PCA, estimation & outlier detection

December 5, 2021

An initial goal of this work is to provide robust estimates of the composixtion of each rock. This problem is made difficult by the presence of both missing values (laboratories did not measure any composition) and outliers (the measures reported are extreme).

Setting aside the problem of missing values for now, one may look at a subcomposition $\mathbf{x} \in \mathbb{S}^D$ of oxides of major elements.

One begin by importing a dataset, for example GeoPT48, a monzonite.

```
setwd("/home/max/Documents/MStatistics/MA2/Thesis/Repository/")
data <- read_csv("data/raw/GeoPT48 -84Ra.csv")

##
## - Column specification ----------------------------
## cols(
##   .default = col_double(),
##   Laboratory = col_character(),
##   Au = col_logical(),
##   N = col_logical(),
##   Os = col_logical()
## )
## i Use `spec()` for the full column specifications.
```

Then, one looks at the subcomposition of major elements :

```
sel <-c("SiO2","TiO2","Al2O3","Fe2O3T","MnO","MgO","CaO","Na2O",
        "K2O","P2O5")
df.majors <- select(data,all_of(sel))
# closure operation
df.majors <- data.frame(clo(df.majors))
```

Then missing values are imputed by the column geometric mean

```
geomean.v <- sapply(rbind(df.majors),geomean)
for (i in 1:ncol(df.majors)){
  df.majors[,i][is.na(df.majors[,i])] <- geomean.v[i]
}
```

One then transform the dataset using the CLR transformation which is a mapping $\mathbb{S}^D - > \mathbb{U}^D$:

$$\mathbf{z} = clr(\mathbf{x} = [log(x_1/g(x)), .., log(x_D/g(x))] \tag{1}$$

Where $\mathbb{U}^D$ is a subspace of $\mathbb{R}^D$ defined as :

$$U^D = \left\{[u_1, .., u_D] : \sum_{i=1}^{D} = 0\right\}$$

```
cr.df <- data.frame()
clr.df <- data.frame()
# cr.df, divide each entries in a column by the geometric mean of this column
cr.df <- sweep(df.majors,MARGIN = 2,FUN="/",STATS = geomean.v)
# clr.df is the natural logarithm of cr.df. Now this dataframe contains clr component
clr.df <- log(cr.df)
```

Then principal component analysis is conducted. Z denotes the mean-centered data matrix X :

$$z_{ij} = x_{ij} - \mu_j$$

Where μ_j denotes the arithmetic mean of the j-th column. Recall that here using the arithmetic mean is justified because X now lives in a subspace of $\mathbb{R}^D$ which is no longer constrained by the unit sum.

Perform Singular Value Decomposition on clr.df :

```
pca.clr <- prcomp(clr.df,scale = T,rank. = ncol(clr.df)-1 )
summary(pca.clr)

## Importance of first k=9 (out of 10) components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5272 1.1288 1.0464 0.72257 0.59620 0.48113 0.23209
## Proportion of Variance 0.6387 0.1274 0.1095 0.05221 0.03555 0.02315 0.00539
## Cumulative Proportion  0.6387 0.7661 0.8756 0.92779 0.96334 0.98649 0.99187
##                            PC8     PC9
## Standard deviation     0.18604 0.17566
## Proportion of Variance 0.00346 0.00309
## Cumulative Proportion  0.99533 0.99842
```