Microbial Data Analysis

Fangting Zhou

We coexist with our microbiota as mutualists. High-throughout sequencing technology such as 16S rRNA sequencing has been widely used to quantify the microbial composition in order to explore its relationship with human health. A key question in such microbiome studies is to identify latent features like diseases that are associated with certain microbes.

Description of the Data Set

In the current work, the count data set **seedLev2Counts** consisting of 162 rows and 12 columns. Columns represent infants from two groups, with BF for breast feeding and FF for formula feeding respectively. Rows represent microbial operational taxonomic units (OTUs, pragmatic proxies for microbial species at different taxonomic levels) after aligning to the second level SEED subsystem. The raw data is highly right skewed and sparse with many zeros.

```
library(reshape2)
library(ggplot2)

load('~/Desktop/Advanced Applied Statistics/project/Microbial-Data.RData')

nrow(seedLev2Counts)
```

```
## [1] 162
```

```
ncol(seedLev2Counts)
```

```
## [1] 12
```

```
head(row.names(seedLev2Counts))
```

```
## [1] "Amino Acids and Derivatives_Alanine, serine, and glycine"
## [2] "Amino Acids and Derivatives_Amino Acids and Derivatives"
## [3] "Amino Acids and Derivatives_Arginine; urea cycle, polyamines"
## [4] "Amino Acids and Derivatives_Aromatic amino acids and derivatives"
## [5] "Amino Acids and Derivatives_Branched-chain amino acids"
## [6] "Amino Acids and Derivatives_Glutamine, glutamate, aspartate, asparagine; ammonia assimilation"
```

```
tail(row.names(seedLev2Counts))
```

```
## [1] "Virulence_Regulation of virulence"
## [2] "Virulence_Resistance to antibiotics and toxic compounds"
## [3] "Virulence_Toxins and superantigens"
## [4] "Virulence_Type III, Type IV, ESAT secretion systems"
## [5] "Virulence_Type VI secretion systems"
## [6] "Virulence_Virulence"
```

```
colnames(seedLev2Counts)
```

```
##  [1] "BMS8"  "BMS10" "BMS16" "BF3"   "BF4"   "BF6"   "FF2"   "FF3"
##  [9] "FF7"   "FF13"  "FF15"  "FF5"
```

```
str(seedLev2Counts)
```

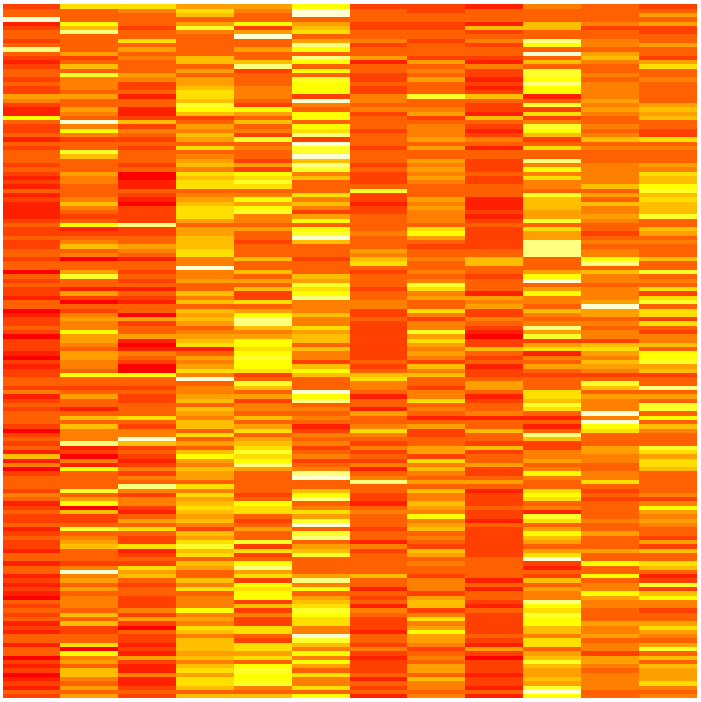
```
## 'data.frame':    162 obs. of  12 variables:
##  $ BMS8 : int  113 0 188 287 189 31 130 637 11 75 ...
```

```
##  $ BMS10: int  250 0 429 359 319 98 350 913 55 156 ...
##  $ BMS16: int  120 0 217 228 221 75 112 563 16 164 ...
##  $ BF3  : int  286 0 468 581 472 90 361 1345 51 180 ...
##  $ BF4  : int  296 0 387 708 480 122 478 1596 36 205 ...
##  $ BF6  : int  387 0 541 441 342 115 170 1092 105 151 ...
##  $ FF2  : int  101 0 234 339 225 66 149 700 18 89 ...
##  $ FF3  : int  228 0 347 409 404 72 271 1112 31 153 ...
##  $ FF7  : int  97 0 197 270 247 41 133 721 6 81 ...
##  $ FF13 : int  379 2 677 469 396 88 271 1284 129 179 ...
##  $ FF15 : int  193 0 389 428 332 80 198 1042 56 175 ...
##  $ FF5  : int  233 0 367 580 380 92 292 1323 36 185 ...
```

```
sum(seedLev2Counts == 0) / ncol(seedLev2Counts) / nrow(seedLev2Counts)
```

```
## [1] 0.1342593
```

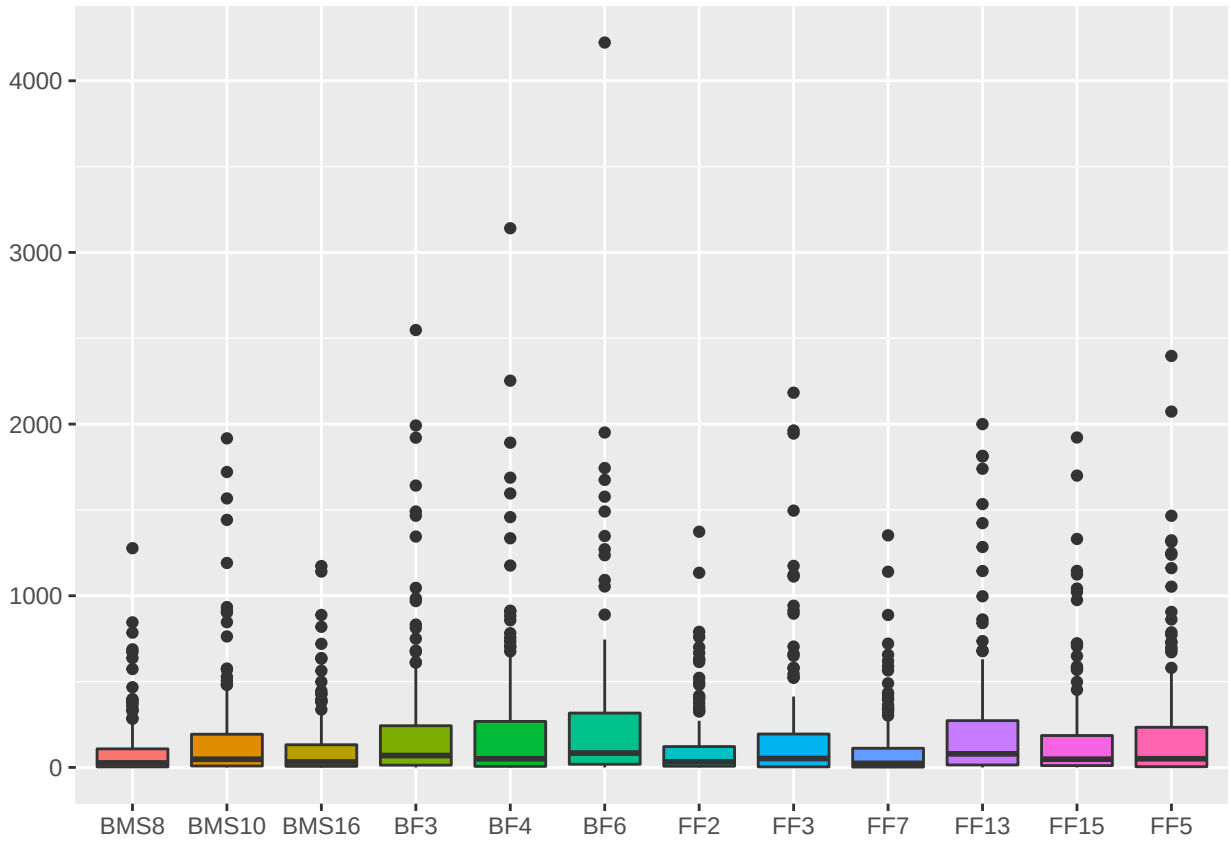
```
heatmap(as.matrix(seedLev2Counts), Rowv = NA, Colv = NA, labRow = NA, labCol = NA)
```



```
plotData = melt(seedLev2Counts)
```

```
## No id variables; using all as measure variables
```

```
ggplot(plotData, aes(x = variable, y = value)) + geom_boxplot(aes(fill = variable)) +
  theme(legend.position = "none") + labs(x = NULL, y = NULL)
```



Goal of the Project

For the current data set, the problem is specific to find features that are closely associated with certain infant microbiomes. Features are defined as some biological functions that are related to a set of microbiomes, for instance, immune function, defense function, digestive function and so on. Relevant gene expression data in **pairedExprSumm** will be used to validate the association analysis. Some work has been done in the report of He (2018) using sparse canonical correlation analysis (CCA). Meanwhile, features are associated with infant samples, which I expect to behave similar within group and different between groups. The project will help improve infant diets and verify whether breast feeding benefits the infant development.

```
nrow(pairedExprSumm)
```

```
## [1] 15856
```

```
ncol(pairedExprSumm)
```

```
## [1] 12
```

```
head(row.names(pairedExprSumm))
```

```
## [1] "15E1.2" "2'-PDE" "3.8-1"  "76P"    "7A5"    "A1BG"
```

```
tail(row.names(pairedExprSumm))
```

```
## [1] "ZYG11A"  "ZYG11B"  "ZYG11BL" "ZYX"     "ZZEF1"   "ZZZ3"
```

```
colnames(pairedExprSumm)
```

```
##  [1] "BMS8"  "BMS10" "BMS16" "BF3"   "BF4"   "BF6"   "FF2"   "FF3"
##  [9] "FF7"   "FF13"  "FF15"  "FF5"
```

Methods

To this end, I generalize the bayesian double feature allocation (DFA) proposed in Ni (2018) originally used on categorical data to the current data set. The generalized DFA is a probability model on the count data matrix based on bayesian hierarchical methods, which allow us to infer the latent structure. To be specific, the sampling model of each infant $x_i = (x_{i1}, \ldots, x_{ip})$ follows the multinomial distribution with the case-specific parameter θ_{ij} depending on the latent binary indicator z_{ij}. The latent variable z_{ij} is used to indicate rich ($z_{ij} = 1$) or rare ($z_{ij} = 0$) of the i-th mircobiome in the j-th infant sample. For modeling details of the indicator matrix Z using categorical matrix factorization, please refer to Ni (2018).