Code for Analyses

Contents

1. Preparations

Load data

```
IBD.original <- read.csv2('./HL_DHL_data.csv')
```

Select columns of relevance and define categorical vectors for summary

```
library(dplyr)
IBD <- IBD.original[, c(2:3, 5:7, 9:15, 20:22, 25,
                        31, 35:42, 54, 63, 67:82, 119)]
# Recode nominal variables to binary variables
IBD$native_language <- ifelse(IBD$native_language %in% c(0, 7, 14, 19), 0, 1)
IBD$ASA_5 <- ifelse(IBD$ASA_5 %in% c(1:3), 1, 0)
# Reduce categories according to Montreal classification (L4 and B4 for CD)
IBD <- IBD %>%
  mutate(CD_localisation = case_when(CD_localisation %in% 4:7 ~ 4,
                                     TRUE ~ CD_localisation))
IBD <- IBD %>%
  mutate(CD_behaviour = case_when(CD_behaviour %in% 4:6 ~ 4,
                                  TRUE ~ CD_behaviour))
# Define categorical variables
categorical <- c("gender", "education", "marital_status", "work_status",
```

```
                "native_language", "diagnose", "surgery", "UC_localisation",
                "CD_localisation", "CD_behaviour", "IBD_disease_activity",
                "treatment", "ASA_5", "immunosuppressive", "biological",
                "corticosteroids", "EQ5D_mobility", "EQ5D_self_care",
                "EQ5D_usual_activities", "EQ5D_pain_discomfort",
                "EQ5D_anxiety_depression")
for (variable_name in categorical) {
  IBD[, variable_name] <- as.factor(IBD[, variable_name])
}
# View summary for the total data set and for each of the diagnoses
CD <- subset(IBD, diagnose == 1)
UC <- subset(IBD, diagnose == 2)
summary(IBD)
```

```
##  gender       age       education  marital_status work_status native_language
##  1:178   Min.   :18.0   0   : 11   0:119          0: 86       0:339
##  2:202   1st Qu.:31.0   1   : 96   1:261          1:294       1: 41
##          Median :42.0   2   :130
##          Mean   :43.6   3   :142
##          3rd Qu.:55.0   NA's:  1
##          Max.   :87.0
##
##  diagnose disease_duration surgery UC_localisation CD_localisation CD_behaviour
##  1:207    Min.   : 0.00    0:256   0:208           0:173           0:183
##  2:173    1st Qu.: 6.00    1:124   1: 31           1: 53           1: 89
##           Median :12.00            2: 41           2: 32           2: 37
##           Mean   :14.83            3:100           3:101           3: 14
##           3rd Qu.:23.00                            4: 21           4: 57
##           Max.   :63.00
##
##  IBD_disease_activity treatment ASA_5   immunosuppressive biological
##  0   :245             0: 19     0:306   0:364             0: 82
##  1   :129             1:361     1: 74   1: 16             1:298
##  NA's:  6
##
##
##
##
##  corticosteroids  Calprotectin     EQ5D_mobility EQ5D_self_care
##  0:347           Min.   :   4.00   1:317         1:358
##  1: 33           1st Qu.:  19.75   2: 49         2: 13
##                  Median :  73.50   3: 10         3:  8
##                  Mean   : 333.09   4:  4         4:  1
##                  3rd Qu.: 320.00
##                  Max.   :6000.00
##                  NA's   :100
##  EQ5D_usual_activities EQ5D_pain_discomfort EQ5D_anxiety_depression
##  1:248                 1:140                1   :209
##  2: 98                 2:168                2   :108
##  3: 24                 3: 53                3   : 48
##  4:  8                 4: 18                4   : 12
##  5:  2                 5:  1                5   :  2
##                                             NA's:  1
##
```

```
##     EQ5Dvas1        GSE_sum         BIPQ_sum       HLQ.Scale1     HLQ.Scale2
##  Min.   :  6.0   Min.   :10.00   Min.   : 4.00   Min.   :1.00   Min.   :1.250
##  1st Qu.: 60.0   1st Qu.:28.00   1st Qu.:26.00   1st Qu.:2.50   1st Qu.:2.500
##  Median : 75.0   Median :32.00   Median :34.00   Median :3.00   Median :3.000
##  Mean   : 71.4   Mean   :31.44   Mean   :34.97   Mean   :2.96   Mean   :2.859
##  3rd Qu.: 89.0   3rd Qu.:36.00   3rd Qu.:43.25   3rd Qu.:3.25   3rd Qu.:3.250
##  Max.   :100.0   Max.   :40.00   Max.   :72.00   Max.   :4.00   Max.   :4.000
##  NA's   :11
##    HLQ.Scale3      HLQ.Scale4      HLQ.Scale5      HLQ.Scale6
##  Min.   :1.400   Min.   :1.000   Min.   :1.000   Min.   :1.600
##  1st Qu.:2.600   1st Qu.:2.600   1st Qu.:2.400   1st Qu.:3.400
##  Median :3.000   Median :2.800   Median :2.800   Median :4.000
##  Mean   :2.906   Mean   :2.884   Mean   :2.736   Mean   :3.825
##  3rd Qu.:3.200   3rd Qu.:3.200   3rd Qu.:3.000   3rd Qu.:4.200
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :5.000
##
##    HLQ.Scale7      HLQ.Scale8      HLQ.Scale9     eHLQ.Domain1
##  Min.   :1.500   Min.   :1.600   Min.   :2.400   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:3.200   1st Qu.:3.600   1st Qu.:2.600
##  Median :3.500   Median :3.800   Median :4.000   Median :3.000
##  Mean   :3.482   Mean   :3.691   Mean   :3.985   Mean   :2.957
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.200   3rd Qu.:3.250
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.000
##
##   eHLQ.Domain2    eHLQ.Domain3  eHLQ.Domain4    eHLQ.Domain5    eHLQ.Domain6
##  Min.   :2.000   Min.   :1.6   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.800   1st Qu.:3.0   1st Qu.:2.800   1st Qu.:2.600   1st Qu.:2.333
##  Median :3.000   Median :3.2   Median :3.000   Median :3.000   Median :2.667
##  Mean   :3.086   Mean   :3.3   Mean   :3.119   Mean   :2.927   Mean   :2.650
##  3rd Qu.:3.400   3rd Qu.:3.8   3rd Qu.:3.600   3rd Qu.:3.200   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.0   Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##   eHLQ.Domain7     OMAS37_sum
##  Min.   :1.000   Min.   : 0.00
##  1st Qu.:2.500   1st Qu.: 0.00
##  Median :3.000   Median : 2.00
##  Mean   :2.865   Mean   : 3.86
##  3rd Qu.:3.250   3rd Qu.: 5.00
##  Max.   :4.000   Max.   :38.00
##                  NA's   :52
```

```
summary(CD)
```

```
##  gender       age        education marital_status work_status native_language
##  1: 99   Min.   :18.00   0   : 8   0: 76          0: 51       0:178
##  2:108   1st Qu.:29.50   1   :55   1:131          1:156       1: 29
##          Median :43.00   2   :73
##          Mean   :43.71   3   :70
##          3rd Qu.:55.00   NA's: 1
##          Max.   :78.00
##
##  diagnose disease_duration surgery UC_localisation CD_localisation CD_behaviour
##  1:207    Min.   : 0.00    0:101   0:207           0:  0           0:10
##  2:  0    1st Qu.: 7.00    1:106   1:  0           1: 53           1:89
##           Median :14.00            2:  0           2: 32           2:37
```

```
##           Mean   :16.69            3:  0           3:101           3:14
##           3rd Qu.:25.50                            4: 21           4:57
##           Max.   :63.00
##
##  IBD_disease_activity treatment ASA_5   immunosuppressive biological
##  0   :105             0: 11     0:202   0:197             0: 21
##  1   : 99             1:196     1:  5   1: 10             1:186
##  NA's:  3
##
##
##
##
##  corticosteroids  Calprotectin     EQ5D_mobility EQ5D_self_care
##  0:198            Min.   :   4.00   1:174         1:189
##  1:  9            1st Qu.:  20.25   2: 28         2: 11
##                   Median :  70.00   3:  2         3:  6
##                   Mean   : 193.49   4:  3         4:  1
##                   3rd Qu.: 233.00
##                   Max.   :1974.00
##                   NA's   :53
##  EQ5D_usual_activities EQ5D_pain_discomfort EQ5D_anxiety_depression
##  1:138                 1:67                1   :111
##  2: 51                 2:99                2   : 60
##  3: 12                 3:31                3   : 27
##  4:  5                 4:10                4   :  6
##  5:  1                 5: 0                5   :  2
##                                            NA's:  1
##
##     EQ5Dvas1         GSE_sum         BIPQ_sum       HLQ.Scale1
##  Min.   :  6.00   Min.   :14.00   Min.   : 6.00   Min.   :1.000
##  1st Qu.: 60.00   1st Qu.:27.00   1st Qu.:27.00   1st Qu.:2.667
##  Median : 77.00   Median :31.00   Median :35.00   Median :3.000
##  Mean   : 71.84   Mean   :30.92   Mean   :35.41   Mean   :2.988
##  3rd Qu.: 90.00   3rd Qu.:36.00   3rd Qu.:43.00   3rd Qu.:3.375
##  Max.   :100.00   Max.   :40.00   Max.   :72.00   Max.   :4.000
##  NA's   :8
##    HLQ.Scale2      HLQ.Scale3      HLQ.Scale4      HLQ.Scale5
##  Min.   :1.250   Min.   :1.400   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.500   1st Qu.:2.600   1st Qu.:2.500   1st Qu.:2.400
##  Median :3.000   Median :3.000   Median :3.000   Median :2.800
##  Mean   :2.887   Mean   :2.898   Mean   :2.888   Mean   :2.735
##  3rd Qu.:3.250   3rd Qu.:3.200   3rd Qu.:3.200   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##    HLQ.Scale6      HLQ.Scale7      HLQ.Scale8      HLQ.Scale9
##  Min.   :1.600   Min.   :1.500   Min.   :1.600   Min.   :2.600
##  1st Qu.:3.400   1st Qu.:3.083   1st Qu.:3.200   1st Qu.:3.600
##  Median :4.000   Median :3.500   Median :3.800   Median :4.000
##  Mean   :3.824   Mean   :3.499   Mean   :3.694   Mean   :3.964
##  3rd Qu.:4.200   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.200
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##   eHLQ.Domain1    eHLQ.Domain2    eHLQ.Domain3    eHLQ.Domain4
##  Min.   :1.000   Min.   :2.000   Min.   :1.600   Min.   :1.000
```

```
##  1st Qu.:2.600   1st Qu.:2.800   1st Qu.:3.000   1st Qu.:2.800
##  Median :3.000   Median :3.000   Median :3.200   Median :3.000
##  Mean   :2.939   Mean   :3.102   Mean   :3.279   Mean   :3.093
##  3rd Qu.:3.200   3rd Qu.:3.400   3rd Qu.:3.800   3rd Qu.:3.600
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##   eHLQ.Domain5    eHLQ.Domain6    eHLQ.Domain7     OMAS37_sum
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:2.600   1st Qu.:2.250   1st Qu.:2.500   1st Qu.: 0.000
##  Median :3.000   Median :2.667   Median :3.000   Median : 2.000
##  Mean   :2.905   Mean   :2.666   Mean   :2.874   Mean   : 3.661
##  3rd Qu.:3.200   3rd Qu.:3.000   3rd Qu.:3.250   3rd Qu.: 5.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :33.000
##                                                  NA's   :33
```

```
summary(UC)
```

```
##  gender      age        education marital_status work_status native_language
##  1:79   Min.   :18.00   0: 3      0: 43          0: 35       0:161
##  2:94   1st Qu.:32.00   1:41      1:130          1:138       1: 12
##         Median :41.00   2:57
##         Mean   :43.46   3:72
##         3rd Qu.:53.00
##         Max.   :87.00
##
##  diagnose disease_duration surgery UC_localisation CD_localisation CD_behaviour
##  1:  0    Min.   : 0.00    0:155   0:  1           0:173           0:173
##  2:173    1st Qu.: 5.00    1: 18   1: 31           1:  0           1:  0
##           Median :10.00            2: 41           2:  0           2:  0
##           Mean   :12.61            3:100           3:  0           3:  0
##           3rd Qu.:17.00                            4:  0           4:  0
##           Max.   :53.00
##
##  IBD_disease_activity treatment ASA_5   immunosuppressive biological
##  0   :140             0:  8     0:104   0:167             0: 61
##  1   : 30             1:165     1: 69   1:  6             1:112
##  NA's:  3
##
##
##
##
##  corticosteroids  Calprotectin     EQ5D_mobility EQ5D_self_care
##  0:149           Min.   :   4.00   1:143         1:169
##  1: 24           1st Qu.:  19.25   2: 21         2:  2
##                  Median :  78.50   3:  8         3:  2
##                  Mean   : 503.71   4:  1         4:  0
##                  3rd Qu.: 424.50
##                  Max.   :6000.00
##                  NA's   :47
##  EQ5D_usual_activities EQ5D_pain_discomfort EQ5D_anxiety_depression
##  1:110                 1:73                1:98
##  2: 47                 2:69                2:48
##  3: 12                 3:22                3:21
##  4:  3                 4: 8                4: 6
##  5:  1                 5: 1                5: 0
```

```
##
##
##     EQ5Dvas1         GSE_sum         BIPQ_sum       HLQ.Scale1
##  Min.   : 15.00   Min.   :10.00   Min.   : 4.00   Min.   :1.000
##  1st Qu.: 60.00   1st Qu.:29.00   1st Qu.:25.00   1st Qu.:2.500
##  Median : 75.00   Median :33.00   Median :33.00   Median :3.000
##  Mean   : 70.88   Mean   :32.05   Mean   :34.44   Mean   :2.927
##  3rd Qu.: 85.00   3rd Qu.:37.00   3rd Qu.:44.00   3rd Qu.:3.250
##  Max.   :100.00   Max.   :40.00   Max.   :71.00   Max.   :4.000
##  NA's   :3
##    HLQ.Scale2      HLQ.Scale3      HLQ.Scale4     HLQ.Scale5      HLQ.Scale6
##  Min.   :1.250   Min.   :1.600   Min.   :1.20   Min.   :1.400   Min.   :2.200
##  1st Qu.:2.500   1st Qu.:2.600   1st Qu.:2.60   1st Qu.:2.400   1st Qu.:3.400
##  Median :3.000   Median :3.000   Median :2.80   Median :2.800   Median :4.000
##  Mean   :2.825   Mean   :2.916   Mean   :2.88   Mean   :2.738   Mean   :3.827
##  3rd Qu.:3.000   3rd Qu.:3.200   3rd Qu.:3.20   3rd Qu.:3.000   3rd Qu.:4.200
##  Max.   :4.000   Max.   :4.000   Max.   :4.00   Max.   :4.000   Max.   :5.000
##
##    HLQ.Scale7      HLQ.Scale8      HLQ.Scale9    eHLQ.Domain1    eHLQ.Domain2
##  Min.   :1.500   Min.   :2.000   Min.   :2.40   Min.   :1.400   Min.   :2.000
##  1st Qu.:3.000   1st Qu.:3.200   1st Qu.:3.80   1st Qu.:2.600   1st Qu.:2.800
##  Median :3.500   Median :3.800   Median :4.00   Median :3.000   Median :3.000
##  Mean   :3.461   Mean   :3.687   Mean   :4.01   Mean   :2.978   Mean   :3.067
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.20   3rd Qu.:3.400   3rd Qu.:3.400
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :4.000   Max.   :4.000
##
##   eHLQ.Domain3    eHLQ.Domain4   eHLQ.Domain5    eHLQ.Domain6    eHLQ.Domain7
##  Min.   :1.800   Min.   :1.00   Min.   :1.400   Min.   :1.000   Min.   :1.250
##  1st Qu.:3.000   1st Qu.:2.80   1st Qu.:2.600   1st Qu.:2.333   1st Qu.:2.500
##  Median :3.400   Median :3.20   Median :3.000   Median :2.667   Median :3.000
##  Mean   :3.326   Mean   :3.15   Mean   :2.954   Mean   :2.631   Mean   :2.853
##  3rd Qu.:3.800   3rd Qu.:3.60   3rd Qu.:3.400   3rd Qu.:3.000   3rd Qu.:3.250
##  Max.   :4.000   Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##    OMAS37_sum
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 2.000
##  Mean   : 4.084
##  3rd Qu.: 5.000
##  Max.   :38.000
##  NA's   :19
```

KNN imputation of missing values

(leaving out OMAS-37 due to MNAR)

```
library(impute)
IBD <- as.matrix(IBD)
IBD.imputed <- impute.knn(IBD[, -44])
IBD.imp <- IBD.imputed$data
IBD.imp <- as.data.frame(IBD.imp)
IBD.imp[, 1:27] <- round(IBD.imp[, 1:27], digits = 0)
```

```r
for (variable_name in categorical) {
  IBD.imp[, variable_name] <- as.factor(IBD.imp[, variable_name])
}
summary(IBD.imp)
```

```
##  gender       age       education marital_status work_status native_language
##  1:178   Min.   :18.0   0: 11     0:119          0: 86       0:339
##  2:202   1st Qu.:31.0   1: 96     1:261          1:294       1: 41
##          Median :42.0   2:131
##          Mean   :43.6   3:142
##          3rd Qu.:55.0
##          Max.   :87.0
##  diagnose disease_duration surgery UC_localisation CD_localisation CD_behaviour
##  1:207    Min.   : 0.00    0:256   0:208           0:173           0:183
##  2:173    1st Qu.: 6.00    1:124   1: 31           1: 53           1: 89
##           Median :12.00            2: 41           2: 32           2: 37
##           Mean   :14.83            3:100           3:101           3: 14
##           3rd Qu.:23.00                            4: 21           4: 57
##           Max.   :63.00
##  IBD_disease_activity treatment ASA_5   immunosuppressive biological
##  0:249                0: 19     0:306   0:364             0: 82
##  1:131                1:361     1: 74   1: 16             1:298
##
##
##
##
##  corticosteroids  Calprotectin     EQ5D_mobility EQ5D_self_care
##  0:347           Min.   :   4.00   1:317         1:358
##  1: 33           1st Qu.:  30.75   2: 49         2: 13
##                  Median : 105.00   3: 10         3:  8
##                  Mean   : 325.97   4:  4         4:  1
##                  3rd Qu.: 343.00
##                  Max.   :6000.00
##  EQ5D_usual_activities EQ5D_pain_discomfort EQ5D_anxiety_depression
##  1:248                 1:140                1:209
##  2: 98                 2:168                2:109
##  3: 24                 3: 53                3: 48
##  4:  8                 4: 18                4: 12
##  5:  2                 5:  1                5:  2
##
##     EQ5Dvas1         GSE_sum         BIPQ_sum       HLQ.Scale1
##  Min.   :  6.00   Min.   :10.00   Min.   : 4.00   Min.   :1.00
##  1st Qu.: 60.00   1st Qu.:28.00   1st Qu.:26.00   1st Qu.:2.50
##  Median : 75.00   Median :32.00   Median :34.00   Median :3.00
##  Mean   : 71.38   Mean   :31.44   Mean   :34.97   Mean   :2.96
##  3rd Qu.: 87.25   3rd Qu.:36.00   3rd Qu.:43.25   3rd Qu.:3.25
##  Max.   :100.00   Max.   :40.00   Max.   :72.00   Max.   :4.00
##    HLQ.Scale2      HLQ.Scale3      HLQ.Scale4      HLQ.Scale5
##  Min.   :1.250   Min.   :1.400   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.500   1st Qu.:2.600   1st Qu.:2.600   1st Qu.:2.400
##  Median :3.000   Median :3.000   Median :2.800   Median :2.800
##  Mean   :2.859   Mean   :2.906   Mean   :2.884   Mean   :2.736
##  3rd Qu.:3.250   3rd Qu.:3.200   3rd Qu.:3.200   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
```

```
##    HLQ.Scale6      HLQ.Scale7      HLQ.Scale8      HLQ.Scale9
##  Min.   :1.600   Min.   :1.500   Min.   :1.600   Min.   :2.400
##  1st Qu.:3.400   1st Qu.:3.000   1st Qu.:3.200   1st Qu.:3.600
##  Median :4.000   Median :3.500   Median :3.800   Median :4.000
##  Mean   :3.825   Mean   :3.482   Mean   :3.691   Mean   :3.985
##  3rd Qu.:4.200   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.200
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##   eHLQ.Domain1    eHLQ.Domain2    eHLQ.Domain3  eHLQ.Domain4    eHLQ.Domain5
##  Min.   :1.000   Min.   :2.000   Min.   :1.6   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.600   1st Qu.:2.800   1st Qu.:3.0   1st Qu.:2.800   1st Qu.:2.600
##  Median :3.000   Median :3.000   Median :3.2   Median :3.000   Median :3.000
##  Mean   :2.957   Mean   :3.086   Mean   :3.3   Mean   :3.119   Mean   :2.927
##  3rd Qu.:3.250   3rd Qu.:3.400   3rd Qu.:3.8   3rd Qu.:3.600   3rd Qu.:3.200
##  Max.   :4.000   Max.   :4.000   Max.   :4.0   Max.   :4.000   Max.   :4.000
##   eHLQ.Domain6    eHLQ.Domain7
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:2.333   1st Qu.:2.500
##  Median :2.667   Median :3.000
##  Mean   :2.650   Mean   :2.865
##  3rd Qu.:3.000   3rd Qu.:3.250
##  Max.   :4.000   Max.   :4.000
```

Preparations for canonical correlation analysis

Select eligible continuous variables for each dataset (D.HL = health literacy and digital health literacy; CDP = clinical, demographic and PROM characteristics)

```
CDP <- IBD.imp[, c(2, 8, 19, 25:27)]
D.HL <- IBD.imp[, c(28:43)]
```

Control for multicollinearity

```
library(car)
resp.Y <- rnorm(nrow(D.HL))
modelY <- lm(resp.Y ~ ., data = D.HL)
vif(modelY)
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##     1.854643     2.627968     1.304085     1.738087     2.106030     3.495599
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##     4.302503     3.819794     2.449637     3.888483     2.549767     2.382375
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##     1.858773     2.557966     2.943072     3.252997
```

```
resp.X <- rnorm(nrow(CDP))
modelX <- lm(resp.X ~ ., data = CDP)
vif(modelX)
```

```
##             age disease_duration     Calprotectin         EQ5Dvas1
##        1.255106         1.249108         1.130455         1.514587
##         GSE_sum         BIPQ_sum
##        1.261009         1.635204
```

Converting datasets to matrices with correct vector types

```
CDP <- as.matrix(sapply(CDP, as.numeric))
D.HL <- as.matrix(sapply(D.HL, as.numeric))
```

Standardize data to z-scores

```
CDP <- scale(CDP)
D.HL <- scale(D.HL)
```

2. Canonical correlation analysis

```
library(candisc)
cca.out <- candisc::cancor(CDP, D.HL)
# View results
cca.out
```

```
##
## Canonical correlation analysis of:
##   6   X  variables:  age, disease_duration, Calprotectin, EQ5Dvas1, GSE_sum, BIPQ_sum
##  with   16  Y variables: HLQ.Scale1, HLQ.Scale2, HLQ.Scale3, HLQ.Scale4, HLQ.Scale5, HLQ.Scale6, HLQ.S
##
##     CanR  CanRSQ   Eigen percent    cum                            scree
## 1 0.6248 0.39040 0.64041  58.170  58.17 ********************************
## 2 0.4432 0.19641 0.24442  22.201  80.37 ***********
## 3 0.3106 0.09646 0.10676   9.698  90.07 *****
## 4 0.2582 0.06666 0.07143   6.488  96.56 ***
## 5 0.1559 0.02430 0.02491   2.263  98.82 *
## 6 0.1133 0.01284 0.01300   1.181 100.00 *
##
## Test of H0: The canonical correlations in the
## current row and all that follow are zero
##
##      CanR LR test stat approx F numDF  denDF   Pr(> F)
## 1 0.62482      0.39789   3.7445    96 2035.1 < 2.2e-16 ***
## 2 0.44318      0.65271   2.1410    75 1723.8 1.025e-07 ***
## 3 0.31059      0.81224   1.3754    56 1402.5   0.03628 *
## 4 0.25819      0.89896   1.0046    39 1069.8   0.46360
## 5 0.15590      0.96317   0.5713    24  724.0   0.95127
## 6 0.11330      0.98716   0.4291    11  363.0   0.94276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Permutation test over significant canonical correlations

Initialise number of permutations and seed for reproducibility

```
n.perm <- 10000
set.seed(0)
```

Create list for storing permutations and vectors for visualizing permutation distributions

```
perm.cancor <- vector("list", length = n.perm)
perm.cancor1 <- numeric()
```

```
perm.cancor2 <- numeric()
perm.cancor3 <- numeric()
```

Permute data

```
for (i in 1:n.perm) {
  perm.cancor[[i]] <- numeric()
  Y.perm <- cca.out$Y[sample(nrow(cca.out$Y)), ]
  perm.cca <- candisc::cancor(cca.out$X, Y.perm)
  perm.cancor[[i]][1:3] <- perm.cca$cancor[1:3]
  perm.cancor1 <- c(perm.cancor1, perm.cancor[[i]][1])
  perm.cancor2 <- c(perm.cancor2, perm.cancor[[i]][2])
  perm.cancor3 <- c(perm.cancor3, perm.cancor[[i]][3])
}
```

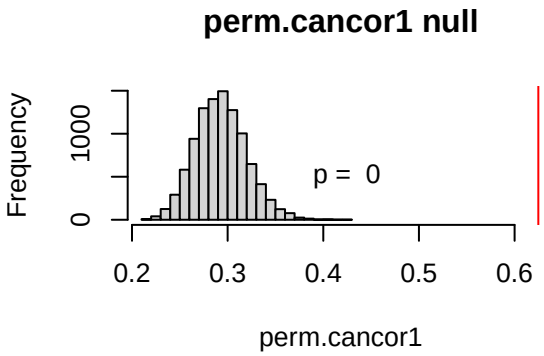
Compute p-value

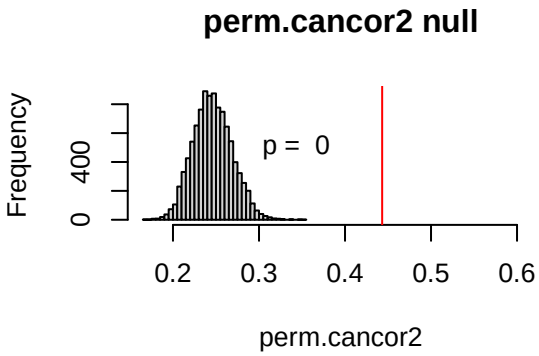
```
obs.3cancor <- cca.out$cancor[1:3]
p.values <- numeric(3)
for (i in 1:3) {
  obs.cancor <- obs.3cancor[i]
  perm.cancors <- unlist(lapply(1:n.perm, function(p) perm.cancor[[p]][i]))
  p.values[i] <- mean(perm.cancors >= obs.cancor)
}
p.values
```

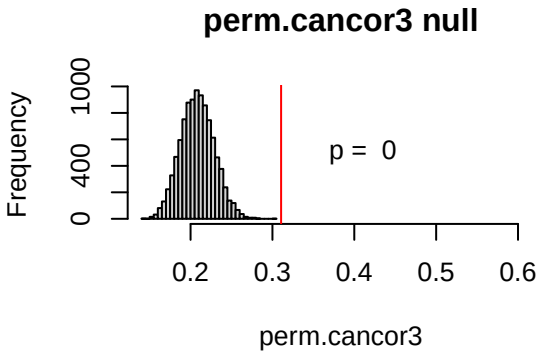
```
## [1] 0 0 0
```

Visualize permutation distribution and empirical canonical correlation value

```
par(mfrow=c(2 ,2))
hist(perm.cancor1, breaks = 30, main = "perm.cancor1 null",
     xlim = range(c(perm.cancor1, obs.3cancor)))
abline(v=obs.3cancor[1], col="red")
text(obs.3cancor[1] - 0.2, 500, paste('p = ', round(p.values[1], 2)))
hist(perm.cancor2, breaks = 30, main = "perm.cancor2 null",
     xlim = range(c(perm.cancor2, obs.3cancor)))
abline(v=obs.3cancor[2], col="red")
text(obs.3cancor[2] - 0.1, 500, paste('p = ', round(p.values[2], 2)))
hist(perm.cancor3, breaks = 30, main = "perm.cancor3 null",
     xlim = range(c(perm.cancor3, obs.3cancor)))
abline(v=obs.3cancor[3], col="red")
text(obs.3cancor[3] + 0.1, 500, paste('p = ', round(p.values[3], 2)))
```







10-fold cross-validation of significant canonical correlations

```
library(caret)
library(candisc)
```

Initalise settings

```
XY <- IBD.imp[, c(2, 8, 19, 25:43)]
XY <- as.matrix(sapply(XY, as.numeric))
XY <- scale(XY)
rep = 100
k = 10
set.seed(1)
```

Create empty vectors for iterations

```
train.canR1 <- matrix(0, rep, k)
train.canR2 <- matrix(0, rep, k)
train.canR3 <- matrix(0, rep, k)
test.canR1 <- matrix(0, rep, k)
test.canR2 <- matrix(0, rep, k)
test.canR3 <- matrix(0, rep, k)
```

repeat 10-fold cross-validation over 100 iterations

```
for (i in 1:rep) {
  folds <- createFolds(XY[, 1], k = k)
  for (j in 1:k) {
    test.XY <- XY[folds[[j]], ]
    train.XY <- XY[-folds[[j]], ]
    Xtrain <- scale(train.XY[, 1:6])
    Ytrain <- scale(train.XY[, 7:22])

    trainCCA <- candisc::cancor(Xtrain, Ytrain)

    train.canR1[i, j] <- trainCCA$cancor[1]
    train.canR2[i, j] <- trainCCA$cancor[2]
    train.canR3[i, j] <- trainCCA$cancor[3]

    Xtest <- scale(test.XY[, 1:6], center = attr(Xtrain, "scaled:center"),
                   scale = attr(Xtrain, "scaled:scale"))
    Ytest <- scale(test.XY[, 7:22], center = attr(Ytrain, "scaled:center"),
                   scale = attr(Ytrain, "scaled:scale"))

    test.canR1.X <- as.matrix(Xtest) %*% trainCCA$coef$X[, 1]
    test.canR1.Y <- as.matrix(Ytest) %*% trainCCA$coef$Y[, 1]
    test.canR1[i, j] <- cor(test.canR1.X, test.canR1.Y)

    test.canR2.X <- as.matrix(Xtest) %*% trainCCA$coef$X[, 2]
    test.canR2.Y <- as.matrix(Ytest) %*% trainCCA$coef$Y[, 2]
    test.canR2[i, j] <- cor(test.canR2.X, test.canR2.Y)

    test.canR3.X <- as.matrix(Xtest) %*% trainCCA$coef$X[, 3]
    test.canR3.Y <- as.matrix(Ytest) %*% trainCCA$coef$Y[, 3]
    test.canR3[i, j] <- cor(test.canR3.X, test.canR3.Y)
  }
}
```

View results

```
mean(train.canR1)
```

```
## [1] 0.6281267
```

```
mean(test.canR1)
```

```
## [1] 0.550082
```

```
mean(train.canR2)
```

```
## [1] 0.4488519
```

```
mean(test.canR2)
```

```
## [1] 0.3405044
```

```
mean(train.canR3)
```

```
## [1] 0.3204393
```

```
mean(test.canR3)
```

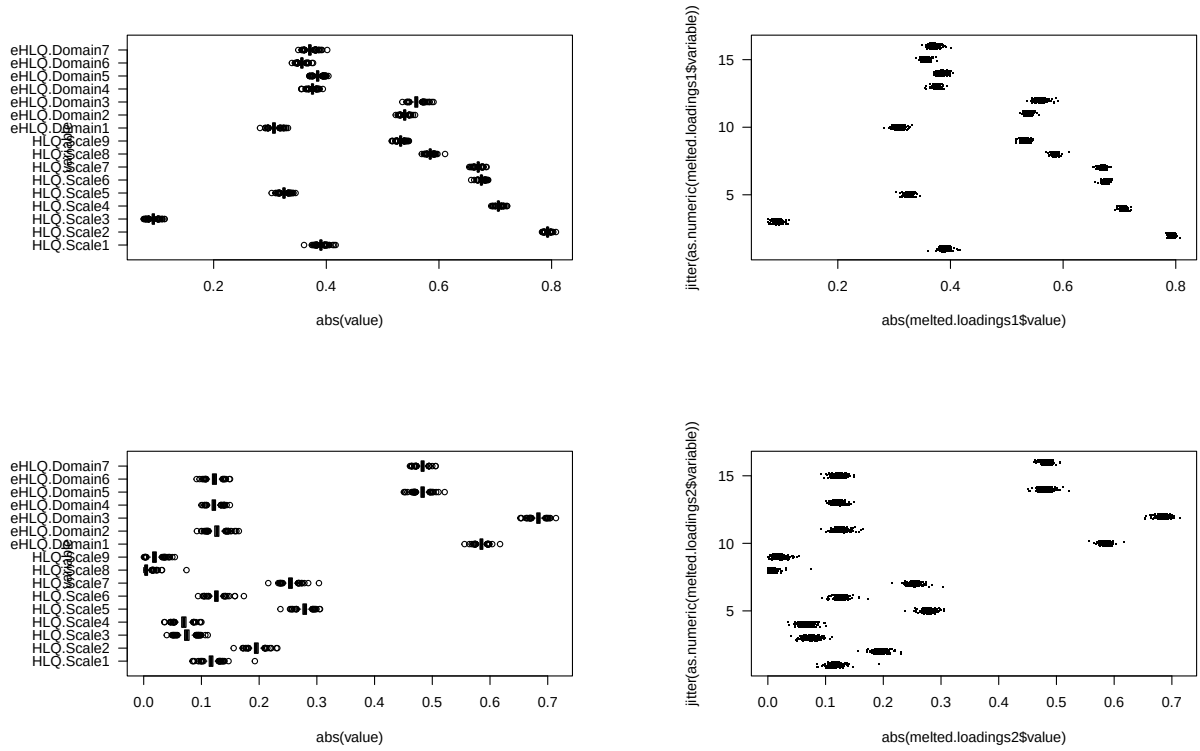
```
## [1] 0.114844
```

Jackknife cross-validation of canonical structure correlations

```
library(candisc)
library(foreach)
library(doParallel)
# Function to compute canonical variates:
predict.cancor <- function(cancor.obj, X, Y){
  pred.X <- as.matrix(X) %*% cancor.obj$coef$X
  pred.Y <- as.matrix(Y) %*% cancor.obj$coef$Y
  pred.XY <- list(pred.X, pred.Y)
  names(pred.XY) <- c("pred.X", "pred.Y")
  return(pred.XY)
}

# Initialize parallel backend
cl <- makeCluster(detectCores() - 1)
registerDoParallel(cl)

# Create function to perform jackknife
njack <- nrow(CDP)
jack.res <- foreach(i=1:njack) %dopar% {
  model <- candisc::cancor(CDP[-i, ], D.HL[-i, ])
  selected.vars <- model$names$X
  # Ensure CDP[i, selected.vars] is properly formatted
  X.test <- CDP[i, selected.vars, drop = FALSE]
  Y.test <- D.HL[i, , drop = FALSE]

  prediction <- predict.cancor(model, X.test, Y.test)
  list(prediction, model)
  }

# Load jackknife
jack.results <- lapply(jack.res, function(x){return(x[[1]])})
jack.X <- lapply(jack.results, function(x){return(x[[1]])})
jack.X <- as.data.frame(do.call(rbind, jack.X))
jack.Y <- lapply(jack.results, function(x){return(x[[2]])})
jack.Y <- as.data.frame(do.call(rbind, jack.Y))

# Retrieve loadings from saved jackknife models
jack.models <- lapply(jack.res, function(x){return(x[[2]])})
jack.loadings1 <- lapply(jack.models, function(model){
  return(model$structure$Y.yscores[,1])
})
jack.loadings2 <- lapply(jack.models, function(model){
  return(model$structure$Y.yscores[,2])
  })
jack.loadings1 <- as.data.frame(do.call(rbind, jack.loadings1))
jack.loadings2 <- as.data.frame(do.call(rbind, jack.loadings2))

# Stop parallel backend
stopCluster(cl)
```

Plot jackknife results for first and second canonical variate

```
library(reshape)
melted.loadings1 <- melt(jack.loadings1)
melted.loadings2 <- melt(jack.loadings2)

par(mfrow=c(2,2), las=1, mai=c(1.02, 1.3, 0.82, 0.42))
boxplot(abs(value) ~ variable, data=melted.loadings1, horizontal=T)
plot(abs(melted.loadings1$value),
     jitter(as.numeric(melted.loadings1$variable)),
     pch='.')
boxplot(abs(value) ~ variable, data=melted.loadings2, horizontal=T)
plot(abs(melted.loadings2$value),
     jitter(as.numeric(melted.loadings2$variable)),
     pch='.')
```



Compute SD for all variables' loadings over all jackknife iterations

```
jack.loadings1.sd <- apply(jack.loadings1, 2, sd)
jack.loadings2.sd <- apply(jack.loadings2, 2, sd)

# View mean and SD
colMeans(jack.loadings1)
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##    0.3823287    0.7761353    0.0909416    0.6907731    0.3181641    0.6613153
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##    0.6556214    0.5722674    0.5206430    0.3007692    0.5277288    0.5483254
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##    0.3675380    0.3766493    0.3490882    0.3632049
```

```
jack.loadings1.sd
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##   0.07956020   0.16206478   0.01959337   0.14411166   0.06668630   0.13796006
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##   0.13670181   0.11961583   0.10905120   0.06353101   0.11051769   0.11505749
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##   0.07678914   0.07916078   0.07301708   0.07628995
```

```
colMeans(jack.loadings2)
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##  0.103339399  0.173155751  0.065776390 -0.060422657 -0.246464652  0.112465718
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##  0.225801805  0.003156244 -0.016546991 -0.516963638 -0.111879476 -0.604277755
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
## -0.106994549 -0.426976398 -0.107289888 -0.426328170
```

```
jack.loadings2.sd
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##  0.055167058  0.090149153  0.037049161  0.034848306  0.130721899  0.057523855
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##  0.117547128  0.007270604  0.011819398  0.274427350  0.060464944  0.320258142
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##  0.058474216  0.225958375  0.058899888  0.226976125
```

Interpretation of canonical correlation analysis

Focusing on first and second canonical correlation due to low performance on third canonical correlation in 10-fold cross-validation

```
cca.out$cancor[1:2]
```

```
## [1] 0.6248177 0.4431836
```

Inspect redundancy

```
library(candisc)
candisc::redundancy(cca.out)
```

```
##
## Redundancies for the X variables & total X canonical redundancy
##
##     Xcan1     Xcan2     Xcan3     Xcan4     Xcan5     Xcan6 total X|Y
##  0.117400  0.043388  0.011942  0.008155  0.002326  0.001753  0.184964
##
## Redundancies for the Y variables & total Y canonical redundancy
##
##     Ycan1     Ycan2     Ycan3     Ycan4     Ycan5     Ycan6 total Y|X
## 0.1019042 0.0189521 0.0053565 0.0017461 0.0017828 0.0007286 0.1304704
```

Inspect linear relationship from each variable in each data set to the canonical correlation

```
cca.out$structure$X.xscores[, 1]
```

```
##              age disease_duration     Calprotectin         EQ5Dvas1
##      -0.12150388       0.03220333      -0.21273111       0.61364169
```

```
##          GSE_sum         BIPQ_sum
##       0.81818286      -0.83503472
```

```
cca.out$structure$X.yscores[, 1]
```

```
##              age disease_duration     Calprotectin         EQ5Dvas1
##      -0.07591777       0.02012121      -0.13291816       0.38341417
##          GSE_sum         BIPQ_sum
##       0.51121512      -0.52174445
```

```
cca.out$structure$Y.yscores[, 1]
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##   0.39053029   0.79293335   0.09291131   0.70568519   0.32506321   0.67558764
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##   0.66976852   0.58466443   0.53196708   0.30734676   0.53919518   0.56024713
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##   0.37547443   0.38484097   0.35663103   0.37109833
```

```
cca.out$structure$Y.xscores[, 1]
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##   0.24401023   0.49543877   0.05805263   0.44092458   0.20310524   0.42211910
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##   0.41848321   0.36530867   0.33238243   0.19203569   0.33689868   0.35005231
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
##   0.23460306   0.24045544   0.22282937   0.23186880
```

```
cca.out$structure$X.xscores[, 2]
```

```
##              age disease_duration     Calprotectin         EQ5Dvas1
##        0.9344416        0.5295724       -0.2043668        0.1446119
##          GSE_sum         BIPQ_sum
##       -0.1527688       -0.2928559
```

```
cca.out$structure$X.yscores[, 2]
```

```
##              age disease_duration     Calprotectin         EQ5Dvas1
##       0.41412922       0.23469783      -0.09057203       0.06408962
##          GSE_sum         BIPQ_sum
##      -0.06770462      -0.12978892
```

```
cca.out$structure$Y.yscores[, 2]
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##  0.116851608  0.195057865  0.075038281 -0.069349323 -0.278935811  0.126092366
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##  0.254465578  0.002856928 -0.018855101 -0.585302149 -0.126923736 -0.683924816
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
## -0.121815474 -0.483088864 -0.122245889 -0.483018185
```

```
cca.out$structure$Y.xscores[, 2]
```

```
##   HLQ.Scale1   HLQ.Scale2   HLQ.Scale3   HLQ.Scale4   HLQ.Scale5   HLQ.Scale6
##  0.051786721  0.086446455  0.033255739 -0.030734486 -0.123619789  0.055882074
##   HLQ.Scale7   HLQ.Scale8   HLQ.Scale9 eHLQ.Domain1 eHLQ.Domain2 eHLQ.Domain3
##  0.112774982  0.001266144 -0.008356272 -0.259396339 -0.056250524 -0.303104292
## eHLQ.Domain4 eHLQ.Domain5 eHLQ.Domain6 eHLQ.Domain7
## -0.053986626 -0.214097083 -0.054177378 -0.214065759
```

3. Hierarchical cluster analysis of covariance patterns identified in the two first pairs of canonical variates

Initiate cluster analysis by creating objects to store scores for CDP and D.HL data sets

```
scores.CDP1 <- cca.out$X %*% cca.out$coef$X[, 1]
scores.DHL1 <- cca.out$Y %*% cca.out$coef$Y[, 1]
scores.CDP2 <- cca.out$X %*% cca.out$coef$X[, 2]
scores.DHL2 <- cca.out$Y %*% cca.out$coef$Y[, 2]
```

Control whether the correlation between the variates (scores) match the canonical correlations

```
cor(scores.CDP1, scores.DHL1)
```

```
##           [,1]
## [1,] 0.6248177
```

```
cor(scores.CDP2, scores.DHL2)
```

```
##           [,1]
## [1,] 0.4431836
```

Prepare data

```
data <- data.frame(hcscores.X1 = scores.CDP1,
                   hcscores.Y1 = scores.DHL1,
                   hcscores.X2 = scores.CDP2,
                   hcscores.Y2 = scores.DHL2)
data.sc <- scale(data)
```

Define distance measures and linkage methods

```
distances <- c("euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski")
linkages<- c("average", "single", "complete", "ward")
```

Create function to evaluate different agglomerative coefficients with different measures of distance

```
library(cluster)
clustEv <- function(dist.m, link.m) {
  dist.m <- dist(data.sc, method = dist.m)
  res.agnes <- agnes(dist.m, method = link.m)
  return(res.agnes$ac)
}
```

Inspect agglomerative coefficient for different combinations of distance measures and linkage methods

```
library(dplyr)
set.seed(2)
res <- expand.grid(distance = distances, linkage = linkages) %>%
  rowwise() %>%
  mutate(ac = clustEv(distance, linkage)) %>%
  ungroup()

res
```

```
## # A tibble: 24 x 3
##    distance  linkage      ac
##    <fct>     <fct>     <dbl>
##  1 euclidean average   0.867
```

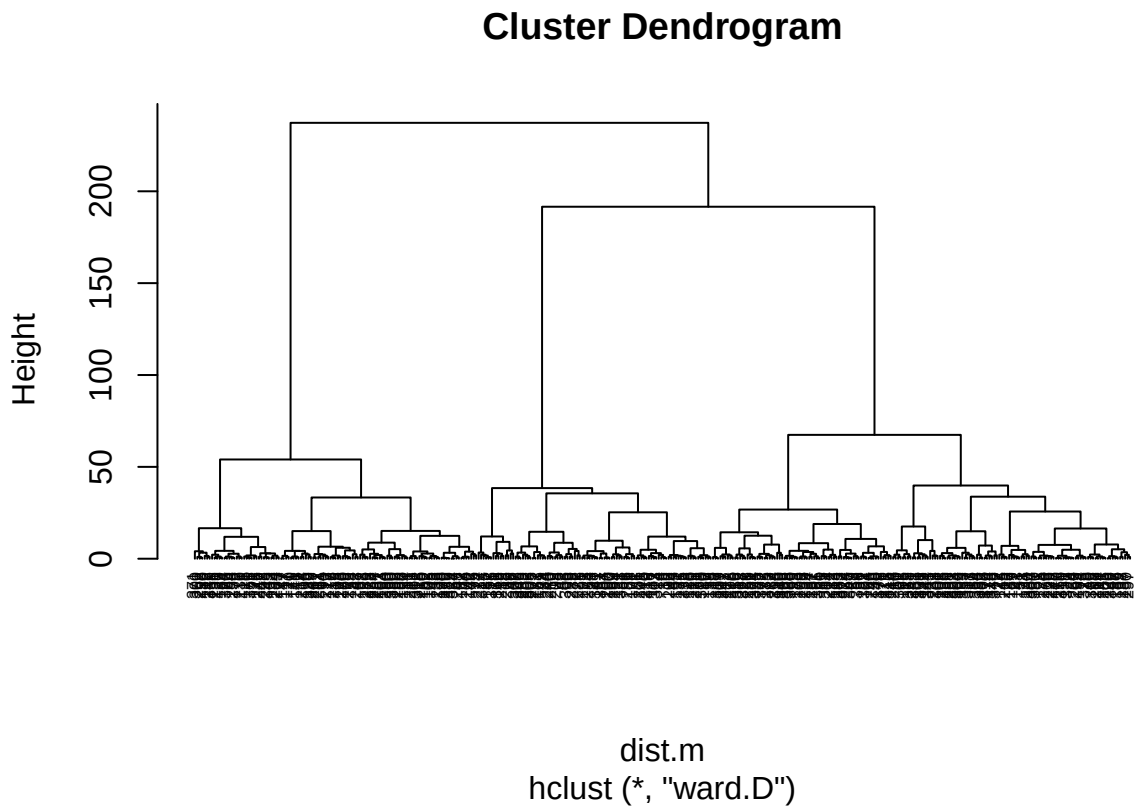
```
##  2 maximum   average   0.835
##  3 manhattan average   0.877
##  4 canberra  average   0.727
##  5 binary    average NaN
##  6 minkowski average   0.867
##  7 euclidean single    0.700
##  8 maximum   single    0.703
##  9 manhattan single    0.694
## 10 canberra  single    0.606
## # i 14 more rows
```

Run hierarchical clustering

```
dist.m <- dist(data.sc, method = "manhattan")
hclust.res <- hclust(dist.m, method = "ward.D")
```

View dendrogram

```
plot(hclust.res, hang = -1, cex = 0.5)
```



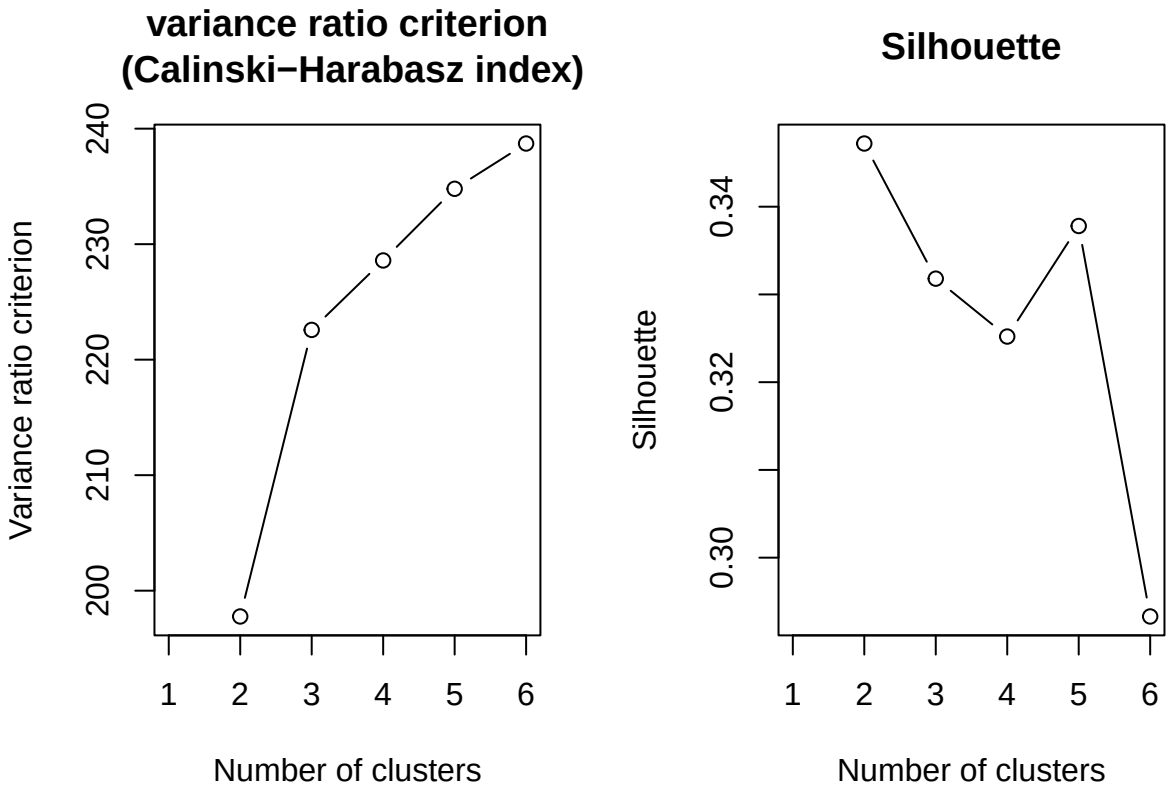
Refine predict.cancor() function for CH and SH index

```
predict.cancor <- function(cancor.obj) {
  pred.X <- cancor.obj$X %*% cancor.obj$coef$X
  pred.Y <- cancor.obj$Y %*% cancor.obj$coef$Y
  pred.XY <- list(pred.X,pred.Y)
  names(pred.XY) <- c("pred.X", "pred.Y")
  return(pred.XY)
}
```

```
canR <- predict.cancor(cca.out)
cca.data <- as.data.frame(cbind(canR$pred.X[, 1], canR$pred.X[, 2]))
```

Inspect CH and SH index

```
library(NbClust)
hcfit.ch <- NbClust(cca.data, distance = "manhattan", method = "ward.D",
                    min.nc = 1, max.nc = 6, index = "ch")
hcfit.sl <- NbClust(cca.data, distance = "manhattan", method = "ward.D",
                    min.nc = 1, max.nc = 6, index = "silhouette")

par(mfrow = c(1,2))
plot(names(hcfit.ch$All.index), hcfit.ch$All.index,
     main = "variance ratio criterion\n (Calinski-Harabasz index)",
     xlab = "Number of clusters", ylab = "Variance ratio criterion", type = 'b')
plot(names(hcfit.sl$All.index), hcfit.sl$All.index,
     main = "Silhouette", xlab = "Number of clusters", ylab = "Silhouette", type = 'b')
```



Cut tree at numbers of clusters indicated by CH and SH index

```
cut.sh <- cutree(hclust.res, k = 2)
cut.ch <- cutree(hclust.res, k = 6)
```

Stability of clusters defined by CH and SH index

Jaccard Similarity Index over bootstrapp samples against original canonical pair of variates

```r
library(fpc)
# Define clustering method for clusterboot()-function
hclust.manhattan <- function(x, k) {
  dist.matrix <- dist(x, method = "manhattan")
  dist.m <- as.matrix(dist.m)
  hc <- hclust(dist.matrix, method = "ward.D")
  clusters <- cutree(hc, k = k)
  clusterlist <- lapply(1:k, function(i) clusters == i)
  list(
    result = hc,
    nc = k,
    clusterlist = clusterlist,
    partition = clusters,
    clustermethod = "hierarchical"
  )
}
```

```r
# Compute mean JSI over 1000 bootstrap samples
clustboot.sh <- clusterboot(cca.data, B=1000, bootmethod = "boot",
                                              clustermethod = hclust.manhattan,
                                              k = 2, count=TRUE)
clustboot.ch <- clusterboot(cca.data, B=1000, bootmethod = "boot",
                                              clustermethod = hclust.manhattan,
                                              k = 6, count=TRUE)
```

```r
# View results
clustboot.sh$bootmean
```

```
## [1] 0.7904199 0.7847414
```

```r
clustboot.ch$bootmean
```

```
## [1] 0.6702875 0.6971898 0.6125810 0.6042625 0.4850620 0.6296715
```

Significance of clusters defined by CH and SH index

Create function that performs hierarchical clustering and returns the highest clustering indexes

```r
library(NbClust)
cluster.test <- function(cca){
  hcfit <- NbClust::NbClust(cca, distance = "manhattan", method = "ward.D",
                            index = "ch", min.nc = 2, max.nc = 6)
  CH.index <- max(hcfit$All.index)
  hcfit <- NbClust::NbClust(cca, distance = "manhattan", method = "ward.D",
                            index = "silhouette", min.nc = 2, max.nc = 6)
  sil.index <- max(hcfit$All.index)
  return(c("CH"=CH.index, "Silhouette"=sil.index))
}
```

Fit a multivariate normal distribution to the original data

```r
sigma <- cov(cca.data)
mu <- colMeans(cca.data)
real.CI <- cluster.test(cca.data)
```

Repeatedly perform parallel hierarchical clustering on resamples to create null distribution of clustering

indices

```
library(parallel)
library(MASS)
null.CI <- list()
n.sims <- 1999
n.cores <- detectCores() - 1
cl <- makeCluster(n.cores)
clusterExport(cl, c("mvrnorm", "mu", "sigma", "nrow", "cca.data", "cluster.test"))

null.CI <- parLapply(cl, 1:n.sims, function(i) {
  set.seed(3 + i)
  rand.sample <- mvrnorm(n = nrow(cca.data), mu = mu, Sigma = sigma)
  cluster.test(rand.sample)
})

stopCluster(cl)
null.CI <- as.data.frame(do.call(rbind, null.CI))
```

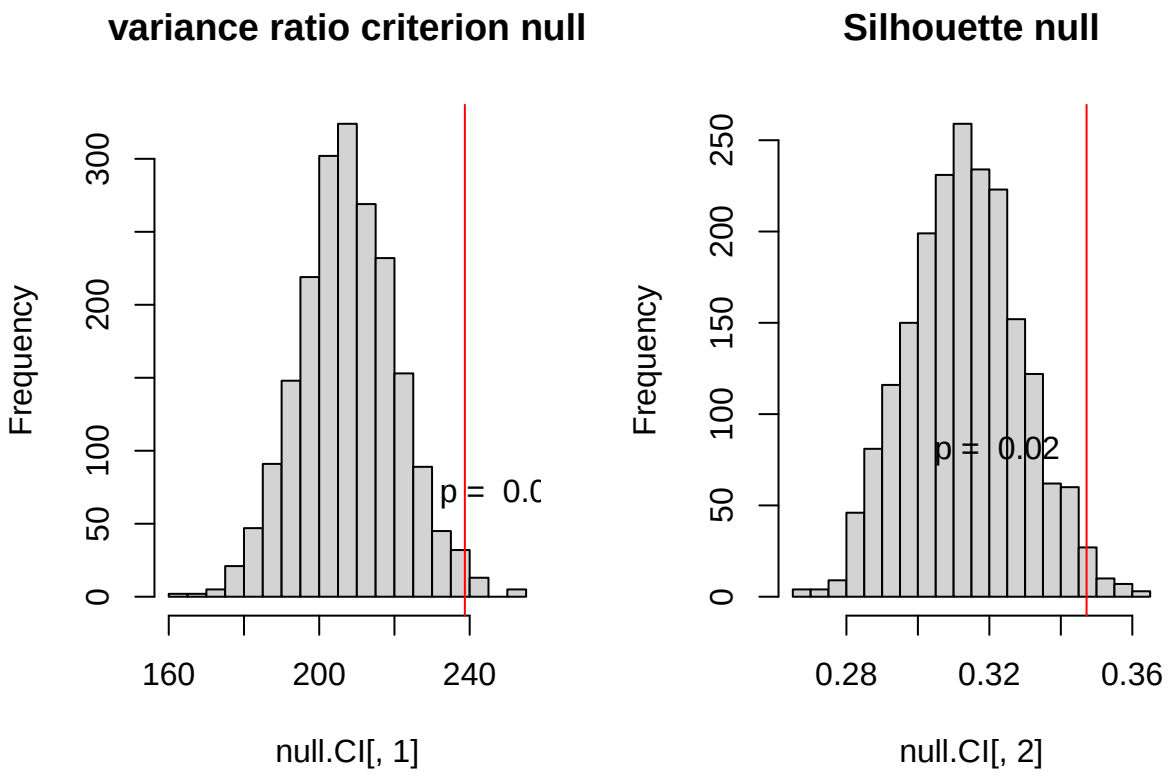
print p-values

```
rank.cv1 <- sum(real.CI[1] < null.CI[, 1]) +1
pval.cv1 <- rank.cv1 / (n.sims+1)

rank.cv2 <- sum(real.CI[2] < null.CI[, 2]) + 1
pval.cv2 <- rank.cv2 / (n.sims + 1)

t(t(c("p.val variance ratio" = pval.cv1, "p.val Silhouette" = pval.cv2)))
```

```
##                         [,1]
## p.val variance ratio 0.0115
## p.val Silhouette     0.0175
```

Visualize p-values

```
par(mfrow=c(1,2))
hist(null.CI[,1], breaks = 30, main = "variance ratio criterion null")
abline(v=real.CI[1], col="red")
text(real.CI[1] + 10, 70, paste('p = ', round(pval.cv1, 2)))

hist(null.CI[,2], breaks = 30, main = "Silhouette null")
abline(v=real.CI[2], col="red")
text(real.CI[2] - 0.025, 80, paste('p = ', round(pval.cv2, 2)))
```



Selecting 2 clusters to merge with data as this appears most stable according to JSI and silhouette

```
library(dplyr)
data <- mutate(data, Cluster = cut.sh)
CCA.cluster <- cbind(IBD, data)
```

4. Comparance between clusters

Refine data for further analysis

```
CCA.cluster$OMAS_37 <- IBD.original$OMAS37_sum
# Define categorical variables
categorical <- c("gender", "education", "marital_status", "work_status",
                 "native_language", "diagnose", "surgery", "UC_localisation",
                 "CD_localisation", "CD_behaviour", "IBD_disease_activity",
                 "treatment", "ASA_5", "immunosuppressive", "biological",
                 "corticosteroids", "EQ5D_mobility", "EQ5D_self_care",
                 "EQ5D_usual_activities", "EQ5D_pain_discomfort",
                 "EQ5D_anxiety_depression", "Cluster")
for (variable.name in categorical) {
  CCA.cluster[, variable.name] <- as.factor(CCA.cluster[, variable.name])
}
# Define values for categorical variables
CCA.cluster$gender <- factor(
```

```
  CCA.cluster$gender, levels = c(1:2),
  labels = c("Female" ,"Male")
  )

CCA.cluster$education <- factor(
  CCA.cluster$education, levels = c(0:3),
  labels = c("Elementary school", "Secondary school",
             "University college or university, up to 4 years",
             "University college or university, over 5 years"),
  ordered = TRUE
  )

CCA.cluster$marital_status <- factor(
  CCA.cluster$marital_status,
  levels = c(0:1),
  labels = c("Single", "In a relationship")
  )

CCA.cluster$work_status <- factor(
  CCA.cluster$work_status, levels = c(0:1),
  labels = c("Not working", "Working")
  )

CCA.cluster$native_language <- factor(
  CCA.cluster$native_language,
  levels = c(0:1),
  labels = c("Norwegian", "Other language")
  )

CCA.cluster$diagnose <- factor(
  CCA.cluster$diagnose, levels = c(1:2),
  labels = c("Crohn's disease", "Ulcerative colitis")
  )

CCA.cluster$surgery <- factor(
  CCA.cluster$surgery, levels = c(0:1),
  labels = c("No surgery", "Surgery")
  )

CCA.cluster$UC_localisation <- factor(
  CCA.cluster$UC_localisation,
  levels = c(0:3),
  labels = c("NA", "Ulcerative proctitis", "Left-sided UC", "Extensive UC")
  )

CCA.cluster$CD_localisation <- factor(
  CCA.cluster$CD_localisation,
  levels = c(0:4),
  labels = c("NA", "Ileal", "Colonic",
             "Ileocolonic", "Upper tract only or modifier")
  )

CCA.cluster$CD_behaviour <- factor(
```

```
  CCA.cluster$CD_behaviour, levels = c(0:4),
  labels = c("NA", "Non-stricturing, non-penetrating",
             "Stricturing", "Penetrating", "Perianal disease")
  )

CCA.cluster$IBD_disease_activity <- factor(
  CCA.cluster$IBD_disease_activity, levels = c(0:1),
  labels = c("Below treshold", "Over treshold")
  )

CCA.cluster$treatment <- factor(
  CCA.cluster$treatment, levels = c(0:1),
  labels = c("No", "Yes")
  )

CCA.cluster$ASA_5 <- factor(
  CCA.cluster$ASA_5, levels = c(0:1),
  labels = c("No", "Yes")
  )

CCA.cluster$immunosuppressive <- factor(
  CCA.cluster$immunosuppressive,
  levels = c(0:1), labels = c("No", "Yes")
  )

CCA.cluster$biological <- factor(
  CCA.cluster$biological, levels = c(0:1),
  labels = c("No", "Yes")
  )

CCA.cluster$corticosteroids <- factor(
  CCA.cluster$corticosteroids, levels = c(0:1),
  labels = c("No", "Yes")
  )

CCA.cluster$EQ5D_mobility <- factor(
  CCA.cluster$EQ5D_mobility, levels = c(1:4),
  labels = c("No problems", "Slight problems",
             "Moderate problems", "Severe problems"),
  ordered = TRUE)

CCA.cluster$EQ5D_self_care <- factor(
  CCA.cluster$EQ5D_self_care, levels = c(1:4),
  labels = c("No problems", "Slight problems",
             "Moderate problems", "Severe problems"),
  ordered = TRUE)

CCA.cluster$EQ5D_usual_activities <- factor(
  CCA.cluster$EQ5D_usual_activities, levels = c(1:5),
  labels = c("No problems", "Slight problems", "Moderate problems",
             "Severe problems", "Unable to do"), ordered = TRUE)

CCA.cluster$EQ5D_pain_discomfort <- factor(
```

```
  CCA.cluster$EQ5D_pain_discomfort, levels = c(1:5),
  labels = c("None", "Slight", "Moderate", "Severe", "Extreme"),
  ordered = TRUE)

CCA.cluster$EQ5D_anxiety_depression <- factor(
  CCA.cluster$EQ5D_anxiety_depression, levels = c(1:5),
  labels = c("None", "Slight", "Moderate", "Severe", "Extreme"),
  ordered = TRUE)
```

Summary of variables within each cluster

```
Cluster.1 <- subset(CCA.cluster, Cluster == 1)
Cluster.2 <- subset(CCA.cluster, Cluster == 2)
summary(Cluster.1)
```

```
##     gender        age
##  Female:132   Length:265
##  Male  :133   Class :character
##               Mode  :character
##
##
##
##
##                                              education            marital_status
##  Elementary school                                : 9   Single           : 84
##  Secondary school                                 :72   In a relationship:181
##  University college or university, up to 4 years:96
##  University college or university, over 5 years :87
##  NA's                                             : 1
##
##
##       work_status        native_language               diagnose
##  Not working: 72   Norwegian     :235    Crohn's disease   :147
##  Working    :193   Other language: 30    Ulcerative colitis:118
##
##
##
##
##
##  disease_duration         surgery                UC_localisation
##  Length:265         No surgery:171   NA                    :148
##  Class :character   Surgery   : 94   Ulcerative proctitis: 26
##  Mode  :character                    Left-sided UC       : 30
##                                      Extensive UC        : 61
##
##
##
##                      CD_localisation                           CD_behaviour
##  NA                          :118    NA                               :125
##  Ileal                       : 38    Non-stricturing, non-penetrating: 61
##  Colonic                     : 24    Stricturing                      : 29
##  Ileocolonic                 : 75    Penetrating                      : 12
##  Upper tract only or modifier: 10    Perianal disease                 : 38
##
```

```
##
##      IBD_disease_activity treatment ASA_5     immunosuppressive biological
##  Below treshold:152       No : 17   No :210   No :254           No : 68
##  Over treshold :108       Yes:248   Yes: 55   Yes: 11           Yes:197
##  NA's          :  5
##
##
##
##
##  corticosteroids Calprotectin                EQ5D_mobility
##  No :239         Length:265         No problems      :210
##  Yes: 26         Class :character   Slight problems  : 42
##                  Mode  :character   Moderate problems:  9
##                                     Severe problems  :  4
##
##
##
##            EQ5D_self_care       EQ5D_usual_activities EQ5D_pain_discomfort
##  No problems      :244    No problems      :150       None    : 73
##  Slight problems  : 12    Slight problems  : 84       Slight  :127
##  Moderate problems:  8    Moderate problems: 22       Moderate: 48
##  Severe problems  :  1    Severe problems  :  7       Severe  : 16
##                           Unable to do     :  2       Extreme :  1
##
##
##  EQ5D_anxiety_depression   EQ5Dvas1           GSE_sum
##  None    :125            Length:265         Length:265
##  Slight  : 87            Class :character   Class :character
##  Moderate: 39            Mode  :character   Mode  :character
##  Severe  : 11
##  Extreme :  2
##  NA's    :  1
##
##    BIPQ_sum          HLQ.Scale1         HLQ.Scale2         HLQ.Scale3
##  Length:265         Length:265         Length:265         Length:265
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HLQ.Scale4         HLQ.Scale5         HLQ.Scale6         HLQ.Scale7
##  Length:265         Length:265         Length:265         Length:265
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HLQ.Scale8         HLQ.Scale9        eHLQ.Domain1       eHLQ.Domain2
##  Length:265         Length:265         Length:265         Length:265
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##  eHLQ.Domain3       eHLQ.Domain4       eHLQ.Domain5       eHLQ.Domain6
##  Length:265         Length:265         Length:265         Length:265
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  eHLQ.Domain7        OMAS37_sum         hcscores.X1       hcscores.Y1
##  Length:265         Length:265         Min.   :-3.4735   Min.   :-3.6186
##  Class :character   Class :character   1st Qu.:-0.8963   1st Qu.:-0.8813
##  Mode  :character   Mode  :character   Median :-0.3636   Median :-0.4288
##                                        Mean   :-0.3651   Mean   :-0.4105
##                                        3rd Qu.: 0.2206   3rd Qu.: 0.1586
##                                        Max.   : 1.6210   Max.   : 1.4288
##
##   hcscores.X2       hcscores.Y2      Cluster    OMAS_37
##  Min.   :-1.8615   Min.   :-3.1314   1:265   Min.   : 0.000
##  1st Qu.:-0.6050   1st Qu.:-0.4588   2:  0   1st Qu.: 0.000
##  Median : 0.3290   Median : 0.2245           Median : 2.000
##  Mean   : 0.2513   Mean   : 0.1436           Mean   : 4.291
##  3rd Qu.: 0.9870   3rd Qu.: 0.7893           3rd Qu.: 5.000
##  Max.   : 2.7924   Max.   : 3.3774           Max.   :38.000
##                                              NA's   :42
```

```
summary(Cluster.2)
```

```
##    gender       age
##  Female:46   Length:115
##  Male  :69   Class :character
##              Mode  :character
##
##
##
##
##                                          education            marital_status
##  Elementary school                            : 2   Single           :35
##  Secondary school                             :24   In a relationship:80
##  University college or university, up to 4 years:34
##  University college or university, over 5 years :55
##
##
##
##       work_status        native_language               diagnose
##  Not working: 14   Norwegian     :104    Crohn's disease   :60
##  Working    :101   Other language: 11    Ulcerative colitis:55
##
##
##
##
##
##  disease_duration         surgery               UC_localisation
```

```
##  Length:115         No surgery:85   NA                  :60
##  Class :character   Surgery   :30   Ulcerative proctitis: 5
##  Mode  :character                   Left-sided UC       :11
##                                     Extensive UC        :39
##
##
##
##                      CD_localisation                            CD_behaviour
##  NA                          :55     NA                                :58
##  Ileal                       :15     Non-stricturing, non-penetrating:28
##  Colonic                     : 8     Stricturing                       : 8
##  Ileocolonic                 :26     Penetrating                       : 2
##  Upper tract only or modifier:11     Perianal disease                  :19
##
##
##      IBD_disease_activity treatment ASA_5    immunosuppressive biological
##  Below treshold:93        No :  2   No :96   No :110           No : 14
##  Over treshold :21        Yes:113   Yes:19   Yes:  5           Yes:101
##  NA's          : 1
##
##
##
##
##  corticosteroids Calprotectin                EQ5D_mobility
##  No :108         Length:115         No problems      :107
##  Yes:  7         Class :character   Slight problems  :  7
##                  Mode  :character   Moderate problems:  1
##                                     Severe problems  :  0
##
##
##
##            EQ5D_self_care       EQ5D_usual_activities EQ5D_pain_discomfort
##  No problems      :114    No problems      :98        None    :67
##  Slight problems  :  1    Slight problems  :14        Slight  :41
##  Moderate problems:  0    Moderate problems: 2        Moderate: 5
##  Severe problems  :  0    Severe problems  : 1        Severe  : 2
##                           Unable to do     : 0        Extreme : 0
##
##
##  EQ5D_anxiety_depression   EQ5Dvas1           GSE_sum
##  None    :84             Length:115         Length:115
##  Slight  :21             Class :character   Class :character
##  Moderate: 9             Mode  :character   Mode  :character
##  Severe  : 1
##  Extreme : 0
##
##
##    BIPQ_sum          HLQ.Scale1         HLQ.Scale2         HLQ.Scale3
##  Length:115         Length:115         Length:115         Length:115
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##   HLQ.Scale4         HLQ.Scale5         HLQ.Scale6         HLQ.Scale7
##  Length:115         Length:115         Length:115         Length:115
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HLQ.Scale8         HLQ.Scale9        eHLQ.Domain1       eHLQ.Domain2
##  Length:115         Length:115         Length:115         Length:115
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  eHLQ.Domain3       eHLQ.Domain4       eHLQ.Domain5       eHLQ.Domain6
##  Length:115         Length:115         Length:115         Length:115
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  eHLQ.Domain7        OMAS37_sum         hcscores.X1       hcscores.Y1
##  Length:115         Length:115         Min.   :-0.7840   Min.   :-0.8410
##  Class :character   Class :character   1st Qu.: 0.4333   1st Qu.: 0.4203
##  Mode  :character   Mode  :character   Median : 0.9400   Median : 0.9202
##                                        Mean   : 0.8414   Mean   : 0.9460
##                                        3rd Qu.: 1.2454   3rd Qu.: 1.4859
##                                        Max.   : 2.3579   Max.   : 2.5955
##
##   hcscores.X2       hcscores.Y2      Cluster    OMAS_37
##  Min.   :-2.3015   Min.   :-1.9356   1:  0   Min.   : 0.000
##  1st Qu.:-1.0733   1st Qu.:-0.7546   2:115   1st Qu.: 0.000
##  Median :-0.5846   Median :-0.3554           Median : 2.000
##  Mean   :-0.5791   Mean   :-0.3309           Mean   : 2.943
##  3rd Qu.:-0.3085   3rd Qu.: 0.1618           3rd Qu.: 4.000
##  Max.   : 2.1909   Max.   : 1.3665           Max.   :18.000
##                                              NA's   :10
```

Comparance between clusters for variables excluded from CCA and hierarchical clustering

Extract external variables

```
ext.IBD <- subset(CCA.cluster, select = c(gender, education, marital_status, work_status,
                                          native_language, diagnose, surgery,
                                          UC_localisation, CD_localisation, CD_behaviour,
                                          IBD_disease_activity, treatment, ASA_5,
                                          immunosuppressive, biological, corticosteroids,
                                          EQ5D_mobility, EQ5D_self_care,
```

```
                                        EQ5D_usual_activities, EQ5D_pain_discomfort,
                                        EQ5D_anxiety_depression, OMAS_37, Cluster))
```

Bivariate analyes

Permuted t-test of continuous variable

```
library(MKinfer)
```

```
## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package
```

```
library(dplyr)

ext.na <- na.omit(ext.IBD)
ext.na %>%
  group_by(Cluster) %>%
  summarise(mean.OMAS_37 = mean(OMAS_37), sd.OMAS_37 = sd(OMAS_37))
```

```
## # A tibble: 2 x 3
##   Cluster mean.OMAS_37 sd.OMAS_37
##   <fct>          <dbl>      <dbl>
## 1 1               4.33       6.73
## 2 2               2.94       3.76
```

```
set.seed(4)
perm.t.OMAS <- perm.t.test(OMAS_37 ~ Cluster, R = 10000, data = ext.na)
perm.t.OMAS
```

```
##
##  Permutation Welch Two Sample t-test
##
## data:  OMAS_37 by Cluster
## number of permutations:  10000
## (Monte-Carlo) permutation p-value = 0.0198
## permutation difference of means (SE) = 1.384981 (0.7052791)
## 95 percent (Monte-Carlo) permutation percentile confidence interval:
##  0.007326007 2.719075630
##
## Results without permutation:
## t = 2.3802, df = 315.83, p-value = 0.0179
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.240580 2.534339
## sample estimates:
## mean in group 1 mean in group 2
##        4.330317        2.942857
```

Chi-square/Fisher's test for nominal variables

```
ext.nom <- subset(ext.IBD, select = c(gender, marital_status, work_status,
                                      native_language, diagnose, surgery,
                                      UC_localisation, CD_localisation,
                                      CD_behaviour, IBD_disease_activity,
```

```
                                    treatment, ASA_5, immunosuppressive,
                                    biological, corticosteroids, Cluster))

nom.vars <- names(ext.nom)
nom.vars <- nom.vars[-16]

# Create empty lists to store results
test.stat <- list()
p.val <- numeric()

# Designate variables to Chi-square/Fisher test depending on contingency table

for (i in 1:length(nom.vars)) {
   cat.tab <- table(ext.nom[, i], ext.nom$Cluster)
  if (any(cat.tab < 5)) {
    fisher <- fisher.test(cat.tab, workspace = 2e8, hybrid = T)
    test.stat[[i]] <- fisher
    p.val[i] <- fisher$p.value
  } else {
    chisq <- chisq.test(cat.tab)
    test.stat[[i]] <- chisq
    p.val[i] <- chisq$p.value
  }
}

# Give corresponding names in list
names(test.stat) <- nom.vars

# Adjust p-values with Hochberg correction
adj.p <- p.adjust(p.val, method = "hochberg")

# View significant results and adjusted p-value
sig.chi <- test.stat[adj.p < .05]
sig.p <- adj.p[adj.p < .05]
sig.chi
```

```
## $work_status
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cat.tab
## X-squared = 9.4611, df = 1, p-value = 0.002099
##
##
## $IBD_disease_activity
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cat.tab
## X-squared = 17.735, df = 1, p-value = 2.539e-05
```

```
sig.p
```

```
## [1] 0.0293822843 0.0003808125
```

```r
# Lower threshold for considering variables into logistic model
mod.chi <- test.stat[adj.p < .1]
mod.chi
```

```
## $work_status
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cat.tab
## X-squared = 9.4611, df = 1, p-value = 0.002099
##
##
## $IBD_disease_activity
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cat.tab
## X-squared = 17.735, df = 1, p-value = 2.539e-05
##
##
## $biological
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cat.tab
## X-squared = 7.8412, df = 1, p-value = 0.005107
```

Repeat procedure with Kruskal-Wallis test for ordinal variables

```r
ext.ord <- subset(ext.IBD, select = c(education, EQ5D_mobility, EQ5D_self_care,
                                      EQ5D_usual_activities, EQ5D_pain_discomfort,
                                      EQ5D_anxiety_depression, Cluster))

ord.vars <- names(ext.ord)
ord.vars <- ord.vars[-7]

kw.test <- list()
kw.p <- numeric()

for (i in 1:length(ord.vars)) {
  kw <- kruskal.test(Cluster ~ ext.ord[, i], data = ext.ord)
  kw.test[[i]] <- kw
  kw.p[i] <- kw$p.value
}

names(kw.test) <- ord.vars
kw.p.adj <- p.adjust(kw.p, method = "hochberg")
sig.kw <- kw.test[kw.p.adj < .05]
sig.kw.p <- kw.p.adj[kw.p.adj < .05]
sig.kw
```

```
## $EQ5D_mobility
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 11.409, df = 3, p-value = 0.00971
##
##
## $EQ5D_usual_activities
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 29.37, df = 4, p-value = 6.574e-06
##
##
## $EQ5D_pain_discomfort
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 37.625, df = 4, p-value = 1.339e-07
##
##
## $EQ5D_anxiety_depression
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 22.275, df = 4, p-value = 0.0001767
```

```
sig.kw.p
```

```
## [1] 2.912913e-02 3.286764e-05 8.034232e-07 7.067479e-04
```

```
mod.kw <- kw.test[kw.p.adj < .1]
mod.kw
```

```
## $education
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 7.8535, df = 3, p-value = 0.04914
##
##
## $EQ5D_mobility
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 11.409, df = 3, p-value = 0.00971
##
##
## $EQ5D_self_care
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 7.4478, df = 3, p-value = 0.05892
```

```
##
##
## $EQ5D_usual_activities
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 29.37, df = 4, p-value = 6.574e-06
##
##
## $EQ5D_pain_discomfort
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 37.625, df = 4, p-value = 1.339e-07
##
##
## $EQ5D_anxiety_depression
##
##  Kruskal-Wallis rank sum test
##
## data:  Cluster by ext.ord[, i]
## Kruskal-Wallis chi-squared = 22.275, df = 4, p-value = 0.0001767
```

Binomial logistic model for cluster membership

Prepare data

```
library(car)

# Assess multicollinearity
response <- rnorm(nrow(ext.IBD))
mc <- lm(response ~., data = ext.IBD)
## vif(mc)
## Error in vif.default(mc) : there are aliased coefficients in the model
alias(mc)
```

```
## Model :
## response ~ gender + education + marital_status + work_status +
##     native_language + diagnose + surgery + UC_localisation +
##     CD_localisation + CD_behaviour + IBD_disease_activity + treatment +
##     ASA_5 + immunosuppressive + biological + corticosteroids +
##     EQ5D_mobility + EQ5D_self_care + EQ5D_usual_activities +
##     EQ5D_pain_discomfort + EQ5D_anxiety_depression + OMAS_37 +
##     Cluster
##
## Complete :
##                                            (Intercept) genderMale education.L
## CD_localisationUpper tract only or modifier  1           0          0
##                                            education.Q education.C
## CD_localisationUpper tract only or modifier  0           0
##                                            marital_statusIn a relationship
## CD_localisationUpper tract only or modifier  0
```

```
##                                                 work_statusWorking
## CD_localisationUpper tract only or modifier  0
##                                                 native_languageOther language
## CD_localisationUpper tract only or modifier  0
##                                                 diagnoseUlcerative colitis
## CD_localisationUpper tract only or modifier -1
##                                                 surgerySurgery
## CD_localisationUpper tract only or modifier  0
##                                                 UC_localisationUlcerative proctitis
## CD_localisationUpper tract only or modifier  0
##                                                 UC_localisationLeft-sided UC
## CD_localisationUpper tract only or modifier  0
##                                                 UC_localisationExtensive UC
## CD_localisationUpper tract only or modifier  0
##                                                 CD_localisationIleal
## CD_localisationUpper tract only or modifier -1
##                                                 CD_localisationColonic
## CD_localisationUpper tract only or modifier -1
##                                                 CD_localisationIleocolonic
## CD_localisationUpper tract only or modifier -1
##                                               CD_behaviourNon-stricturing, non-penetrating
## CD_localisationUpper tract only or modifier  0
##                                                 CD_behaviourStricturing
## CD_localisationUpper tract only or modifier  0
##                                                 CD_behaviourPenetrating
## CD_localisationUpper tract only or modifier  0
##                                                 CD_behaviourPerianal disease
## CD_localisationUpper tract only or modifier  0
##                                                 IBD_disease_activityOver treshold
## CD_localisationUpper tract only or modifier  0
##                                                 treatmentYes ASA_5Yes
## CD_localisationUpper tract only or modifier  0            0
##                                                 immunosuppressiveYes biologicalYes
## CD_localisationUpper tract only or modifier  0                    0
##                                                 corticosteroidsYes EQ5D_mobility.L
## CD_localisationUpper tract only or modifier  0                  0
##                                                 EQ5D_mobility.Q EQ5D_mobility.C
## CD_localisationUpper tract only or modifier  0               0
##                                                 EQ5D_self_care.L EQ5D_self_care.Q
## CD_localisationUpper tract only or modifier  0                0
##                                                 EQ5D_self_care.C
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_usual_activities.L
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_usual_activities.Q
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_usual_activities.C
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_usual_activities^4
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_pain_discomfort.L
## CD_localisationUpper tract only or modifier  0
##                                                 EQ5D_pain_discomfort.Q
## CD_localisationUpper tract only or modifier  0
```

```
##                                           EQ5D_pain_discomfort.C
## CD_localisationUpper tract only or modifier  0
##                                           EQ5D_pain_discomfort^4
## CD_localisationUpper tract only or modifier  0
##                                           EQ5D_anxiety_depression.L
## CD_localisationUpper tract only or modifier  0
##                                           EQ5D_anxiety_depression.Q
## CD_localisationUpper tract only or modifier  0
##                                           EQ5D_anxiety_depression.C
## CD_localisationUpper tract only or modifier  0
##                                           EQ5D_anxiety_depression^4 OMAS_37
## CD_localisationUpper tract only or modifier  0                          0
##                                           Cluster2
## CD_localisationUpper tract only or modifier  0
```

```
#remove CD_localisation
ext.IBD <- subset(ext.IBD, select = - CD_localisation)
response <- rnorm(nrow(ext.IBD))
mc <- lm(response ~., data = ext.IBD)
vif(mc)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## gender                   1.224890  1        1.106748
## education                1.750005  3        1.097758
## marital_status           1.177338  1        1.085052
## work_status              1.524549  1        1.234726
## native_language          1.205062  1        1.097753
## diagnose                99.486985  1        9.974316
## surgery                  1.991081  1        1.411057
## UC_localisation        120.533917  3        2.222550
## CD_behaviour            21.145039  4        1.464371
## IBD_disease_activity     1.846057  1        1.358697
## treatment                2.108456  1        1.452052
## ASA_5                    2.732090  1        1.652903
## immunosuppressive        1.204170  1        1.097347
## biological               3.348487  1        1.829887
## corticosteroids          1.435325  1        1.198050
## EQ5D_mobility            4.078327  3        1.264000
## EQ5D_self_care           7.519911  3        1.399702
## EQ5D_usual_activities    8.980532  4        1.315718
## EQ5D_pain_discomfort     7.337568  4        1.282904
## EQ5D_anxiety_depression  3.503138  4        1.169653
## OMAS_37                  1.364345  1        1.168052
## Cluster                  1.335818  1        1.155776
```

```
# Remove UC_localisation and CD_behaviour
ext.IBD <- subset(ext.IBD, select = -c(UC_localisation, CD_behaviour))
response <- rnorm(nrow(ext.IBD))
mc <- lm(response ~., data = ext.IBD)
vif(mc)
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## gender                  1.142348  1        1.068807
## education               1.557646  3        1.076659
## marital_status          1.158507  1        1.076340
## work_status             1.454780  1        1.206143
```

```
## native_language         1.163806  1        1.078798
## diagnose                1.851584  1        1.360729
## surgery                 1.541336  1        1.241506
## IBD_disease_activity    1.817033  1        1.347974
## treatment               2.077494  1        1.441351
## ASA_5                   2.626751  1        1.620725
## immunosuppressive       1.173508  1        1.083286
## biological              3.239612  1        1.799892
## corticosteroids         1.402909  1        1.184445
## EQ5D_mobility           3.914048  3        1.255368
## EQ5D_self_care          6.808438  3        1.376706
## EQ5D_usual_activities   8.218717  4        1.301219
## EQ5D_pain_discomfort    7.077029  4        1.277119
## EQ5D_anxiety_depression 3.235776  4        1.158103
## OMAS_37                 1.303712  1        1.141802
## Cluster                 1.307567  1        1.143489
```

```
# View summary
summary(ext.IBD)
```

```
##     gender                                              education
##  Female:178   Elementary school                                : 11
##  Male  :202   Secondary school                                 : 96
##               University college or university, up to 4 years:130
##               University college or university, over 5 years :142
##               NA's                                            :  1
##
##
##            marital_status      work_status        native_language
##  Single            :119    Not working: 86   Norwegian     :339
##  In a relationship:261    Working    :294   Other language: 41
##
##
##
##
##
##                diagnose         surgery        IBD_disease_activity treatment
##  Crohn's disease   :207   No surgery:256   Below treshold:245       No : 19
##  Ulcerative colitis:173   Surgery   :124   Over treshold :129       Yes:361
##                                            NA's          :  6
##
##
##
##
##  ASA_5     immunosuppressive biological corticosteroids           EQ5D_mobility
##  No :306   No :364           No : 82    No :347         No problems      :317
##  Yes: 74   Yes: 16           Yes:298    Yes: 33         Slight problems  : 49
##                                                         Moderate problems: 10
##                                                         Severe problems  :  4
##
##
##
##            EQ5D_self_care       EQ5D_usual_activities EQ5D_pain_discomfort
##  No problems      :358    No problems      :248       None    :140
##  Slight problems  : 13    Slight problems  : 98       Slight  :168
```

```
##  Moderate problems:  8    Moderate problems: 24       Moderate: 53
##  Severe problems  :  1    Severe problems  :  8       Severe  : 18
##                           Unable to do     :  2       Extreme :  1
##
##
##  EQ5D_anxiety_depression    OMAS_37      Cluster
##  None    :209            Min.   : 0.00   1:265
##  Slight  :108            1st Qu.: 0.00   2:115
##  Moderate: 48            Median : 2.00
##  Severe  : 12            Mean   : 3.86
##  Extreme :  2            3rd Qu.: 5.00
##  NA's    :  1            Max.   :38.00
##                          NA's   :52
```

Build model including significant variables from bivariate analysis

```
glm.fit0 <- glm(Cluster ~ work_status + IBD_disease_activity + biological +
                  education + EQ5D_mobility + EQ5D_self_care +
                  EQ5D_usual_activities + EQ5D_pain_discomfort +
                  EQ5D_anxiety_depression + OMAS_37,
                data = ext.IBD, family = binomial(link = "logit"))
summary(glm.fit0)
```

```
##
## Call:
## glm(formula = Cluster ~ work_status + IBD_disease_activity +
##     biological + education + EQ5D_mobility + EQ5D_self_care +
##     EQ5D_usual_activities + EQ5D_pain_discomfort + EQ5D_anxiety_depression +
##     OMAS_37, family = binomial(link = "logit"), data = ext.IBD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5905  -0.7922  -0.4491   0.8565   2.3738
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -13.65210  979.41013  -0.014  0.98888
## work_statusWorking                 0.09994    0.44705   0.224  0.82311
## IBD_disease_activityOver treshold -0.79791    0.37669  -2.118  0.03416 *
## biologicalYes                      0.98243    0.37385   2.628  0.00859 **
## education.L                        0.47304    0.62679   0.755  0.45043
## education.Q                        0.02051    0.48438   0.042  0.96623
## education.C                        0.18495    0.32494   0.569  0.56924
## EQ5D_mobility.L                   -8.96851 1135.81805  -0.008  0.99370
## EQ5D_mobility.Q                   -6.37831  846.58898  -0.008  0.99399
## EQ5D_mobility.C                   -3.67906  378.60705  -0.010  0.99225
## EQ5D_self_care.L                  -6.10369 1991.69962  -0.003  0.99755
## EQ5D_self_care.Q                   5.74184 1607.79272   0.004  0.99715
## EQ5D_self_care.C                   9.79162 1096.87215   0.009  0.99288
## EQ5D_usual_activities.L           -9.08320 1000.79960  -0.009  0.99276
## EQ5D_usual_activities.Q           -8.45692  845.83014  -0.010  0.99202
## EQ5D_usual_activities.C           -6.66419  500.40075  -0.013  0.98937
## EQ5D_usual_activities^4           -3.02650  189.13585  -0.016  0.98723
## EQ5D_pain_discomfort.L            -1.41882 1728.94624  -0.001  0.99935
## EQ5D_pain_discomfort.Q            -0.58978 1461.22633   0.000  0.99968
```

```
## EQ5D_pain_discomfort.C              -1.39102  864.47342  -0.002  0.99872
## EQ5D_pain_discomfort^4              -0.92883  326.74089  -0.003  0.99773
## EQ5D_anxiety_depression.L           -0.61221 1728.94610   0.000  0.99972
## EQ5D_anxiety_depression.Q            0.67009 1461.22606   0.000  0.99963
## EQ5D_anxiety_depression.C            0.40063  864.47325   0.000  0.99963
## EQ5D_anxiety_depression^4            0.74892  326.74084   0.002  0.99817
## OMAS_37                              0.01702    0.02866   0.594  0.55248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.74  on 325  degrees of freedom
## Residual deviance: 327.68  on 300  degrees of freedom
##   (54 observations deleted due to missingness)
## AIC: 379.68
##
## Number of Fisher Scoring iterations: 15
```

Model is showing signs of complete separation in EQ5D-variables -> Recoding EQ5D-variables:

```
library(dplyr)
ext.IBD <- ext.IBD %>%
  mutate(EQ5D_mobility = case_when(
    EQ5D_mobility %in% c("No problems",
                         "Slight problems") ~ "None or slight problems",
    EQ5D_mobility == "Moderate problems" ~ "Moderate problems",
    EQ5D_mobility %in% c("Severe problems") ~ "Severe problems"
  ))
ext.IBD$EQ5D_mobility <- factor(
  ext.IBD$EQ5D_mobility,
  levels = c("None or slight problems",
             "Moderate problems",
             "Severe problems"),
  ordered = TRUE)

ext.IBD <- ext.IBD %>%
  mutate(EQ5D_self_care = case_when(
    EQ5D_self_care %in% c("No problems",
                          "Slight problems") ~ "None or slight problems",
    EQ5D_self_care == "Moderate problems" ~ "Moderate problems",
    EQ5D_self_care %in% c("Severe problems") ~ "Severe problems"
  ))
ext.IBD$EQ5D_self_care <- factor(
  ext.IBD$EQ5D_self_care,
  levels = c("None or slight problems",
             "Moderate problems",
             "Severe problems"),
  ordered = TRUE)

ext.IBD <- ext.IBD %>%
  mutate(EQ5D_usual_activities = case_when(
    EQ5D_usual_activities %in% c("No problems",
                                 "Slight problems") ~ "None or slight problems",
    EQ5D_usual_activities == "Moderate problems" ~ "Moderate problems",
```

```
    EQ5D_usual_activities %in% c("Severe problems") ~ "Severe problems"
  ))
ext.IBD$EQ5D_usual_activities <- factor(
  ext.IBD$EQ5D_usual_activities,
  levels = c("None or slight problems",
             "Moderate problems",
             "Severe problems"),
  ordered = TRUE)


ext.IBD <- ext.IBD %>%
  mutate(EQ5D_pain_discomfort = case_when(
    EQ5D_pain_discomfort %in% c("None",
                                "Slight") ~ "None or slight",
    EQ5D_pain_discomfort == "Moderate" ~ "Moderate",
    EQ5D_pain_discomfort %in% c("Severe",
                                "Extreme") ~ "Severe or extreme"
  ))
ext.IBD$EQ5D_pain_discomfort <- factor(
  ext.IBD$EQ5D_pain_discomfort,
  levels = c("None or slight",
             "Moderate",
             "Severe or extreme"),
  ordered = TRUE)

ext.IBD <- ext.IBD %>%
  mutate(EQ5D_anxiety_depression = case_when(
    EQ5D_anxiety_depression %in% c("None",
                                   "Slight") ~ "None or slight",
    EQ5D_anxiety_depression == "Moderate" ~ "Moderate",
    EQ5D_anxiety_depression %in% c("Severe",
                                   "Extreme") ~ "Severe or extreme"
  ))
ext.IBD$EQ5D_anxiety_depression <- factor(
  ext.IBD$EQ5D_anxiety_depression,
  levels = c("None or slight",
             "Moderate",
             "Severe or extreme"),
  ordered = TRUE)

summary(ext.IBD)
```

```
##     gender                                                education
##  Female:178   Elementary school                                : 11
##  Male  :202   Secondary school                                 : 96
##               University college or university, up to 4 years:130
##               University college or university, over 5 years :142
##               NA's                                            :  1
##
##
##            marital_status      work_status        native_language
##  Single           :119    Not working: 86   Norwegian     :339
##  In a relationship:261    Working    :294   Other language: 41
##
```

```
##
##
##
##
##                diagnose         surgery        IBD_disease_activity treatment
##  Crohn's disease   :207   No surgery:256   Below treshold:245       No : 19
##  Ulcerative colitis:173   Surgery   :124   Over treshold :129       Yes:361
##                                            NA's          :  6
##
##
##
##
##  ASA_5     immunosuppressive biological corticosteroids
##  No :306   No :364           No : 82    No :347
##  Yes: 74   Yes: 16           Yes:298    Yes: 33
##
##
##
##
##
##                  EQ5D_mobility                 EQ5D_self_care
##  None or slight problems:366   None or slight problems:371
##  Moderate problems      : 10   Moderate problems      :  8
##  Severe problems        :  4   Severe problems        :  1
##
##
##
##
##              EQ5D_usual_activities        EQ5D_pain_discomfort
##  None or slight problems:346       None or slight   :308
##  Moderate problems      : 24       Moderate         : 53
##  Severe problems        :  8       Severe or extreme: 19
##  NA's                   :  2
##
##
##
##       EQ5D_anxiety_depression    OMAS_37      Cluster
##  None or slight   :317        Min.   : 0.00   1:265
##  Moderate         : 48        1st Qu.: 0.00   2:115
##  Severe or extreme: 14        Median : 2.00
##  NA's             :  1        Mean   : 3.86
##                               3rd Qu.: 5.00
##                               Max.   :38.00
##                               NA's   :52
```

Repeat model fit

```
glm.fit1 <- glm(Cluster ~ work_status + IBD_disease_activity + biological +
                  education + EQ5D_mobility + EQ5D_self_care +
                  EQ5D_usual_activities + EQ5D_pain_discomfort +
                  EQ5D_anxiety_depression + OMAS_37,
                data = ext.IBD, family = binomial(link = "logit"))
summary(glm.fit1)
```

```
##
```

```
## Call:
## glm(formula = Cluster ~ work_status + IBD_disease_activity +
##     biological + education + EQ5D_mobility + EQ5D_self_care +
##     EQ5D_usual_activities + EQ5D_pain_discomfort + EQ5D_anxiety_depression +
##     OMAS_37, family = binomial(link = "logit"), data = ext.IBD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3498  -0.8407  -0.5133   1.0310   2.5528
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -13.04642  865.02786  -0.015  0.98797
## work_statusWorking                 0.27359    0.42196   0.648  0.51674
## IBD_disease_activityOver treshold -1.05752    0.34383  -3.076  0.00210 **
## biologicalYes                      1.25468    0.35635   3.521  0.00043 ***
## education.L                        0.69331    0.60325   1.149  0.25044
## education.Q                       -0.08186    0.46593  -0.176  0.86054
## education.C                        0.29252    0.30983   0.944  0.34511
## EQ5D_mobility.L                   -9.09548 1173.74382  -0.008  0.99382
## EQ5D_mobility.Q                   -4.96119  677.66187  -0.007  0.99416
## EQ5D_self_care.L                  -2.88648 2063.14900  -0.001  0.99888
## EQ5D_self_care.Q                  10.80275 1438.73311   0.008  0.99401
## EQ5D_usual_activities.L            1.10469    0.99299   1.112  0.26593
## EQ5D_usual_activities.Q            1.14269    0.87958   1.299  0.19390
## EQ5D_pain_discomfort.L             0.17053    0.62328   0.274  0.78439
## EQ5D_pain_discomfort.Q             0.95212    0.59317   1.605  0.10846
## EQ5D_anxiety_depression.L         -0.95882    0.79208  -1.210  0.22609
## EQ5D_anxiety_depression.Q         -0.15612    0.61893  -0.252  0.80085
## OMAS_37                           -0.02090    0.02667  -0.784  0.43326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 408.18  on 323  degrees of freedom
## Residual deviance: 350.43  on 306  degrees of freedom
##   (56 observations deleted due to missingness)
## AIC: 386.43
##
## Number of Fisher Scoring iterations: 15
```

Remove problematic variables

```
glm.fit2 <- glm(Cluster ~ work_status + IBD_disease_activity + biological +
                  education + EQ5D_usual_activities + EQ5D_pain_discomfort +
                  EQ5D_anxiety_depression + OMAS_37,
                data = ext.IBD, family = binomial(link = "logit"))
summary(glm.fit2)
```

```
##
## Call:
## glm(formula = Cluster ~ work_status + IBD_disease_activity +
##     biological + education + EQ5D_usual_activities + EQ5D_pain_discomfort +
##     EQ5D_anxiety_depression + OMAS_37, family = binomial(link = "logit"),
```

```
##     data = ext.IBD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3573  -0.8479  -0.5066   1.0271   2.5817
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -2.76356    0.75865  -3.643 0.000270 ***
## work_statusWorking                 0.31858    0.41705   0.764 0.444940
## IBD_disease_activityOver treshold -0.99906    0.33776  -2.958 0.003098 **
## biologicalYes                      1.23306    0.35395   3.484 0.000494 ***
## education.L                        0.67004    0.60158   1.114 0.265362
## education.Q                       -0.04202    0.46463  -0.090 0.927931
## education.C                        0.28493    0.30836   0.924 0.355478
## EQ5D_usual_activities.L            0.19728    0.85435   0.231 0.817384
## EQ5D_usual_activities.Q            0.72042    0.79525   0.906 0.364989
## EQ5D_pain_discomfort.L            -0.05624    0.62192  -0.090 0.927950
## EQ5D_pain_discomfort.Q             0.79855    0.56493   1.414 0.157499
## EQ5D_anxiety_depression.L         -0.93230    0.78007  -1.195 0.232030
## EQ5D_anxiety_depression.Q         -0.11346    0.60660  -0.187 0.851622
## OMAS_37                           -0.02462    0.02642  -0.932 0.351288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 408.18  on 323  degrees of freedom
## Residual deviance: 353.09  on 310  degrees of freedom
##   (56 observations deleted due to missingness)
## AIC: 381.09
##
## Number of Fisher Scoring iterations: 5
```

```
confint(glm.fit2)
```

```
## Waiting for profiling to be done...
##                                       2.5 %      97.5 %
## (Intercept)                       -4.3970727 -1.36951152
## work_statusWorking                -0.4805353  1.16749789
## IBD_disease_activityOver treshold -1.6839749 -0.35364678
## biologicalYes                      0.5647917  1.96091156
## education.L                       -0.4284666  2.03539598
## education.Q                       -1.0816679  0.81323262
## education.C                       -0.3092242  0.91460093
## EQ5D_usual_activities.L           -1.9772848  1.69291097
## EQ5D_usual_activities.Q           -0.8880950  2.45056308
## EQ5D_pain_discomfort.L            -1.4873252  1.06211373
## EQ5D_pain_discomfort.Q            -0.3127663  1.97549447
## EQ5D_anxiety_depression.L         -3.0272567  0.35677955
## EQ5D_anxiety_depression.Q         -1.5013240  1.03389228
## OMAS_37                           -0.0806025  0.02418014
```

View diagnostics

```r
library(car)
vif.glm <- vif(glm.fit2)
cat("Variance inflation factor:\n")
```

```
## Variance inflation factor:
```

```r
vif.glm
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## work_status              1.219427  1        1.104277
## IBD_disease_activity     1.164792  1        1.079255
## biological               1.061894  1        1.030482
## education                1.205325  3        1.031614
## EQ5D_usual_activities    1.310593  2        1.069959
## EQ5D_pain_discomfort     1.387117  2        1.085246
## EQ5D_anxiety_depression 1.169929  2        1.040016
## OMAS_37                  1.048192  1        1.023812
```

```r
lev <- hatvalues(glm.fit2)
avg.lev <- mean(lev)
high <- avg.lev*3
high.lev <- which(lev > high)
cat("\nLeverage three times greater than mean leverage:\n")
```

```
##
## Leverage three times greater than mean leverage:
```

```r
lev[high.lev]
```

```
##        44        56        77       118       132       153       170       185
## 0.2955610 0.1888001 0.1858884 0.1445420 0.2272857 0.1445420 0.1830015 0.1579274
##       202       205       214       227       256       261       268       294
## 0.1445420 0.2959983 0.2548185 0.1387734 0.1624059 0.1740727 0.1421859 0.1366284
##       300       305       316       327
## 0.1826156 0.1924093 0.1445420 0.1315507
```

```r
library(car)
n = nrow(ext.IBD)
cooks <- cooks.distance(glm.fit2)
critical <- which(cooks > 4/n)
cat("\nCook's distance above 4/N:\n")
```

```
##
## Cook's distance above 4/N:
```

```r
cooks[critical]
```

```
##         43         44         76         77         79         95        115
## 0.05207460 0.01222691 0.07546648 0.08608121 0.01077050 0.01063384 0.01698905
##        129        141        142        146        152        153        196
## 0.01406280 0.03664757 0.07037655 0.01057813 0.01443031 0.03580726 0.01266450
##        205        214        230        261        300
## 0.01263100 0.12322918 0.01478817 0.02795747 0.03796440
```

10-fold CV to assess model performance

```r
library(caret)
library(doParallel)
library(pROC)
```

```r
set.seed(5)
# Define labels for caret::train()
ext.IBD$Cluster <- factor(ext.IBD$Cluster, labels = c("Cluster_1", "Cluster_2"))
# Removing missing values for cross-validation
ext.na <- na.omit(ext.IBD)

cl <- makeCluster(detectCores()-1)
registerDoParallel(cl)
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 100,
                     classProbs = TRUE, summaryFunction = twoClassSummary,
                     savePredictions = TRUE)
glm.model <- train(Cluster ~ work_status + IBD_disease_activity + biological +
                     education + EQ5D_usual_activities + EQ5D_pain_discomfort +
                     EQ5D_anxiety_depression + OMAS_37, data = ext.na,
                   method = "glm", family = binomial(link = "logit"),
                   trControl = ctrl)
stopCluster(cl)

glm.model
```

```
## Generalized Linear Model
##
## 324 samples
##   8 predictor
##   2 classes: 'Cluster_1', 'Cluster_2'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 100 times)
## Summary of sample sizes: 291, 291, 291, 292, 292, 292, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.6990026  0.8378831  0.3560182
```

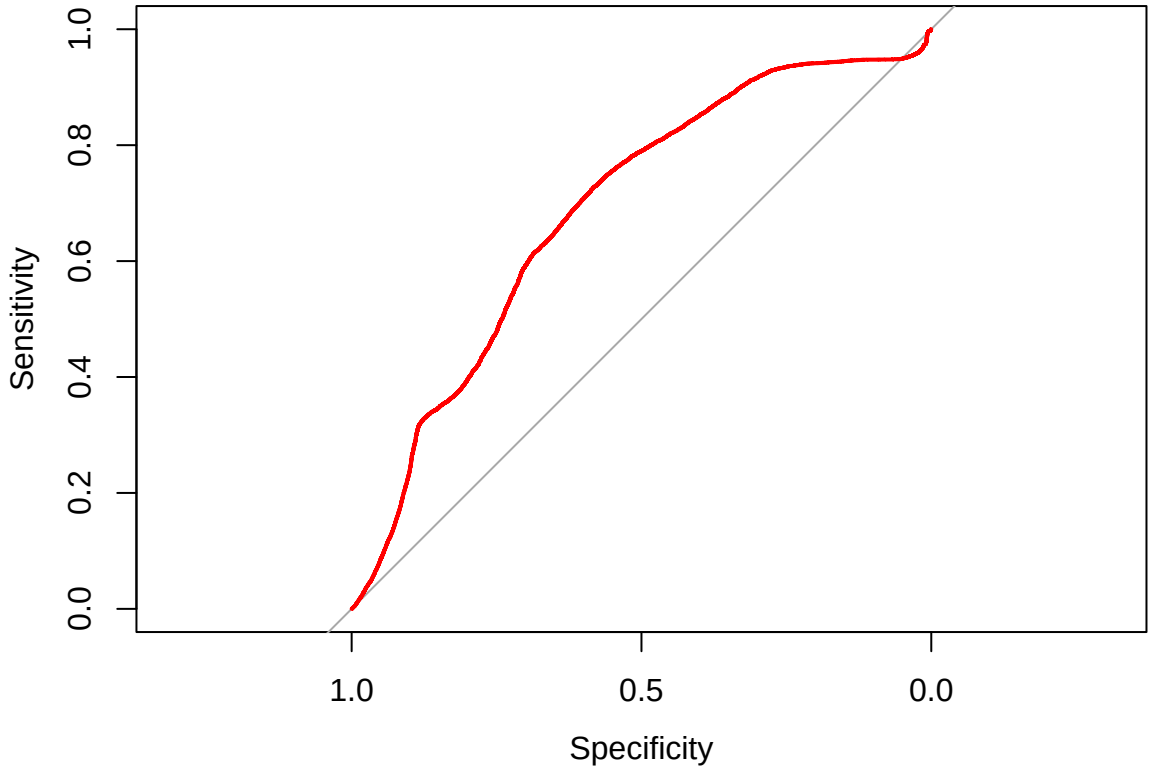
```r
cat("Standard deviation for ROC-AUC:\n", sd(glm.model$resample$ROC))
```

```
## Standard deviation for ROC-AUC:
##  0.09182288
```

```r
cat("\nStandard error for ROC-AUC:\n", sd(glm.model$resample$ROC) /
      sqrt(length(glm.model$resample$ROC)))
```

```
##
## Standard error for ROC-AUC:
##  0.002903694
```

```r
# plot
preds <- glm.model$pred
roc.curve <- roc(preds$obs, preds$Cluster_2)
plot(roc.curve, col = "red")
```



Permutation test of model

```
library(coin)
perm.glm <- independence_test(Cluster ~ gender + work_status +
                                IBD_disease_activity + biological +
                                EQ5D_mobility + EQ5D_pain_discomfort +
                                EQ5D_anxiety_depression, data = ext.IBD,
                              teststat = "maximum",
                              distribution = approximate(nresample = 10000))

pvalue(perm.glm)
```

```
## [1] 2e-04
## 99 percent confidence interval:
##  1.034992e-05 9.270420e-04
```