components is generally used more to describe the data rather than doing inference, and so doesn't assume that the data are multivariate normal, although the ideas are easier to visualize when the data is multivariate normal and 2-dimensional.

Principal Components

A crude example for the chile data, is that if we just looked at length and width, we might construct two new variables:

$$size = length + width$$

$$shape = length - width$$

Here, we've transformed the two variables of *length* and *width* into two new variables, *size* and *shape*. This doesn't reduce the dimensions of the data, but these two new variables might give a nice way to interpret the variation in the data, and they don't lose any of teh information in the original data.

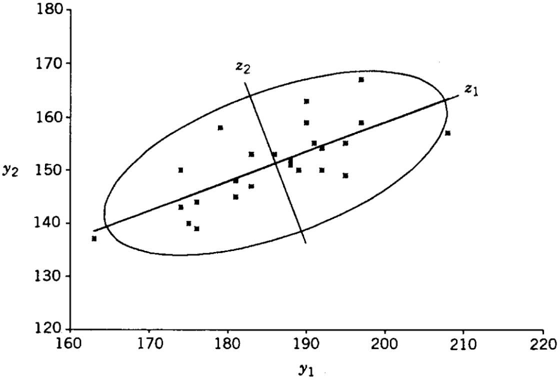
Typically, with principal components, we transform n observations of p variables into n observations of a new set of p variables, where the new variables are linear combinations of the old variables. The coefficients will be real numbers, and not usually as easy to interpret as 1 and -1.

Principal Components

The goal of principal compoents is not to choose new variables yourself, but rather to let the principal components use the data to determine the best linear combinations of the original variables.

For bivariate data, the “best” linear combinations create new axes which go through the ellipsoidal cloud of data. One axis goes through the major, longer axis, of the ellipse, while the other axis goes through the minor, shorter axis of the ellipse. The two axes are orthogonal. Usual we hope that the new linear combinations can be roughly interpreted, for example as an average or as a contrast between aspects of the variables.

Chile data from Chimayo



Chile data from Chimayo

```
> y <- x[23:33,2:3] #subset for Chimayo and
# only length and width
> plot(y)
> plot(y,cex.axes=1.3,cex.lab=1.3)
> a <- prcomp(y,scale=TRUE)
> a
Standard deviations:
[1] 1.2385587 0.6826216

Rotation:
              PC1        PC2
Length -0.7071068 -0.7071068
Width  -0.7071068  0.7071068
> summary(a)
Importance of components:
                         PC1    PC2
Standard deviation     1.239 0.6826
Proportion of Variance 0.767 0.2330
```

Chile data from Chimayo

Another function that does principal components in R is in the stats library and is called princomp() but works similarly. To scale the data, you have to scale it yourself, which you can do in the function call.

```
> b <- princomp(scale(y))
> b$loadings

Loadings:
       Comp.1 Comp.2
Length -0.707  0.707
Width  -0.707 -0.707

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

Chile data from Chimayo

Here the Rotation matrix is in `b$loadings` (the coefficients to transform the variables are sometimes called factor loadings). This gives a slightly different matrix from the `prcomp()` function, with a change in the sign of the bottom right coefficient. This doesn't affect the interpretation much. Shape here is measured by $Length - Width$ instead of $Width - Length$, so one is just the negative of the other. However, this matrix actually corresponds to a rotation matrix in the linear algebra sense, so I would use this instead of the `prcomp()` function to get the rotation matrix.

Chile data from Chimayo

There are two principal components which are the following (based on `princomp()`:

$$PC1 = -\frac{1}{\sqrt{2}} Length - \frac{1}{\sqrt{2}} Width$$

$$PC2 = \frac{1}{\sqrt{2}} Length - \frac{1}{\sqrt{2}} Width$$

Essentially, the first principle component is minus the overall size measure we had earlier, but scaled, and the second principle component is the shape also scaled by $\sqrt{2}$.

Chile data from Chimayo

A common thing to do is to look at the proportion of the variance due to each component. This is output directly using `prcomp()` but not using `princomp()`. To get this information from `princomp()`, you can use

```
> b <- princomp(scale(y))
> cumsum(b$sdev^2)/sum(b$sdev^2)
   Comp.1    Comp.2
0.7670139 1.0000000
```

Note that the sum of the squared coefficients is $(1/\sqrt{2})^2 + (1/\sqrt{2})^2 = 1$. Also, looking at the output for the importance of the components, we see that the first principle component accounts for about 77% of the variance, and the second principle component accounts for about 23% of the variance.

Chile data from Chimayo

For the R code, the option `scale.=TRUE` divides observations by their standard deviation for the variable, making the sample standard deviation equal to 1, which is considered advisable, but the default is to not scale the data. Not scaling the data will change the coefficients as well as the relative importance of the variables. The option `center=TRUE` centers the variables to each have mean 0, but doesn't change the standard deviation of the variables.

Chile data from Chimayo

```
> b <- prcomp(y)
> b
Standard deviations:
[1] 2.3344877 0.4611584

Rotation:
              PC1        PC2
Length -0.9913864 -0.1309693
Width  -0.1309693  0.9913864
```

Chile data from Chimayo

We could scale the data outside of the `prcomp` function.

```
> y2 <- scale(y)
> y2
        Length     Width
23  0.51832106  0.6363636
24  1.16622237  0.6363636
25 -0.12958026 -2.0909091
...
> c <- prcomp(y2)
> c
Standard deviations:
[1] 1.2385587 0.6826216

Rotation:
              PC1        PC2
Length -0.7071068 -0.7071068
Width  -0.7071068  0.7071068
```

Chile data from Chimayo

Chile data from Chimayo

The Rotation matrix gives the cosine of the angle between the x axis (the first row) and the first principal component (the first column). Thus

$$\cos\theta = -\frac{1}{\sqrt{2}}\theta \Rightarrow \theta = \cos^{-1}\frac{1}{\sqrt{2}}$$

So what was the angle of rotation?

Chile data from Chimayo

Admittedly, it is easy to forget trigonometry for statisticians...but the answer is that $\cos^{-1} \frac{1}{\sqrt{2}} = 3\pi/4 = 135^{\circ}$.

Chile data from Chimayo

To check this with the chile data, we can try rotating the data. When the standard rotation matrix for points in $\mathbb{R}^2$ from linear algebra has $\theta = 3\pi/4$ plugged in, we get

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \mathbf{R} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}$$

Thus, a rotated data point $(Length_i, Width_i)$ is

$$\frac{1}{\sqrt{2}} \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} Length_i \\ Width_i \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -Length_i - Width_i \\ Length_i - Width_i \end{pmatrix}$$

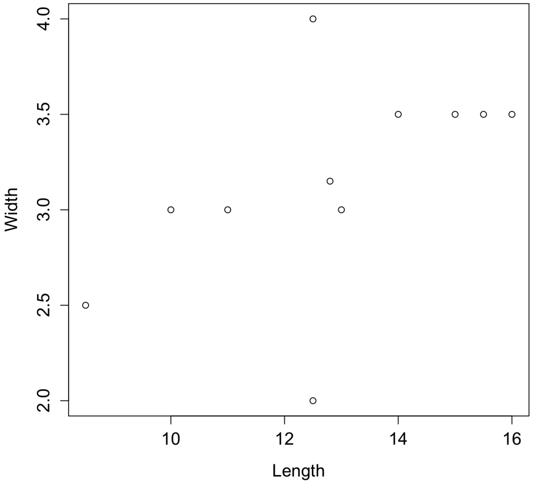
Chile data from Chimayo

Thus, the rotated data is equivalent to taking the rotation matrix $\mathbf{R}$, equivalent to the loadings matrix output from `princomp()` and multiplying by the matrix of the data. Thus

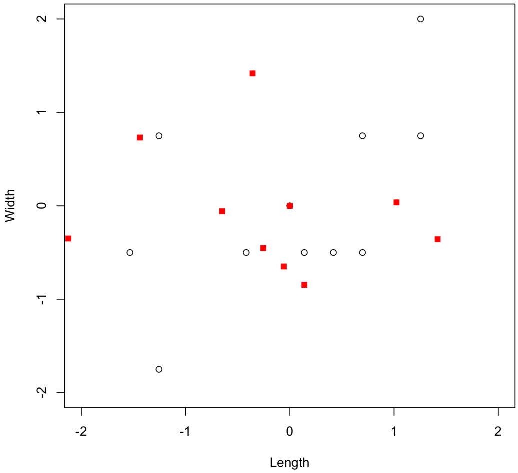
$$\mathbf{R}\mathbf{Y}'$$

gives the rotated data, where the new x-axis is the measure of size, and the new y axis is the measure of shape.

Chile data from Chimayo



Chile data from Chimayo with rotated data



Principal components

The point of finding how to rotate the data is to find new x and y axes such that the new x axis is a linear combination of the variables such that it is has the highest variance. The new y axis is then the linear combination of variables that is orthogonal to the x axis.

The point of this is largely to figure out where most of the variability in the data lies. For bivariate data, once the data is transformed, you rotate by a multiple of 45 degrees and scale by $\frac{1}{\sqrt{2}}$. If you don't scale, you might get different numbers. However, principal compoenents is sensitive to the scaling, so scaling is often done. This is particularly important if the data are on different scales.

Chile data from Chimayo

Here, I'll try principal components on the three variables of *Length*, *Width*, and *Thickness*.

```
> y <- x[23:33,2:4]
> b <- princomp(y)
Standard deviations:
   Comp.1    Comp.2    Comp.3
1.4130783 0.7837540 0.3408988

 2  variables and  11 observations.
> b$loadings

> b$loadings

Loadings:
          Comp.1 Comp.2 Comp.3
Length    -0.474  0.863  0.177
Width     -0.641 -0.200 -0.741
Thickness -0.604 -0.465  0.648
```

Chile data from Chimayo

If we use `prcmp()` instead, then we get

```
> c <- prcomp(scale(y))
> c
Standard deviations:
[1] 1.4820490 0.8220082 0.3575377

Rotation:
                 PC1        PC2        PC3
Length    -0.4738953  0.8625419 -0.1773266
Width     -0.6411163 -0.1999115  0.7409490
Thickness -0.6036499 -0.4648191 -0.6477268
```

The standard deviations are a bit different, but similar, and the rotations (loadings) are the same except that PC3 is is multiplied by a factor of -1. `princomp()` and `prcmp()` use different algorithms, but the total proportion of variance is the same.

Chile data from Chimayo

```
> summary(b) # from princomp()
Importance of components:
                          Comp.1    Comp.2     Comp.3
Standard deviation     1.4130783 0.7837540 0.34089882
Proportion of Variance 0.7321565 0.2252325 0.04261107
Cumulative Proportion  0.7321565 0.9573889 1.00000000
> summary(c) # from prcomp()
Importance of components:
                          PC1    PC2     PC3
Standard deviation     1.4820 0.8220 0.35754
Proportion of Variance 0.7322 0.2252 0.04261
Cumulative Proportion  0.7322 0.9574 1.00000
```

Chile data from Chimayo

Note that with three variables, the principal components are harder to interpret. For PC1, all variables have the same sign and have similar magnitudes, so this is similar to to taking a sum or average of the three variables, and is still a measure of size. The second PCA contrasts length with width and thickness, and could still be a measure of shape. Here a thin chile pepper will have a larger PC2 if it also not very thick. A long but wide and thick chile pepper will have PCA2 close to 0. A long, narrow, and thin chile pepper will have a large PCA2.

PCA3 contrasts width with length and thickness, so a wide, short, and thin-walled pepper will have a large PCA3. However, PCA3 contributes little to the overall variation in the chile data for Chimayo. A dimension reduction technique is to only use PCA1 and PCA2 and ignore PCA3. Since PCA1 and PCA2 use all three variables, this allows linear combinations of the three variables to contribute to a two-dimensional represenation of the data.

Principal components: matrix approach

We'll take a look at what's going on with principal components from a matrix point of view.

First, in PCA, we find the rotation of the axes (after centering) that leads to maximal variance along the first axes (the axis of the first principal component). The observation vectors $\mathbf{y}_i$ will be assumed to have already been centered, so that $\overline{\mathbf{y}}_i = \mathbf{0}$.

PCA: matrix approach

The rotation of the centered data is done by a $p \times p$ orthogonal matrix $\mathbf{A}$ (meaning that columns are orthogonal so that dot products of distinct columns are equal to 0) and for a column $\mathbf{a}_i$, we have $\mathbf{a}_i'\mathbf{a}_i = 1$. We can also say that $\mathbf{A}'\mathbf{A} = \mathbf{I}$.

We then let $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$. Then

$$\mathbf{z}_i'\mathbf{z} = (\mathbf{A}\mathbf{y}_i)'\mathbf{A}\mathbf{y}_i = y_i'\mathbf{A}'\mathbf{A}y_i = \mathbf{y}_i'\mathbf{y}_i$$

Thus, the rotated observation vectors $\mathbf{z}_i$ have the same distance to the origin as the observation vectors $\mathbf{y}_i$.

PCA: matrix approach

The new variables $\mathbf{z}_i$ must be uncorrelated (i.e., we rotate the data so that the cloud is no longer tilted), which means that the covariance matrix for $\mathbf{z}$ has 0s on the off diagonal. If the covariance matrix for $\mathbf{y}$ is $\mathbf{S}$, then

$$cov(\mathbf{z}) = cov(\mathbf{Ay}) = \mathbf{ASA}' = \begin{pmatrix} s_{z_1}^2 & 0 & \cdots & 0 \\ 0 & s_{z_2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{z_p}^2 \end{pmatrix}$$

PCA: matrix approach

The sample variances of $\mathbf{z}_i$ are the eigenvalues of $\mathbf{S}$ with $\mathbf{a}_i$ being the eigenvectors of $\mathbf{S}$. Thus, we get

$$s_{z_i}^2 = \lambda_i$$

This is a general property for orthogonal matrices, that if $\mathbf{C}$ is orthogonal, then $\mathbf{C}'\mathbf{S}\mathbf{C} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Thus, for $\mathbf{C} = \mathbf{A}'$, we get the result.

The proportion of variance explained by the first k principle components is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

The denominator can also be represented by $\mathrm{tr}(\mathbf{S})$. If the proportion variance explained by the first k components is large, then it is reasonable to represent the p-dimensional data using the first k principal components, meaning that the data is nearly k-dimensional. For example, if you had (x, y, z) coordinates of houses in a city, then this is strictly speaking 3-dimensional, but if elevation is relatively flat, the third dimension

PCA: matrix approach

A more algebraic (but still matrix-based) approach is to note that the sample variance for a linear combination $z = \mathbf{a}'\mathbf{y}$ is $\mathbf{a}'\mathbf{S}\mathbf{a}$. The goal is to find the linear combination $\mathbf{a}$ that maximizes

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}$$

and this is found by solving

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$$

PCA: matrix approach

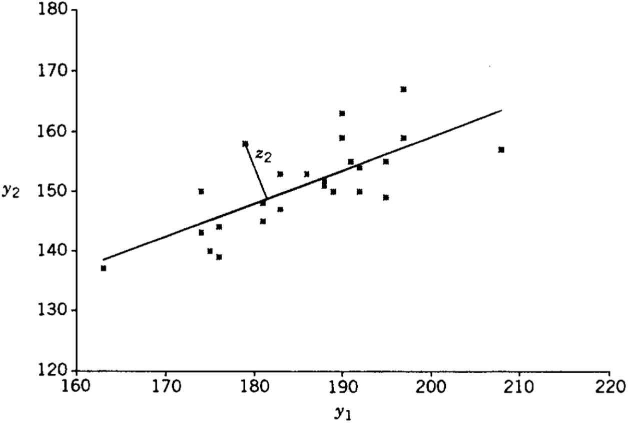
PCA can handle the case that there are more variables than observations ($p > n$) because the inverse of $\mathbf{S}$ is not needed, so that $\mathbf{S}$ can be singular. In this case, some eigenvalues are 0 and can be ignored.

PCA: matrix approach

A nice interpretation of the first principle component for two-dimensional data is that it minimizes the sum of perpendicular distances from the observations to the first principal component axis.

This is contrast to linear regression, which minimizes the sum of vertical distances from points to the regression line. Why should you minmize vertical distances rather than perpendicular distances?

PCA: perpendicular distances



PCA: perpendicular distances

When you rotate the data, the perpendicular distance from (y_{1i}, y_{2i}) to the line becomes the vertical distance from the point to the z_1 axis, which is z_{2i} (the z_2 coordinate of the ith data point. Thus, the sum of the squared perpendicular distances is (assuming **y** is already centered)

$$\begin{aligned}\sum_{i=1}^{n} z_{2i}^2 &= (\mathbf{a}_2'\mathbf{y}_i)'(\mathbf{a}_2'\mathbf{y}_i) \\ &= \mathbf{a}_2' \left[\sum_i \mathbf{y}_i\mathbf{y}_i' \right] \mathbf{a}_2 \\ &= (n-1)\mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \\ &= (n-1)\lambda_2\end{aligned}$$

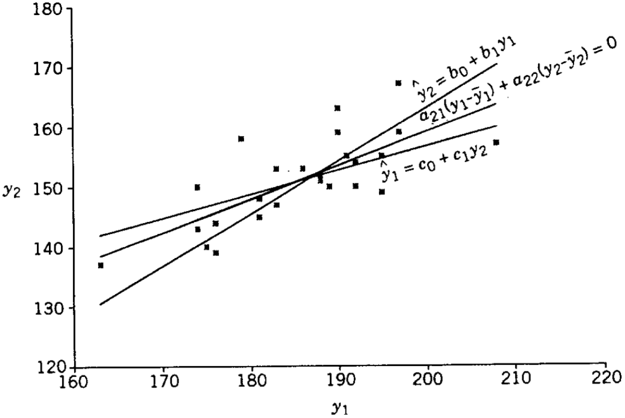
which is minimized since the first principal component is in the direction of maximal variance and the second principal component is in the direction of minimum variance.

PCA: perpendicular distances

Given two variables, such as head length versus head width, should regress length against width, or width against length? The choice seems arbitrary, but they result in different answers, and this is annoying (to me, anyway). One thing you might think of doing is to take the two regression lines, express both as say functions of head length (so regress length on width, then solve for width as a function of length). You might then take the average of the two regression lines. This way your answer doesn't treat one variable as the response more than the other one. (I have never heard of anyone doing this, so I don't recommend it.)

On the other hand, this solution is not equivalent to minimizing the perpendicular distances, although minimizing the perpendicular distances gives you a regression line that is in between the other simple linear regression lines.

PCA: perpendicular distances



Using PCA to detect outliers and nonnormality

If you have data that you want to be multivariate normal, you can use PCA to help detect outliers. In particular, since all principal components are linear combinations of the original variables, if the original data is multivariate normal, then so are the PCs, and the first two PCs will be bivariate normal. They are also independent, so you should see a cloud of points that doesn't have an angle with the z_1 axis.

It is also possible that outliers will show up in PCA plots that were hard to detect using a scatterplot matrix of the original data, or that points will cluster in ways that wouldn't be expected from multivariate normal data (this is particularly the case if data is combined from different populations).

Plotting PCs

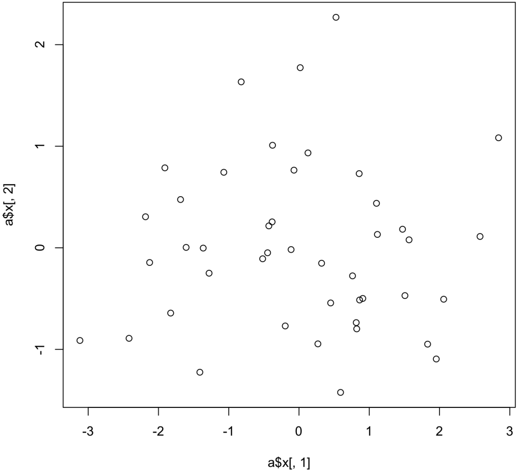
To plot PCs in R, you can run a PCA function and get the new rorated data values. For the chile data (using all locations), I did the following to plot

```
> a <- prcomp(scale(y))
> names(a)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> head(a$x)
             PC1         PC2        PC3
[1,] -0.44691332 -0.04905345  0.2214757
[2,]  0.01737361  1.77329701  0.6517343
[3,]  0.85639612  0.73028423  0.2401843
[4,]  1.11583041  0.13152051  1.5130950
[5,]  1.10281969  0.43792077  0.1154351
[6,]  0.12776230  0.93423415 -0.5312896
> a
Standard deviations:
[1] 1.4165809 0.8187573 0.5682738
```

Plotting PCs

¿ b ¡- princomp(scale(y)) ¿ head(b$scores)Comp.1Comp.2Comp.3[1,]0.446913320.04905345 - 0.2214757[2,] - 0.01737361 - 1.77329701 - 0.6517343[3,] - 0.85639612 - 0.73028423 - 0.2401843[4,] - 1.11583041 - 0.13152051 - 1.5130950[5,] - 1.10281969 - 0.43792077 - 0.1154351[6,] - 0.12776230 - 0.934234150.5312896 > head(a$x) PC1 PC2 PC3 [1,] -0.44691332 -0.04905345 0.2214757 [2,] 0.01737361 1.77329701 0.6517343 [3,] 0.85639612 0.73028423 0.2401843 [4,] 1.11583041 0.13152051 1.5130950 [5,] 1.10281969 0.43792077 0.1154351 [6,] 0.12776230 0.93423415 -0.5312896

plotting PCs: chile example

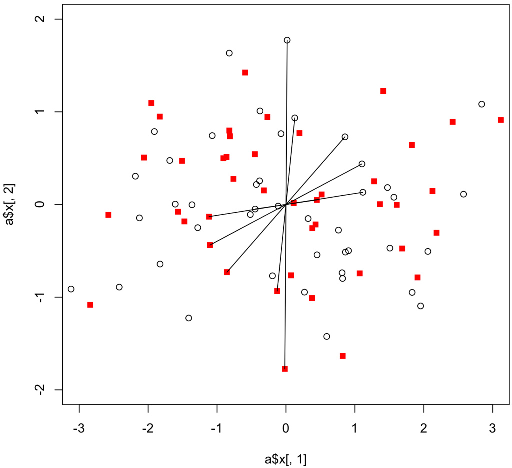


Plotting PCs: chile example

Here I plot the PCs from both `prcomp()` (black circles) and `princomp` (red squares). I show lines connecting corresponding points (for just a few points) to show that they are mirror images of each other. The code for this is

```
> plot(a$x[,1],a$x[,2],xlim=c(-3,3),ylim=c(-2,2))
> points(b$scores[,1],b$scores[,2],col="red",pch=15)
> for(i in 1:6) {
+ lines(c(a$x[i,1],b$scores[i,1]),c(a$x[i,2],b$scores[i,2]))
+ }
```

plotting PCs: chile example

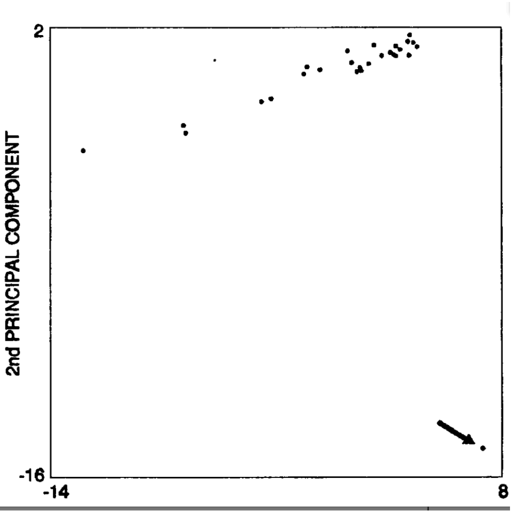


plotting PCs

There don't appear to be any obvious outliers in this data. The book gives some other interesting examples of how PCA was used to find unusual observations.

The first example has 14 economic variables obtained on 29 chemical companies. Note that the correlation between the first two PCs is 0, but in this case, the outlier is so different from the other observations, that the observations without the outlier are highly correlated.

plotting PCs: chemical companies

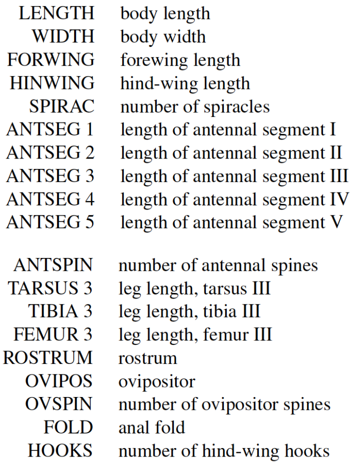


plotting PCs

The second example combines data from different samples of insects, and the question is to determine the number of distinct “taxa”, which means species, or subspecies, or populations within species. Species boundaries can be very difficult to determine, so biologists sometimes talk about “taxonomic distinctiveness” as well as “species delimitation”.

In this case, there were 40 individusal aphids sampled, and 19 variables were measured. You might expect some of these variables to be highly correlated, such as the lengths of different legs. I would guess, since there are three pairs of legs, there were two measurements taken per leg pair, one for each side, and the results were averaged. Note that you could have taken six different leg measurements, with even higher correlations.

plotting PCs: variables for the insect example



LENGTH	body length
WIDTH	body width
FORWING	forewing length
HINWING	hind-wing length
SPIRAC	number of spiracles
ANTSEG 1	length of antennal segment I
ANTSEG 2	length of antennal segment II
ANTSEG 3	length of antennal segment III
ANTSEG 4	length of antennal segment IV
ANTSEG 5	length of antennal segment V
ANTSPIN	number of antennal spines
TARSUS 3	leg length, tarsus III
TIBIA 3	leg length, tibia III
FEMUR 3	leg length, femur III
ROSTRUM	rostrum
OVIPOS	ovipositor
OVSPIN	number of ovipositor spines
FOLD	anal fold
HOOKS	number of hind-wing hooks

plotting PCs: insect example

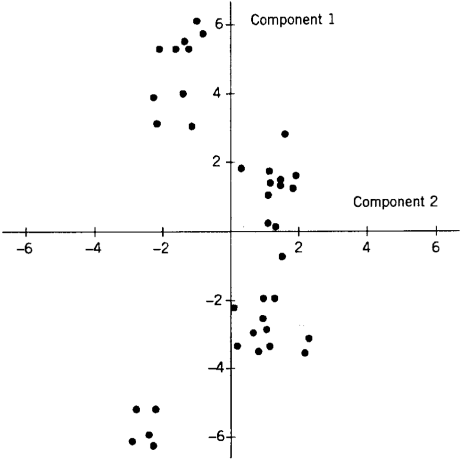


Figure 12.7. Plotted values of the first two components for individual insects.

plotting PCs

There do appear to be different clusters of points, perhaps 4 clusters. Biologists can use these types of plots to help judge whether different insects should be regarded as belonging to different species or subspecies. These days, genetic data would likely to be used to help make such judgments, but for many decades, only morphological data was available.

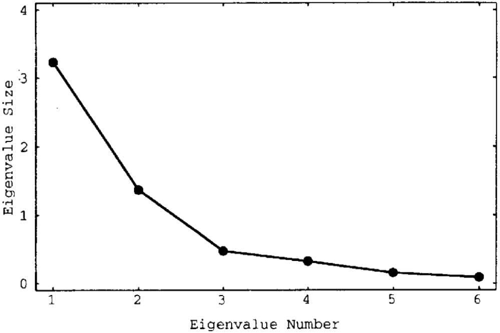
Keeping k principal components

If you are interested in dimension reduction, you might want to keep more than two principal compoents but less than p principal components. How many should you keep?

There are four typical criteria

1. Keep enough principal components so that the proportion of variance explained meets a threshold, e.g., 80% or 90%.
2. Keep components with larger than average eigenvalues, $\overline{\lambda} = (1/p)\sum_i \lambda_i$ (often used in software)
3. Use a scree graph, plotting λ_i against i, and see where there is a large break in the eigenvalues
4. Test for significance of larger components

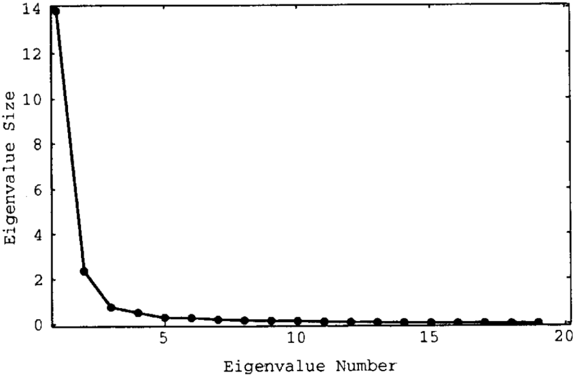
scree graph



Scree graph

The use of a scree graph is a little bit subjective, but the idea is to keep cases where the slope appears to be changing. In this example, the last four points are roughly on a line, so you would keep just the first two principal components.

scree graph: aphid example



Testing for statistical significance of components

To test whether or not to keep some of the principal components, you can test whether the last k eigenvalues are the same. If the last components have no information, then they are essentially noise, and their eigenvalues will be similar. To do this, let

$$\overline{\lambda} = \sum_{i=p-k+1}^{p} \lambda_i$$

$$u = \left(n - \frac{2p+11}{6}\right)\left(k \ln \overline{\lambda} - \sum_{i=p-n+1}^{p} \ln \lambda_i\right)$$

Then u is approximately χ^2 with degrees of freedom $(k-1)(k+2)/2$ if the original data is multivariate normal. The assumption of multivariate normality is needed for testing, but not for principal components generally, and the first three methods are often used instead of this formal testing procedure. The testing procedure often leads to retaining more components than more informal procedures.

Using the last few principal components

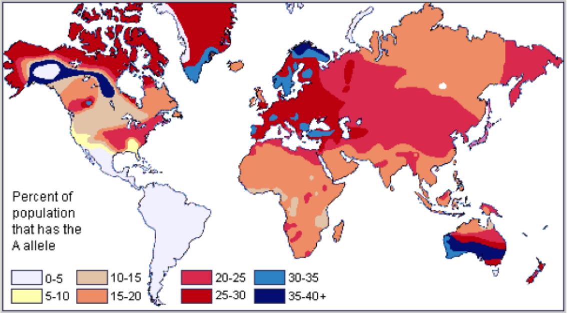
Often the last few principal components are ignored, but if the variance is close to 0 for the last (few) principal component(s), this suggests multicollinearity in the data, which could be useful to know in a regression problem.

Application of PCA to genetic data

PCA is very frequently applied to different types of genetic data. This was first done in a paper from nearly 40 years ago: Menozzi P, Piazza A, Cavalli-Sforza L (1978) "Synthetic maps of human gene frequencies in Europeans". *Science* 201: 786792.

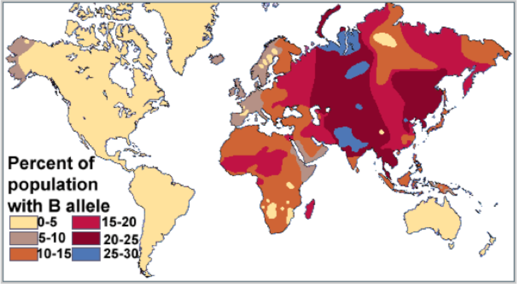
Many different types of genetic data can be used for PCA, including frequencies of different genotypes for different genes (for example, the frequency of the type O allele, the frequency of the type A allele for blood type). Gene frequencies are quantitative variables (like leg length for insects) and tend to vary for different populations even within species. When many genes are examined together, populations can tend to cluster, much like the insect example showed clustering into 4-5 groups.

Blood type frequencies



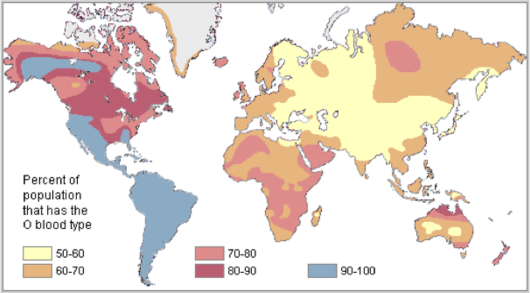
Map of allele A among native populations

Blood type frequencies



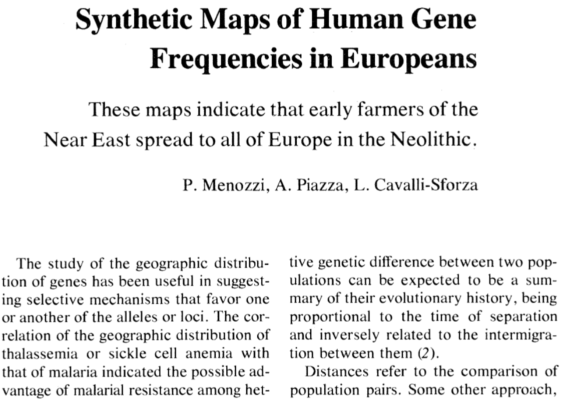
Map of allele B among native populations

Blood type frequencies



Map of allele O among native populations

PCA genetics paper



Synthetic Maps of Human Gene Frequencies in Europeans

These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic.

P. Menozzi, A. Piazza, L. Cavalli-Sforza

The study of the geographic distribution of genes has been useful in suggesting selective mechanisms that favor one or another of the alleles or loci. The correlation of the geographic distribution of thalassemia or sickle cell anemia with that of malaria indicated the possible advantage of malarial resistance among het-

tive genetic difference between two populations can be expected to be a summary of their evolutionary history, being proportional to the time of separation and inversely related to the intermigration between them (*2*).

Distances refer to the comparison of population pairs. Some other approach,

Application of PCA to genetic data

For blood type, there are three alleles, different versions of the gene: A, B, and O. You carry two copies of each gene in each cell (except for sex cells — egg or sperm, which only carry one copy per cell), and the combination determines your genotype. Frequently, geneticists use the frequencies of the individual alleles in the population. Type O is the most frequent for essentially all human populations, but the frequency still varies. For example, in the Middle East, the frequency of O is lower than in Europe, and the frequency of type O in Europe is lower than in the Americas.

A data set of gene frequencies would have separate columns for different alleles, and these will be highly correlated. Given the frequencies of say, A and B, the frequency of O is not needed since $f_A + f_B + f_O = 1$, but the frequencies of just A and B are also correlated. (Question: do you think they are positively or negatively correlated?)

Application of PCA to genetic data

In addition to blood type, you could also use other genetic markers, such as frequencies of being Rh negative or Rh positive (another aspect of blood), and the *MN* blood type (just another blood type marker. Early uses of PCA often used genetics related to blood samples, so these types of markers were used.

Instead of plotting the frequency of a single allele and needing separately plots for each allele, such as in the Wikipedia plots, the idea is to find a linear combination of the allele frequencies that will create the largest amount of variation between populations. In this case, the value of each principle component is plotted as a function of the geographical location, and placed on a map.

PCA genetics paper

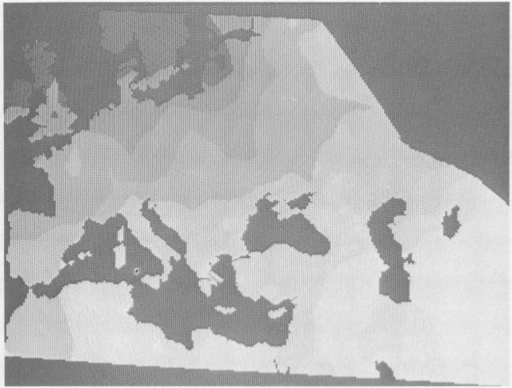


Fig. 1. The first principal component of gene frequencies from 38 independent alleles at the human loci: ABO, Rh, MNS, Le, Fy, Hp, PGM$_1$, HLA-A, and HLA-B. Shades indicate different intensities of the first principal component, which accounts for 27 percent of the total variation and is represented with green shades in the photograph on the cover.

PCA genetics paper

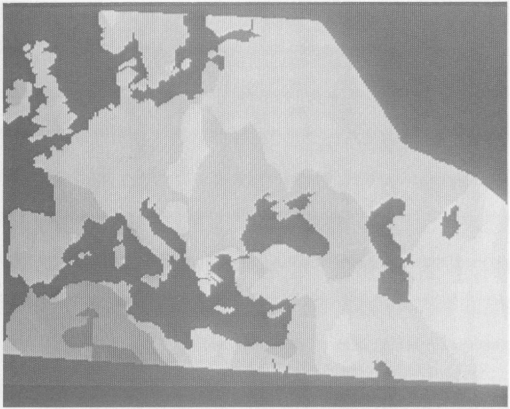


Fig. 2. The second principal component of gene frequencies from 38 independent alleles at the human loci: ABO, Rh, MNS, Le, Fy, Hp, PGM_1, HLA-A, and HLA-B. Shades indicate different intensities of the second principal component which accounts for 18 percent of the total variation and is represented with blue shades on the cover.

PCA genetics paper

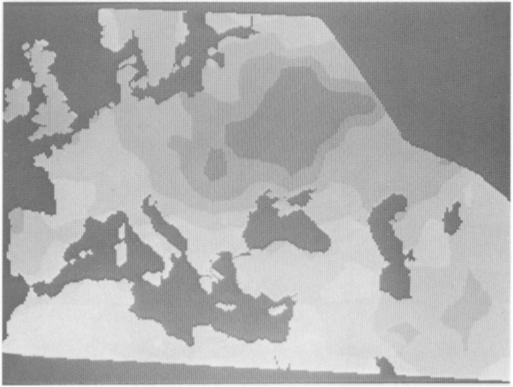


Fig. 3. The third principal component of gene frequencies from 38 independent alleles at the human loci: ABO, Rh, MNS, Le, Fy, Hp, PGM$_1$, HLA-A, and HLA-B. Shades indicate different intensities of the third principal component which accounts for 11 percent of the total variation and is represented with red shades on the cover.

PCA for genetics

At the time of this study (1978), it was difficult to get many genetic markers, but since genetic technology has improved, we can now get genetic information across the whole genome. PCA today in genetics is often applied to individual letters of DNA instead of frequencies of alleles. This gives more detailed information because there might be several genetic variants of an allele that can be distinguished genetically but not phenotypically (not all O alleles necessarily have exactly the same DNA sequence, so we can just use the sequence instead of noting the allele).

PCA for genetics

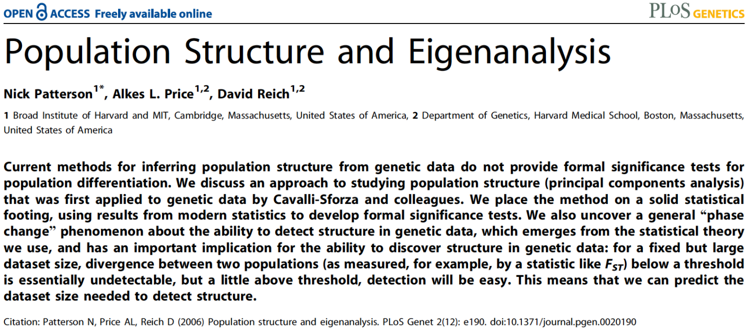
In addition, we can use genomic locations that don't correspond to functional genes and have no phenotypic results at all. It is possible to find over 1 million positions in the genome (out of over 3 billion letters of DNA in our genomes) where there is some variability. These are often called SNPs (Single Nucleotide Polymorphisms). DNA can be represented by sequences of four letters: A, C, G, and T. A SNP is a specific location, such as position 131,007 on the long arm of chromosome 7, where there is some variability so that some humans have one DNA letter and other humans have another DNA letter. For roughly 99.9% of genomic locations, there is no variability, and all humans have the same DNA letter. But with 3 billion letters, 0.1% of locations having variability means that there are a few million locations where there is variability. This also means that you need on average to observe about 1000 DNA letters from two individuals to find any genetic differences between them.

PCA for genetics

A paper in 2006 advocated using PCA with SNP data and was able to do so efficiently for hundreds of thousands of SNPS on 1000s of individuals. This was a bit different from the 1978 approach because: (1) the variables were binary rather than continuous, (2) they used individuals rather than summaries from populations.

Although you can use summaries of SNP frequencies from different populations, this doesn't tell you if some individuals have unusual patterns. In particular, the authors of the 2006 paper were interested in admixed populations — people with ancestry from geographically distinct areas. Using individuals is also useful when the population that an individual is from is unclear.

PCA genetics paper



OPEN ACCESS Freely available online

PLoS GENETICS

Population Structure and Eigenanalysis

Nick Patterson[1*], **Alkes L. Price**[1,2], **David Reich**[1,2]

1 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, **2** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

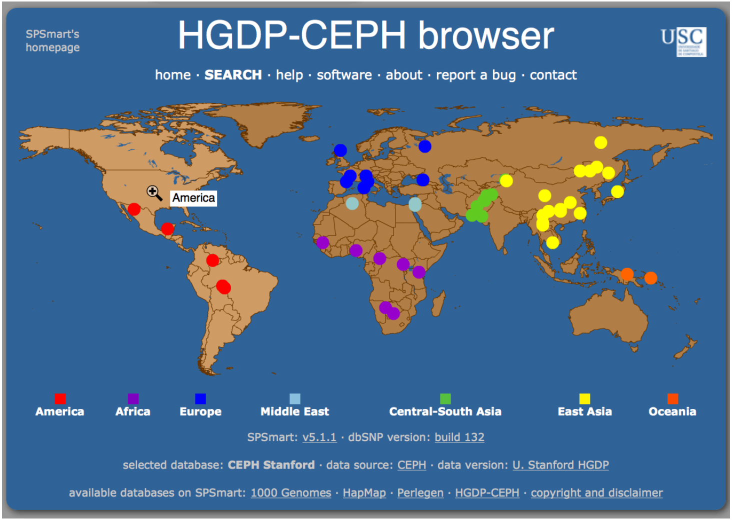
Current methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation. We discuss an approach to studying population structure (principal components analysis) that was first applied to genetic data by Cavalli-Sforza and colleagues. We place the method on a solid statistical footing, using results from modern statistics to develop formal significance tests. We also uncover a general "phase change" phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory we use, and has an important implication for the ability to discover structure in genetic data: for a fixed but large dataset size, divergence between two populations (as measured, for example, by a statistic like F_{ST}) below a threshold is essentially undetectable, but a little above threshold, detection will be easy. This means that we can predict the dataset size needed to detect structure.

Citation: Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190

PCA genetics paper

This paper applied principal components to the CEPH-HGDP (Centre dEtude du Polymorphisme Human Genetic Diversity Project) data, a famous data set which intensively sampled 1050 individuals from 52 populations around the world. The data set has nearly 1 million SNPs and other types of genetic markers, including microsatellites (short segments of DNA that get repeated a variable number of times), insertion/deletion events, and CNV (copy-number variants, longer stretches of DNA that occur a variable number of times).

PCA paper: genetic structure in East Asia



SPSmart's homepage

HGDP-CEPH browser

USC

home · **SEARCH** · help · software · about · report a bug · contact

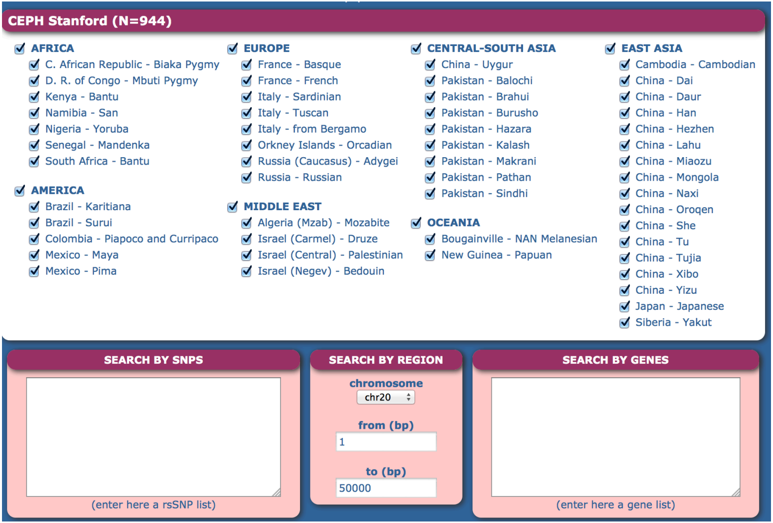
🟥 America 🟪 Africa 🟦 Europe ▨ Middle East 🟩 Central-South Asia 🟨 East Asia 🟧 Oceania

SPSmart: v5.1.1 · dbSNP version: build 132

selected database: **CEPH Stanford** · data source: CEPH · data version: U. Stanford HGDP

available databases on SPSmart: 1000 Genomes · HapMap · Perlegen · HGDP-CEPH · copyright and disclaimer

SNPs in HGDP-browser



CEPH Stanford (N=944)

☑ **AFRICA**
- ☑ C. African Republic - Biaka Pygmy
- ☑ D. R. of Congo - Mbuti Pygmy
- ☑ Kenya - Bantu
- ☑ Namibia - San
- ☑ Nigeria - Yoruba
- ☑ Senegal - Mandenka
- ☑ South Africa - Bantu

☑ **AMERICA**
- ☑ Brazil - Karitiana
- ☑ Brazil - Surui
- ☑ Colombia - Piapoco and Curripaco
- ☑ Mexico - Maya
- ☑ Mexico - Pima

☑ **EUROPE**
- ☑ France - Basque
- ☑ France - French
- ☑ Italy - Sardinian
- ☑ Italy - Tuscan
- ☑ Italy - from Bergamo
- ☑ Orkney Islands - Orcadian
- ☑ Russia (Caucasus) - Adygei
- ☑ Russia - Russian

☑ **MIDDLE EAST**
- ☑ Algeria (Mzab) - Mozabite
- ☑ Israel (Carmel) - Druze
- ☑ Israel (Central) - Palestinian
- ☑ Israel (Negev) - Bedouin

☑ **CENTRAL-SOUTH ASIA**
- ☑ China - Uygur
- ☑ Pakistan - Balochi
- ☑ Pakistan - Brahui
- ☑ Pakistan - Burusho
- ☑ Pakistan - Hazara
- ☑ Pakistan - Kalash
- ☑ Pakistan - Makrani
- ☑ Pakistan - Pathan
- ☑ Pakistan - Sindhi

☑ **OCEANIA**
- ☑ Bougainville - NAN Melanesian
- ☑ New Guinea - Papuan

☑ **EAST ASIA**
- ☑ Cambodia - Cambodian
- ☑ China - Dai
- ☑ China - Daur
- ☑ China - Han
- ☑ China - Hezhen
- ☑ China - Lahu
- ☑ China - Miaozu
- ☑ China - Mongola
- ☑ China - Naxi
- ☑ China - Oroqen
- ☑ China - She
- ☑ China - Tu
- ☑ China - Tujia
- ☑ China - Xibo
- ☑ China - Yizu
- ☑ Japan - Japanese
- ☑ Siberia - Yakut

SEARCH BY SNPS

(enter here a rsSNP list)

SEARCH BY REGION

chromosome chr20

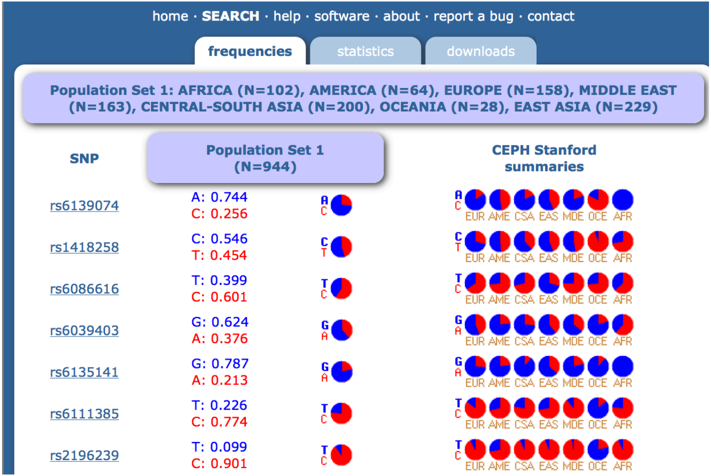
from (bp) 1

to (bp) 50000

SEARCH BY GENES

(enter here a gene list)

SNPs in HGDP-browser



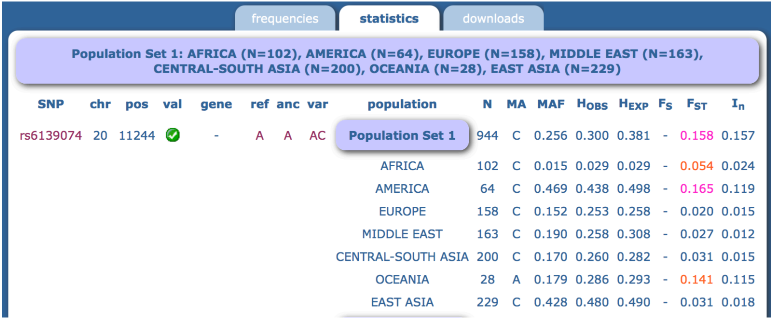
home · **SEARCH** · help · software · about · report a bug · contact

frequencies | statistics | downloads

Population Set 1: AFRICA (N=102), AMERICA (N=64), EUROPE (N=158), MIDDLE EAST (N=163), CENTRAL-SOUTH ASIA (N=200), OCEANIA (N=28), EAST ASIA (N=229)

SNP	Population Set 1 (N=944)		CEPH Stanford summaries
rs6139074	A: 0.744 C: 0.256	A C	A C EUR AME CSA EAS MDE OCE AFR
rs1418258	C: 0.546 T: 0.454	C T	C T EUR AME CSA EAS MDE OCE AFR
rs6086616	T: 0.399 C: 0.601	T C	T C EUR AME CSA EAS MDE OCE AFR
rs6039403	G: 0.624 A: 0.376	G A	G A EUR AME CSA EAS MDE OCE AFR
rs6135141	G: 0.787 A: 0.213	G A	G A EUR AME CSA EAS MDE OCE AFR
rs6111385	T: 0.226 C: 0.774	T C	T C EUR AME CSA EAS MDE OCE AFR
rs2196239	T: 0.099 C: 0.901	T C	T C EUR AME CSA EAS MDE OCE AFR

SNPs in HGDP-browser



Population Set 1: AFRICA (N=102), AMERICA (N=64), EUROPE (N=158), MIDDLE EAST (N=163), CENTRAL-SOUTH ASIA (N=200), OCEANIA (N=28), EAST ASIA (N=229)

SNP	chr	pos	val	gene	ref	anc	var	population	N	MA	MAF	H_{OBS}	H_{EXP}	F_S	F_{ST}	I_n
rs6139074	20	11244	✓	-	A	A	AC	Population Set 1	944	C	0.256	0.300	0.381	-	0.158	0.157
								AFRICA	102	C	0.015	0.029	0.029	-	0.054	0.024
								AMERICA	64	C	0.469	0.438	0.498	-	0.165	0.119
								EUROPE	158	C	0.152	0.253	0.258	-	0.020	0.015
								MIDDLE EAST	163	C	0.190	0.258	0.308	-	0.027	0.012
								CENTRAL-SOUTH ASIA	200	C	0.170	0.260	0.282	-	0.031	0.015
								OCEANIA	28	A	0.179	0.286	0.293	-	0.141	0.115
								EAST ASIA	229	C	0.428	0.480	0.490	-	0.031	0.018

SNPs in HGDP-browser

The number of SNPs detected in a 50 kb window (i.e., 50,000 letter stretch of DNA) was 15, a bit less than 0.1%; however, this data set sampled a large number of SNPs but not the entire genomes of these individuals. The SNPs chosen were based on earlier efforts to find SNPs in the CEPH data set which had a smaller number of individuals and populations sampled.

The original CEPH data was based on: "The DNA samples for the HapMap will come from a total of 270 people: from the Yoruba people in Ibadan, Nigeria (30 both-parent-and-adult-child trios), Japanese in Tokyo (45 unrelated individuals), Han Chinese in Beijing (45 unrelated individuals), and the CEPH (30 trios). " (http://hapmap.ncbi.nlm.nih.gov/abouthapmap.html). Thus, there is some ascertainment bias in the CEPH-HGDP data in that SNPs are more likely to be detected that came from the original populations in the CEPH study (which didn't include Native Americans, for example).

More recently, entire human genomes have been sequenced, so this should reduce any sampling bias issues.

PCA paper: genetic structure in East Asia

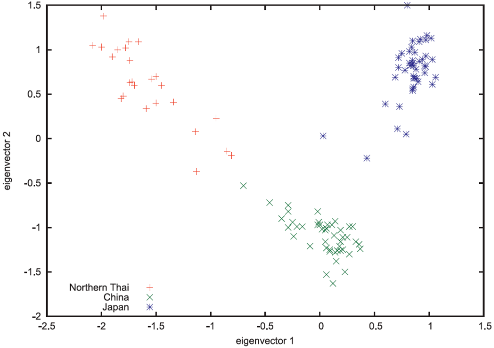


Figure 5. Three East Asian Populations

Plots of the first two eigenvectors for a population from Thailand and Chinese and Japanese populations from the International Haplotype Map [32]. The Japanese population is clearly distinguished (though not by either eigenvector separately). The large dispersal of the Thai population, along a line where the Chinese are at an extreme, suggests some gene flow of a Chinese-related population into Thailand. Note the similarity to the simulated data of Figure 8.

doi:10.1371/journal.pgen.0020190.g005

PCA paper: genetic structure in East Asia

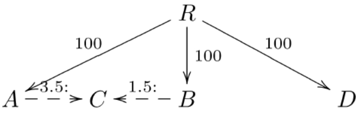


Figure 7. Simulation of an Admixed Population

We show a simple demography generating an admixed population. Populations *A,B,D* trifurcated 100 generations ago, while population *C* is a recent admixture of *A* and *B*. Admixture weights for the proportion of population *A* in population *C* are Beta-distributed with parameters (3.5,1.5). Effective population sizes are 10,000.
doi:10.1371/journal.pgen.0020190.g007

PCA paper: genetic structure in East Asia

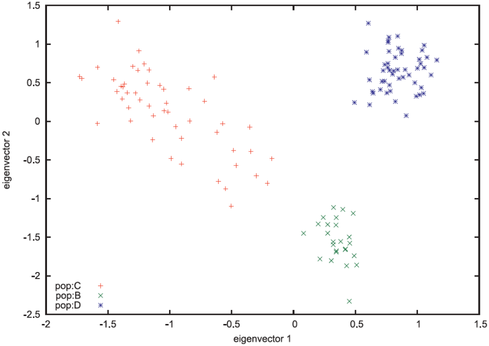


Figure 8. A Plot of a Simulation Involving Admixture (See Main Text for Details)

We plot the first two principal components. Population C is a recent admixture of two populations, B and a population not sampled. Note the large dispersion of population C along a line joining the two parental populations. Note the similarity of the simulated data to the real data of Figure 5.
doi:10.1371/journal.pgen.0020190.g008