Multi-view Banded Spectral Clustering (mvBSC)

```
library(data.table)
source(file = "code/groupingEvaluationMetrics.R")
```

```
get_U <- function(cosK, h, k, R) {
  cosK_band <- cosK * (R <= h)
  eigen_cosK_band <- eigen(cosK_band)
  idx <- order(abs(eigen_cosK_band$values), decreasing = TRUE)
  U <- eigen_cosK_band$vectors[, idx[1:k]]
  rownames(U) <- rownames(cosK)
  return(U)
}

get_Z <- function(codes, labels) {
  uniqueLabels <- sort(unique(labels))
  Z <- 1 * outer(labels, uniqueLabels, "==")
  rownames(Z) <- codes
  colnames(Z) <- uniqueLabels
  return(Z)
}

mvbsc_fit <- function(codes, distance, similarity, ncluster, weights, delta, band, seed) {
  h <- band
  k <- ncluster
  R <- distance[codes, codes]
  W <- 0
  for (i in 1:length(weights)) {
    if (weights[i] > 0) {
      cosK <- similarity[[i]][codes, codes]
      tmp <- get_U(cosK = cosK, h = h, k = k, R = R)
      W <- W + weights[i] * tcrossprod(tmp)
    }
  }
  W_eigen <- eigen(W)
  idx <- order(abs(W_eigen$values), decreasing = TRUE)
  U <- W_eigen$vectors[, idx[1:k]]
  rownames(U) <- rownames(W)
  set.seed(seed)
  fit <- kmeans(U, k, iter.max = 100, nstart = 25)
  tbl <- data.frame(cluster = 1:k, size = NA, max_dist = NA)
  for (i in 1:k) {
    v <- names(fit$cluster)[fit$cluster == i]
    tbl$size[i] <- length(v)
    tbl$max_dist[i] <- max(R[v, v])
  }
  tbl2 <- tbl[order(tbl$max_dist), ]
  rownames(tbl2) <- 1:k
  list(cluster = fit$cluster, cluster_info = tbl2)
}
```

```
h1.cat <- "I00-I25" # focused on ICD category I00-I25
va_cosK <- readRDS(paste0("data/va_", h1.cat, "_cosineMat.rds"))
```

```
bio_cosK <- readRDS(paste0("data/biobank_", h1.cat, "_cosineMat.rds"))
this.R <- readRDS(paste0("data/", h1.cat, "_distR_wt_avg_1.rds"))
icd.info <- readRDS("data/rollable_new_icd_info_20190130.rds")
codes_in_use <- colnames(bio_cosK)
codes_with_phecode <- codes_in_use[!is.na(icd.info[codes_in_use]$PheCode)]
pheCodes <- icd.info[codes_with_phecode]$PheCode
Z0 <- get_Z(codes_with_phecode, pheCodes)
# delta0 <- min(apply(this.R, 1, max)) / 2
delta0 <- 5
delta0
```

```
## [1] 5
```

```
initial <- mvbsc_fit(
  codes = rownames(this.R),
  distance =  this.R,
  similarity = list(va_cosK, bio_cosK),
  ncluster = 30,
  weights = c(va.wt = 0.5, bio.wt = 0.5),
  delta = delta0,
  band = delta0 / 2,
  seed = 123)
initial
```

```
## $cluster
##   394.9   395.9   396.0   396.1   396.2   396.3   396.8   396.9   397.0
##      28      17      25      25      25      25      25      25      17
##   397.1   397.9  398.99   401.0   401.1   401.9  402.00  402.01  402.10
##       4       4       4      30      30      30      21      21      21
##  402.11  402.90  402.91  403.00  403.01  403.10  403.11  403.90  403.91
##      21      21      21      11      10      14      14      14      10
##  404.00  404.01  404.02  404.03  404.10  404.11  404.12  404.13  404.90
##       9      11       9       9      11       9      11      11      11
##  404.91  404.92  404.93  405.01  405.09  405.11  405.19  405.91  405.99
##      11       9       9      29      29      29      30      29      29
##  410.00  410.01  410.02  410.10  410.11  410.12  410.20  410.21  410.22
##       5       5       5       5       5       5      16      16      16
##  410.30  410.31  410.32  410.40  410.41  410.42  410.50  410.51  410.52
##      16      16      16      16      16      16       5       5      16
##  410.60  410.61  410.62  410.70  410.71  410.72  410.80  410.81  410.82
##      16      16      16      30       3       3       5      16       3
##  410.90  410.91  410.92   411.1  411.89   413.0   413.9  414.00  414.01
##      16      16      16      19      23      19      19      23      23
##  414.02  414.03  414.04  414.05  414.06  414.07  414.10  414.19   414.8
##      26      26      26      26      12      12      26      26      26
##   414.9   429.2  429.79    I00.   I01.0   I01.1   I01.2   I01.8   I01.9
##      12      23      30      30      13      13      13      13      13
##   I02.0   I02.9   I05.0   I05.1   I05.2   I05.8   I05.9   I06.0   I06.1
##      30      30      28      28      28      28      28       1       1
##   I06.2   I06.8   I06.9   I07.1   I07.2   I07.8   I07.9   I08.0   I08.1
##       1      30      30      30      30      27      27      27      27
##   I08.2   I08.3   I08.8   I08.9   I09.1   I09.2  I09.81  I09.89   I09.9
##      27      27      27      30       4      30       4       4       4
##    I10.   I11.0   I11.9   I12.0   I12.9   I13.0  I13.10  I13.11   I13.2
##      30      30      30      30      18      18      18      18      18
```

```
##   I15.0   I15.1   I15.2   I15.8   I15.9   I20.0   I20.1   I20.8   I20.9
##      24      24      24      24      24       2      30       2       2
##  I21.01  I21.02  I21.09  I21.11  I21.19  I21.21  I21.29   I21.3   I21.4
##       8       8       8       8       8       8       8       8       8
##   I22.0   I22.1   I22.2   I22.8   I22.9   I23.1   I23.2   I23.6   I23.7
##      20      20      20      20      20       7       7       7       7
##   I23.8   I24.0   I24.1   I24.8   I24.9  I25.10 I25.110 I25.111 I25.118
##       7      23      23      15      15      15      15      15      15
## I25.119   I25.2   I25.3  I25.41  I25.42   I25.5   I25.6 I25.700 I25.701
##      15       6      12      15      15      15      12       6       6
## I25.708 I25.709 I25.710 I25.711 I25.718 I25.719 I25.720 I25.721 I25.728
##       6      30      30      22       6      22      30      30      22
## I25.729 I25.739 I25.750 I25.758 I25.759 I25.760 I25.769 I25.790 I25.791
##      30      22      30      22      22      30      22      30      22
## I25.798 I25.799 I25.810 I25.811 I25.812  I25.82  I25.83  I25.84  I25.89
##      22      22       6      30      30       6       6       6       6
##   I25.9
##       6
##
## $cluster_info
##    cluster size max_dist
## 1       10    2     0.00
## 2       22   10     0.08
## 3        3    3     0.31
## 4       16   17     0.39
## 5        1    3     0.40
## 6        5    9     0.40
## 7        9    6     0.40
## 8        8    9     0.79
## 9       26    7     1.01
## 10      12    5     1.20
## 11       6   11     1.40
## 12       7    5     1.40
## 13       4    7     1.60
## 14      11    7     1.60
## 15      18    5     1.60
## 16      29    5     1.60
## 17      19    3     1.70
## 18       2    3     1.80
## 19      13    5     1.80
## 20      14    3     1.80
## 21      20    5     1.80
## 22      21    6     1.80
## 23      24    5     1.80
## 24      25    6     1.80
## 25      28    6     1.80
## 26      17    2     2.03
## 27      15   10     2.40
## 28      27    7     3.00
## 29      23    6     3.20
## 30      30   30    96.61
```

```
cluster0 <- subset(initial$cluster_info, max_dist > delta0)$cluster
cluster0
```

```
## [1] 30
```

```
regroup <- vector("list", length(cluster0))
names(regroup) <- paste0("initial_", cluster0)
```

```
for (i in 1:length(cluster0)) {
  tmp <- names(initial$cluster[initial$cluster == cluster0[i]])
  for (k in 2:(length(tmp) - 1)) {
    try <- mvbsc_fit(
      codes = tmp,
      distance =  this.R,
      similarity = list(va_cosK, bio_cosK),
      ncluster = k,
      weights = c(va.wt = 0.5, bio.wt = 0.5),
      delta = delta0,
      band = delta0 / 2,
      seed = 123)
    if (all(try$cluster_info$max_dist <= delta0)) break
  }
  regroup[[i]] <- try
}
regroup
```

```
## $initial_30
## $initial_30$cluster
##   401.0   401.1   401.9  405.19  410.70  429.79    I00.   I02.0   I02.9
##      18       6      18      17       5      13       4       1       1
##   I06.8   I06.9   I07.1   I07.2   I08.9   I09.2    I10.   I11.0   I11.9
##       7       7       3       7      16       9      10      15      15
##   I12.0   I20.1 I25.709 I25.710 I25.720 I25.721 I25.729 I25.750 I25.760
##       8      11       2       2       2       2       2       2      14
## I25.790 I25.811 I25.812
##       2      14      12
##
## $initial_30$cluster_info
##    cluster size max_dist
## 1        3    1     0.00
## 2        4    1     0.00
## 3        5    1     0.00
## 4        6    1     0.00
## 5        8    1     0.00
## 6        9    1     0.00
## 7       10    1     0.00
## 8       11    1     0.00
## 9       12    1     0.00
## 10      13    1     0.00
## 11      16    1     0.00
## 12      17    1     0.00
## 13      18    2     0.00
## 14       2    7     0.09
## 15      14    2     0.15
## 16       1    2     1.80
## 17       7    3     1.80
## 18      15    2     1.80
```