Multivariate statistical analysis of metagenomic data

Boris Shilov

Introduction

The aim of this analysis is to explore the metagenomic dataset assembled by Forslund et al. (n.d.).

Data format

```r
library(tidyverse)
set.seed(100)
data = read_tsv("data.r")
data
```

```
## # A tibble: 1,178,352 x 6
##    Sample Dataset Status  FeatureType Feature Abundance
##    <chr>  <chr>   <chr>   <chr>       <chr>       <dbl>
##  1 MH0443 MHD     ND CTRL GMM         MF0119        902
##  2 MH0443 MHD     ND CTRL GMM         MF0090        302
##  3 MH0443 MHD     ND CTRL GMM         MF0097          0
##  4 MH0443 MHD     ND CTRL GMM         MF0092          0
##  5 MH0443 MHD     ND CTRL GMM         MF0011       1076
##  6 MH0443 MHD     ND CTRL GMM         MF0083          3
##  7 MH0443 MHD     ND CTRL GMM         MF0110          0
##  8 MH0443 MHD     ND CTRL GMM         MF0063          4
##  9 MH0443 MHD     ND CTRL GMM         MF0042          1
## 10 MH0443 MHD     ND CTRL GMM         MF0024        185
## # ... with 1,178,342 more rows
```

We can see that the dataset consists of six columns. The first is sample IDs, which are not unique keys since every sample may have multiple features. The second is the dataset of origin, which is either Danish (MHD), Swedish (SWE) or Han Chinese (CHN). The third column is treatment status of a particular patient, where they are either healthy (ND CTRL), are having type 2 diabetes treated with metformin included (T2D metformin+), are having type 2 diabetes treated without metformin (T2D metformin-) or are having type 1 diabetes treated (T1D). The fourth column specifies the type of feature in the fifth column. This is either gut microbial/metabolic module (GMM), SEED database annotation, bacterial family, bacterial genus or metagenomic operational taxonomic units (Motu). The fifth column contains the feature names which are coded differently depending on the feature type. Finally, the sixth column contains the abundance of a given feature.

The feature types are split into two spaces: the taxonomic space represented by the genuses, families and taxonomic units, and the functional space as annotated by SEED and GMM. The full SEED and GMM feature code annotations can be found in the appendix files, taken from Supplementary Table 10 of Forslund et al. (n.d.). We are only interested in exploring the structure of the functional space.

```r
functional_space = filter(data, FeatureType %in% c("GMM", "SEED"))
functional_space
```

```
## # A tibble: 413,168 x 6
##    Sample Dataset Status  FeatureType Feature Abundance
##    <chr>  <chr>   <chr>   <chr>       <chr>       <dbl>
##  1 MH0443 MHD     ND CTRL GMM         MF0119        902
##  2 MH0443 MHD     ND CTRL GMM         MF0090        302
##  3 MH0443 MHD     ND CTRL GMM         MF0097          0
##  4 MH0443 MHD     ND CTRL GMM         MF0092          0
##  5 MH0443 MHD     ND CTRL GMM         MF0011       1076
##  6 MH0443 MHD     ND CTRL GMM         MF0083          3
##  7 MH0443 MHD     ND CTRL GMM         MF0110          0
##  8 MH0443 MHD     ND CTRL GMM         MF0063          4
##  9 MH0443 MHD     ND CTRL GMM         MF0042          1
## 10 MH0443 MHD     ND CTRL GMM         MF0024        185
## # ... with 413,158 more rows
```

In order to actually perform analysis, we transform the data into the data matrix X, with every row being a sample and every column a particular feature. First we combine redundant columns.

```
united_functional_space = functional_space %>%
  unite(SampleInfo, Sample, Dataset, Status) %>%
  unite(FeatureInfo, FeatureType, Feature)
```

```
X = united_functional_space %>% group_by(FeatureInfo) %>% spread(FeatureInfo, Abundance)
```

There is significant missingness in 25 of the individuals, and we exclude them from further analysis.

```
X = drop_na(X)
X
```

```
## # A tibble: 759 x 528
##    SampleInfo GMM_MF0001 GMM_MF0002 GMM_MF0003 GMM_MF0004 GMM_MF0005
##    <chr>           <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
##  1 BGI-06A_C~        466          0          3          3          3
##  2 BGI-15A_C~        195          0          0          0          0
##  3 BGI-17A_C~        304          0          0          0          0
##  4 BGI-27A_C~        148          0          0          0          0
##  5 BGI-28A_C~        329          0          5          5          2
##  6 BGI-33A_C~        300          0          7          7          1
##  7 BGI-34A_C~        202          0          9          8          3
##  8 BGI001A_C~         35          0          2          1          1
##  9 BGI002A_C~        191          0          1          1          1
## 10 BGI003A_C~        246          0          4          4          3
## # ... with 749 more rows, and 522 more variables: GMM_MF0006 <dbl>,
## #   GMM_MF0007 <dbl>, GMM_MF0008 <dbl>, GMM_MF0009 <dbl>,
## #   GMM_MF0010 <dbl>, GMM_MF0011 <dbl>, GMM_MF0012 <dbl>,
## #   GMM_MF0014 <dbl>, GMM_MF0015 <dbl>, GMM_MF0016 <dbl>,
## #   GMM_MF0017 <dbl>, GMM_MF0018 <dbl>, GMM_MF0019 <dbl>,
## #   GMM_MF0023 <dbl>, GMM_MF0024 <dbl>, GMM_MF0026 <dbl>,
## #   GMM_MF0027 <dbl>, GMM_MF0028 <dbl>, GMM_MF0029 <dbl>,
## #   GMM_MF0030 <dbl>, GMM_MF0031 <dbl>, GMM_MF0032 <dbl>,
## #   GMM_MF0033 <dbl>, GMM_MF0034 <dbl>, GMM_MF0035 <dbl>,
## #   GMM_MF0036 <dbl>, GMM_MF0037 <dbl>, GMM_MF0039 <dbl>,
## #   GMM_MF0040 <dbl>, GMM_MF0041 <dbl>, GMM_MF0042 <dbl>,
## #   GMM_MF0043 <dbl>, GMM_MF0044 <dbl>, GMM_MF0045 <dbl>,
## #   GMM_MF0046 <dbl>, GMM_MF0047 <dbl>, GMM_MF0048 <dbl>,
## #   GMM_MF0049 <dbl>, GMM_MF0050 <dbl>, GMM_MF0051 <dbl>,
## #   GMM_MF0052 <dbl>, GMM_MF0053 <dbl>, GMM_MF0054 <dbl>,
## #   GMM_MF0055 <dbl>, GMM_MF0056 <dbl>, GMM_MF0058 <dbl>,
## #   GMM_MF0059 <dbl>, GMM_MF0060 <dbl>, GMM_MF0061 <dbl>,
## #   GMM_MF0062 <dbl>, GMM_MF0063 <dbl>, GMM_MF0066 <dbl>,
## #   GMM_MF0070 <dbl>, GMM_MF0071 <dbl>, GMM_MF0073 <dbl>,
## #   GMM_MF0074 <dbl>, GMM_MF0075 <dbl>, GMM_MF0076 <dbl>,
## #   GMM_MF0077 <dbl>, GMM_MF0079 <dbl>, GMM_MF0080 <dbl>,
## #   GMM_MF0081 <dbl>, GMM_MF0082 <dbl>, GMM_MF0083 <dbl>,
## #   GMM_MF0084 <dbl>, GMM_MF0085 <dbl>, GMM_MF0086 <dbl>,
## #   GMM_MF0089 <dbl>, GMM_MF0090 <dbl>, GMM_MF0091 <dbl>,
## #   GMM_MF0092 <dbl>, GMM_MF0094 <dbl>, GMM_MF0095 <dbl>,
## #   GMM_MF0097 <dbl>, GMM_MF0099 <dbl>, GMM_MF0100 <dbl>,
## #   GMM_MF0101 <dbl>, GMM_MF0102 <dbl>, GMM_MF0104 <dbl>,
## #   GMM_MF0106 <dbl>, GMM_MF0107 <dbl>, GMM_MF0108 <dbl>,
## #   GMM_MF0109 <dbl>, GMM_MF0110 <dbl>, GMM_MF0112 <dbl>,
## #   GMM_MF0113 <dbl>, GMM_MF0114 <dbl>, GMM_MF0116 <dbl>,
## #   GMM_MF0117 <dbl>, GMM_MF0118 <dbl>, GMM_MF0119 <dbl>,
## #   GMM_MF0120 <dbl>, GMM_MF0121 <dbl>, GMM_MF0123 <dbl>,
## #   GMM_MF0124 <dbl>, GMM_MF0125 <dbl>, GMM_MF0126 <dbl>,
## #   GMM_MF0127 <dbl>, GMM_MF0128 <dbl>, GMM_MF0129 <dbl>, ...
```

Analysis

PCA

```r
library(nFactors)
```

```
## Loading required package: MASS
```

```
## 
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
## 
##     select
```

```
## Loading required package: psych
```

```
## 
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
## 
##     %+%, alpha
```

```
## Loading required package: boot
```

```
## 
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:psych':
## 
##     logit
```

```
## Loading required package: lattice
```

```
## 
## Attaching package: 'lattice'
```

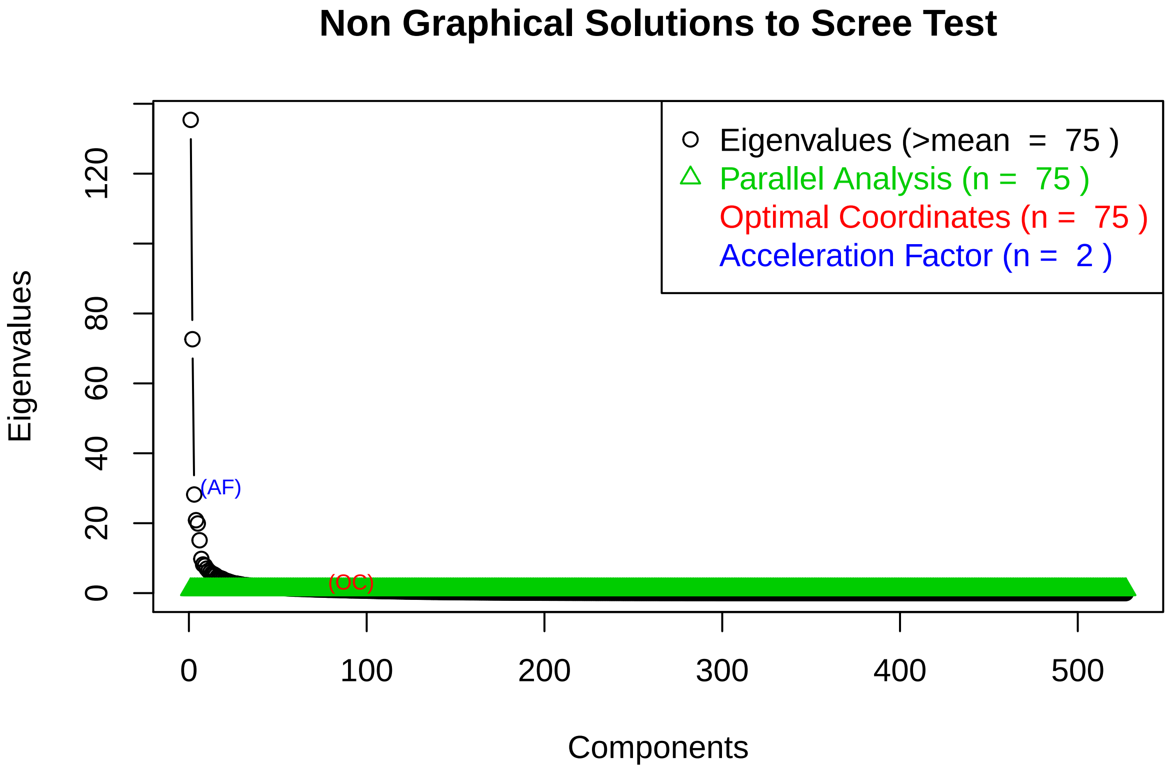
```
## The following object is masked from 'package:boot':
## 
##     melanoma
```

```
##
```

```
## Attaching package: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

```
eigenvals = eigen(cor(X[-1]))
eigenvaldist = parallel(subject=nrow(X[-1]), var=ncol(X[-1]), rep=100, cent=0.05)
nS = nScree(x=eigenvals$values, aparallel=eigenvals$eigen$qevpea)
plotnScree(nS)
```



Bibliography

Forslund, Kristoffer, Falk Hildebrand, Trine Nielsen, Gwen Falony, Le ChatelierEmmanuelle, Shinichi Sunagawa, Edi Prifti, et al. n.d. "Disentangling Type 2 Diabetes and Metformin Treatment Signatures in the Human Gut Microbiota." *Nature* 528: 262 EP. https://doi.org/10.1038/nature15766.