# ON THE USE OF PRINCIPAL COMPONENTS IN CONTEMPORARY POPULATION GENETICS: A CASE STUDY

**M. Gasparini**[1]**and C. Di Gaetano**[2]
[1]Politecnico di Torino, Torino, Italy; *mauro.gasparini@polito.it*
[2] Università di Torino, Torino, Italy

## 1. Introduction

In human Population Genetics, routine applications of principal component techniques are often required. Population biologists make widespread use of certain discrete classifications of human samples into haplotypes, the monophyletic units of phylogenetic trees constructed from several single nucleotide bimorphisms hierarchically ordered. Compositional frequencies of the haplotypes are recorded within the different samples. Principal component techniques are then required as a dimension-reducing strategy to bring the dimension of the problem to a manageable level, say two, to allow for graphical analysis.

Population biologists at large are not aware of the special features of compositional data and normally make use of the crude covariance of compositional relative frequencies to construct principal components. In this short note we present our experience with using traditional linear principal components or compositional principal components based on logratios, with reference to a specific dataset.

## 2. The state of the art in human population genetics

After the recent decoding of the human genome and the latest developments in molecular biology techniques, the technology and the scope of contemporary human population genetics has deeply changed. The number of genetic markers used to describe contemporary populations in order to infer possible past patterns of human migration and colonization has dramatically increased.

Most of the interesting markers used today are single base variations occurring pervasively in the whole of the human genome and called single nucleotide polymorphisms (SNPs). For population genetic purposes, particularly useful are the ones lying on the nonrecombinant portion of the Y chromosome and on mithocondrial DNA. This is so because these portions of the genome are exempt from recombination and allow for the possibility to trace more easily recent ancestry through a paternal or maternal perspective.

Cascades of SNP-related events are studied by phylogenetic techniques in order to reconstruct mutational trees and identify their final leaves as groups of individuals with similar mutational history. These groups are called haplotypes. The frequencies of the different haplotypes in a population may then be used in the same way traditional marker frequencies are, to identify differencies and similarities between the evolutionary histories of the populations. A prime example of this technology is a paper by Underhill and others (2000), a recent review is Jobling and Tyler-smith (2003), and an online reference is site `http://ycc.biosci.arizona.edu/`

Haplotype frequencies are high dimensional data and, as such, often need dimension reduction techniques such as principal components. Principal components are very well known tools in population genetics at least since they had been employed in Menozzi and others (1978) to construct genetic maps based on frequencies of markers of the pre-molecular era. Plots of the first two principal components are informative here because samples are often coming from geographically dispersed populations and correlating geographical and genetical population structure is a primary interest of the population biologist.

The fact that genetic marker frequencies are compositional data, though, does not seem to be often emphasized in the population genetic literature. The recommendation of the compositional data literature, starting from Aitchison (1986), is that principal components for compositional data should be computed after a logratio transformation has been applied to frequencies, in order to preserve the linearity of the technique.

In this short note we present our experience on the analysis of compositional SNP frequency data. The second author is working on a set of populations which extend the results by Semino and others (2000). In order not to present prematurely unpublished data, we refer to a dataset taken from Semino and others (2000) itself, a paper which belongs to the same line of research on the Y chromosome mentioned above. In particular, it contains a map made of the first two principal components. Our experience with the analysis of this dataset is that compositional principal components add an angle to the understanding of the data, but are not strictly necessary, since linear principal components provide the bulk of the needed data reduction. The reason for this partially negative result on compositional principal components is first that, when the dimension of the composition is high, compositional principal components do not differ much from linear principal components and secondly, that it is not clear how to optimally treat structural zeros, which are likely to be numerous with SNP data.

## 3. Compositional or linear principal components

Semino and others (2000) present a table of relative frequencies of 19 Y chromosome haplotypes in 25 European and Middle Eastern samples (the samples considered are then reduced to 22 by excluding populations from the subartic area). As part of the analysis, they also build a two dimensional map, reconstructed in Figure 1, by plotting the first two linear principal components of the frequencies.

It is possible to identify on the map some recurrent themes dear to the population biologist, like the relative accuracy with which the genetic map reproduces the geographical map, the differentiation of Northern and Southern Italy and the clustering of Slavic populations.

An alternative way of computing principal components for this kind of compositional frequency data is the one recommended by Aitchison (1986) and it is based on logcontrasts, i.e. convex linear combinations of logratios of frequencies. The formulas are the same as in Aitchison (1986) and they have been implemented for this study in the statistical free software R. The results are plots in Figures 2 and 3.

The two plots are different because of the different treatment of zeros. In population genetics, it is not unusual to find zero frequencies, since certain haplotypes are strictly associated with certain geographical areas and migrational past events (see Underhill and others 2000). It is therefore of primary importance to have a way to treat zeros, since ratio and log operations can not be performed on them. A standard way is to substitute a small frequency for a zero, under the assumption that if a haplotype has not been observed, that is because of a small frequency in the population. The issue is complicated by the fact that some of the zeros in population genetics may be structural zeros, in the sense that the true frequency in the population may actually be exactly zero. Anyhow, the treatment of zeros as traces is commonplace in compositional data analysis and it has produced the two plots. For the plot of Figure 2 the number 0.9 has been substituted for zero and for the second plot in Figure 3 the number 0.001 has been used instead.

Some empirical remarks can be made on this specific case study. First, the difference between the map made with linear principal components and the first one made with compositional principal components is not large. The two plots actually convey the same bulk of information the population biologist is looking for. That is probably because as the number of members of the composition (the genetic markers in our case) grows larger, the difference between the first pairs of principal components becomes negligible. Secondly, due to the difference between the two compositional plots in Figures 2 and 3, it is not clear if a stable limit for the operation of trace substitution can be obtained as the trace goes to zero. We have not been able to prove an analytical result on the convergence of the map as the trace goes to zero, but it appears that convergence strictly depends on the number of zeros and on their configuration in the data matrix.

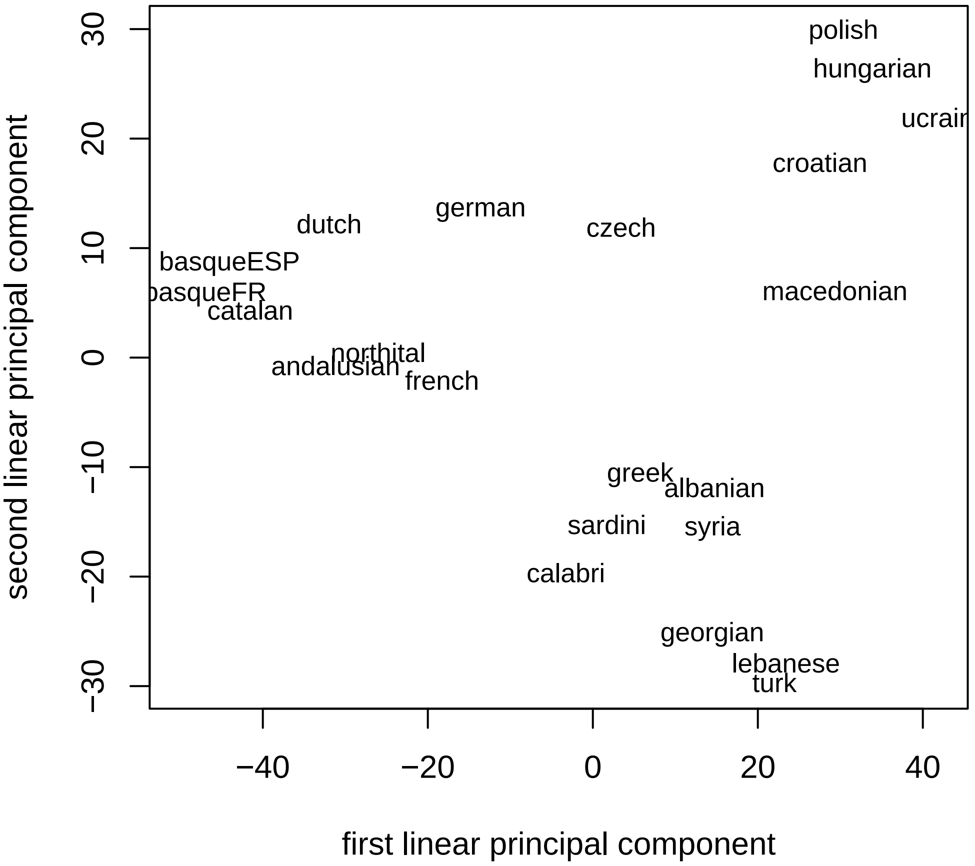

**Figure** 1. First two linear principal components

These issues should be resolved in a more positive sense if compositional principal components are to become a routine tool in population genetics. That is not to deny them any value. Preliminary exploratory results show for example that when an outgroup is introduced in the analysis, compositional principal components may be more suitable to identify it as a separate cluster and to invoke, for example, the necessity of consideration of the third principal component.

## 4. References

Aitchison J., 1986, The statistical analysis of compositional data: Chapman and Hall.

Jobling M. and Tyler-Smith C., 2003, The human Y chromosome: an evolutionary marker comes of age: Nature Reviews — Genetics, v. 4, p. 598–612.

Menozzi P., Piazza A. and Cavalli-Sforza L., 1978, Synthetic maps of human gene frequencies in Europeans: Science, v. 201, p. 786–792.

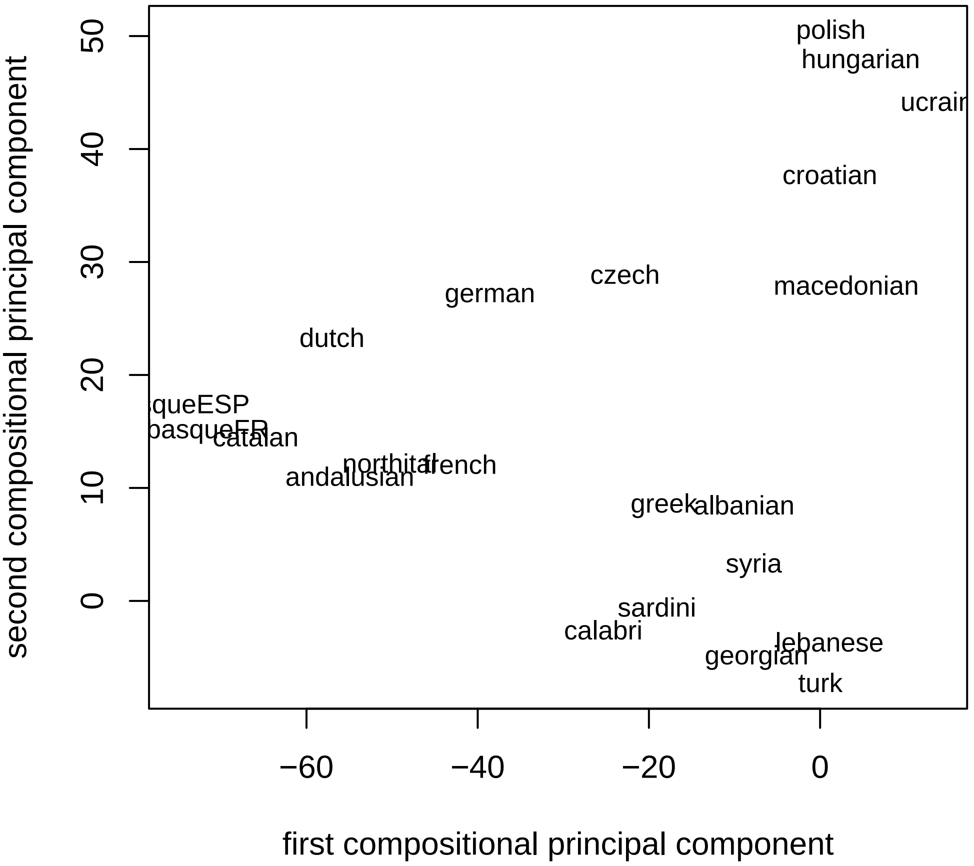

**Figure** 2. First two compositional principal components with trace=0.9

Semino O., Passarino G., Oefner P., Lin A., Arbuzova S., Beckman L., De Benedictis G., Francalacci P., Kouvatsi A., Limborska S., Marcikiæ M., Mika A., Mika B., Primorac D., Santachiara-Benerecetti A., Cavalli-Sforza L. and Underhill P., 2000, The genetic legacy of paleolithic homo sapiens sapiens in extant europeans: A Y chromosome perspective: Science, v. 290, p. 1155–1159.

Underhill P., Shen P., Lin A., Jin L., Passarino G., Yang W., Kauffman E., Bonné-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J., Mehdi S., Seielstad M., Spencer Wells R., Piazza A., Davis R., Feldman M., Cavalli-Sforza L. and Oefner P., 2000, A Y chromosome sequence variation and the history of human populations: Nature Genetics, v. 26, p. 358–361.

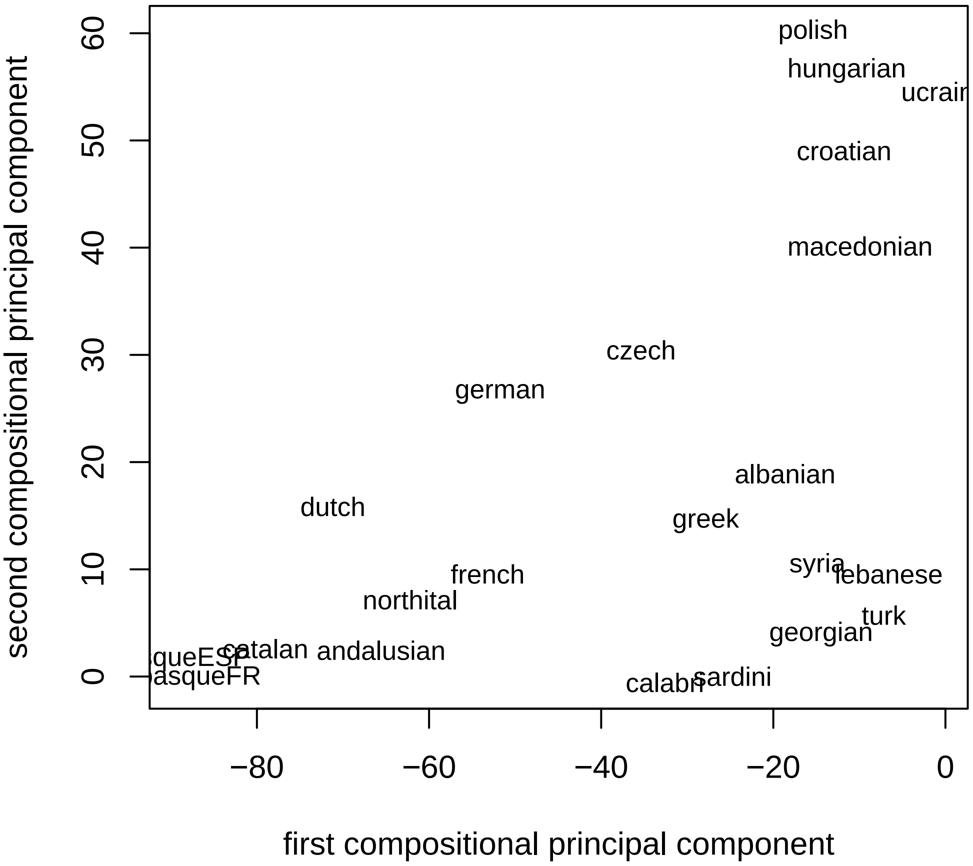

**Figure** 3. First two compositional principal components with trace=0.001