Manual

Manual 1

Install 2

Example data 2

Functions 2

encore_step1: subspace separation and subspace distance calculation 2

encore_step2: consensus clustering 3

encore_big: extend to big data by an supervised methods 5

find_markers: Find group markers 6

Demos 7

mode 1: run subspace separation in kmeans mode with fiexed k and run cluster in subspaces in dbscan mode 7

mode 2: run subspace separation in kmeans mode with fiexed k and run cluster in subspaces in spect mode 10

mode 3: run subspace separation in 'both' mode (try GMM firstly and then kmeans) with fiexed k and run cluster in subspaces in 'both' mode (used in the study) 12

mode 4: run big data 13

Find group markers 15

Visualization density profiles of features in each subspace 15

Suggestions 17

Introduction

Single-cell RNA sequencing enables us to characterize the cellular heterogeneity in single cell resolution with the help of cell type identification algorithm. However, the noise inherent in single-cell RNA-sequencing data severely disturbs the accuracy of cell clustering, marker identification and visualization. We propose that clustering based on feature density profiles can distinguish informative features from noise. We named such strategy as “entropy subspace” separation and designed a cell clustering algorithm called ENtropy subspace separation-based Clustering for nOise REduction (ENCORE) by integrating the “entropy subspace” separation strategy with a consensus clustering method.

We have provided an example data in “Tian_297.txt” and its validated cell type results in “label.txt” at https://github.com/SONG0417/ENCORE_V1.0.git, which might be helpful for users to go through this manual.

Install

#Download the source code from github, and install as fellow

install.packages("the_directory_that_contain_the_package/ENCORE_0.1.0.tar.gz",repos=NULL,type="source")

Example data

We have integrated an example data in the directory of the source codes. Tian_297.txt gave the input expression matrix and labels.txt gave the experimental validated cell types.

Functions

encore_step1: subspace separation and subspace distance calculation

Description:

Using an expression matrix as input, encore_step1 will scale data, separate features into different subspaces, calculate and visualize distance matrices in subspaces.

Usage: encore_step1(

```
data,
method = "kmeans",
start = 50,
end = 0,
nstart = 10,
dens = 512,
thread = 1,
sample = "FALSE",
scale = "TRUE",
sample_size = 10000,
perp = 0,
minPts = 0
```

)

Examples:

```
result1=encore_step1(data,thread=5)
result1=encore_step1(data,method="kmeans",start=10,end=50, thread=20)
result1=encore_step1(data,dis_method="dis",thread=20)
```

encore_step2: consensus clustering

Description:

Using the expression matrix (Not required in sample mode) and the result from encore_step1 as input, encore_step2 will integrate distance information from low-entropy subspace and finish consensus clustering.

Usage: encore_step2(

```
data = NA,
ac = NA,
k = NA,
adjust = "TRUE",
method = "spect",
result = NA,
thread = 4,
```

```
    sample = "FALSE",
    scale = "TRUE",
    minPts = 0,
    perp = 0,
    pc = 1e-04
)
```

Examples:

```
result1=encore_step1(data,method="kmeans",start=50)
result2=encore_step2(data=data,method="both",k=5,result=result1,thread=4)
result2=encore_step2(data=data,method="dbscan",result=result1,thread=4)
```

encore_big: extend to big data by an supervised methods

Description:

Using the expression matrix and the results from encore_step1 and encore_step2 as input, encore_big will extend the cluster results of sampled cells to rest cells using a supervised method.

Usage: encore_big(

```
        data = NA,
        result = NA,
        result1 = NA,
        scale = "TRUE",
        thread = 1,
        perp = 0
    )
```

Examples:

result1=encore_step1(data,dens=2000,thread=2,sample="TRUE")

result2=encore_step2(result=result1,k=c(38,39,40,41,42),sample="TRUE",thread=4)

result=encore_big(data=data,result=result2,result1=result1,thread=2)

find_markers: Find group markers

Description:

Using an expression matrix and cluster result as input, find_markers will output cell group markers.

Usage: find_markers(

data,

temp,

top = 20,

thread = 1,

clus = "all",

pc = 1e-04

heatmap = "FALSE")

Examples:

result1=encore_step1(data,method="kmeans",start=50,thread=4)

result2=encore_step2(data=data,method="both",k=5,result=result1,thread=4)

markers=find_markers(data, temp=result2$temp, thread=10)

Demos

Under the ENCORE package home directory, we put expression matrix of Tian_297, which can be used for testing.

mode 1: run subspace separation in kmeans mode with fiexed k and run cluster in subspaces in dbscan mode

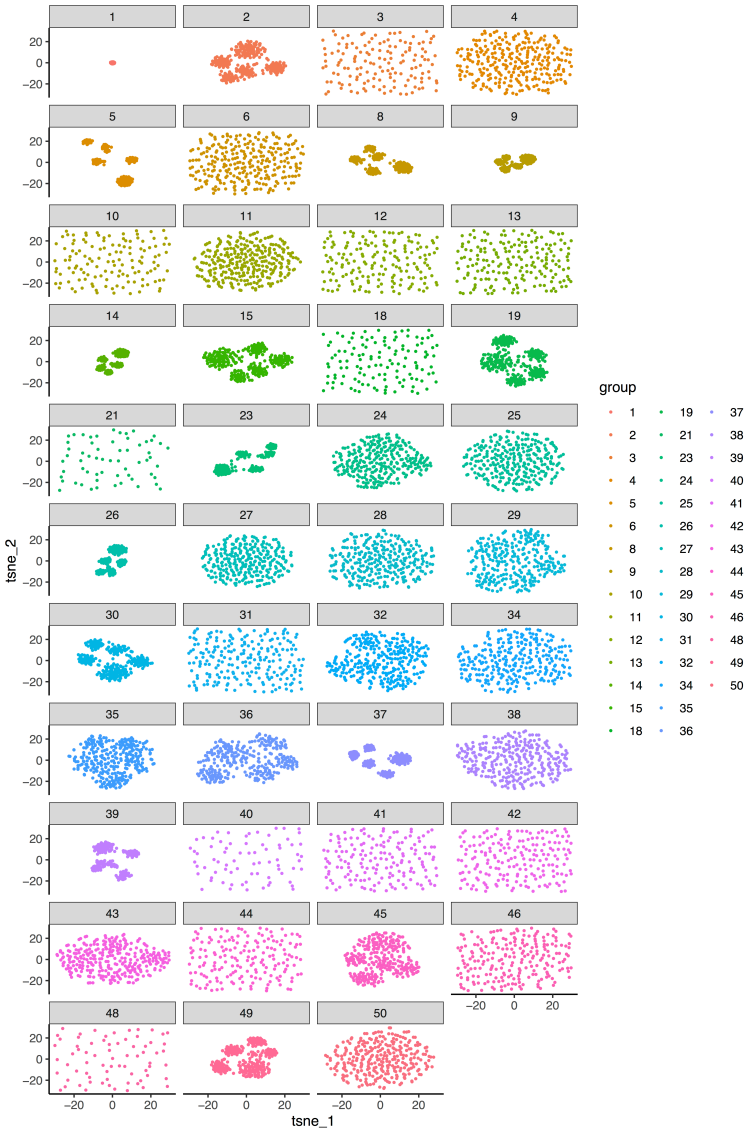
R code:

```
library("ENCORE")
#Read the data
data=read.table("Tian_297.txt")
#1 run subspace separation and distance calculation in subspaces
result1=encore_step1(data,thread=4)
#2 run consensus clustering
result2=encore_step2(data=data,result=result1,method="dbscan",thread=4)
# read the real labels
labels=read.table("labels.txt")
#3 check the result accuracy
adjustedRandIndex(result2$cluster,labels[,1])
#4 visualization cluster results
ggplot(data=result2$temp,aes(x=tsne_1,y=tsne_2,color=clusters))+geom_point()
```

Output results of #1

```
[1] "***The subspace ordered by entropies are:"
 [1]  5 14 37  9 26  8 39 23 30 49  2 15 19 36 45 32 35 29 43 24 34 38 31 46 13
[26] 48 25 40 10 21  1  3  4  6  7 11 12 16 17 18 20 22 27 28 33 41 42 44 47 50
[1] "***The suggest k for clustering in subspaces in spectral mode is 5.5"
[1] "***The t-SNE plots in subspaces can be shown by type result$tsnes"
```

tsnes.pdf or result1$tsnes



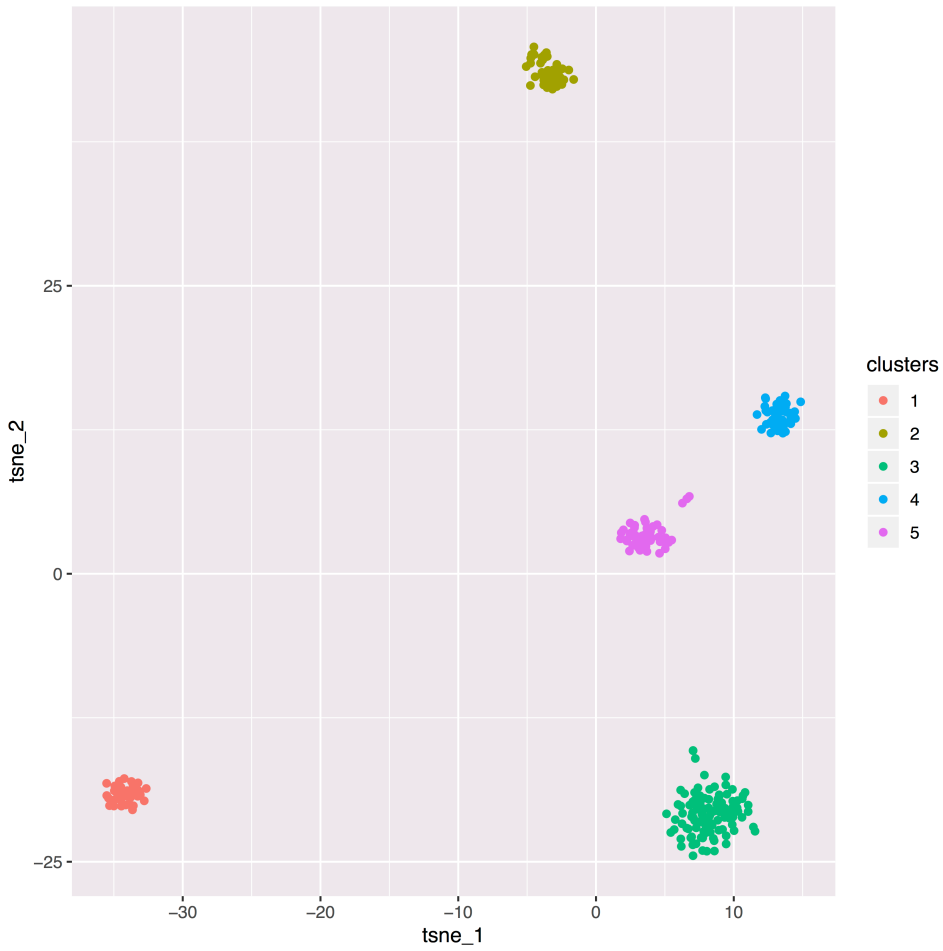
Output of #2

```
> result2=encore_step2(data=data,result=result1,method="dbscan",thread=4)
[1] "***The subspaces used in these step"
[1]  5 14 37  9
[1] "run cluster in subspace in DBSCAN mode"
[1] "Choose best subspaces"
[1]  5 14
[1]  5 14 37
[1]  5 14 37  9
[1] "The largest IAUC is: 0.868624626609928"
[1] "The IAUC score: && the factors:"
[1] "0.868624626609928 1.6"
[1] "0.867060316336302 1.8"
[1] "0.866350839586496 1.6"
[1] "***The best subspaces is c(5, 14, 37)"
```

Output of #3

```
> adjustedRandIndex(result2$cluster,labels[,1])
 [1] 0.9643056
```

output of #4



mode 2: run subspace separation in kmeans mode with fiexed k and run cluster in subspaces in spect mode

R code:

result1=encore_step1(data, thread=4)

run consensus clustering

try1:

result2=encore_step2(data=data, result=result1,method="spect",thread=4)

If you do not assign a value to k, k will automatically use the value inferred by step1

```
> result2=encore_step2(data=data,result=result1,method="spect",thread=4)
[1] "***The subspaces used in these step"
[1]  5 14 37  9
[1] "run cluster in subspace in spectral cluster mode"
[1] "***Used ks in this study:"
[1] 5 6 7
[1] "Start learn distance"
[1] "***spectral cluster, k="
[1] 5 6 7
[1] "***IAUC scores of ks="
[1] 0.8462034 0.7993453 0.7415670
[1] "***The best methods is spectral and k is 5"
[1] "Choose best subspaces"
[1]  5 14
[1]  5 14 37
[1]  5 14 37  9
[1] "The largest IAUC is: 0.846660777463153"
[1] "The IAUC score: && the factors:"
[1] "0.846203378977512 3"
[1] "0.84602591110591 3"
[1] "0.846660777463153 3"
[1] "***The best subspaces is c(5, 14, 37, 9)"
```

try2:

result2=encore_step2(data=data, result=result1,method="spect",k=c(3,5,7), thread=4)

If multiple k values are provided, the best k will choose among these values

```
> result2=encore_step2(data=data, result=result1,method="spect",k=c(3,5,7), thread=4)
[1] "***The subspaces used in these step"
[1]  5 14 37  9
[1] "run cluster in subspace in spectral cluster mode"
[1] "***Used ks in this study:"
[1] 3 5 7
[1] "Start learn distance"
[1] "***spectral cluster, k="
[1] 3 5 7
[1] "***IAUC scores of ks="
[1] 0.8627231 0.8588987 0.7521026
[1] "***The best methods is spectral and k is 3"
[1] "Choose best subspaces"
[1]  5 14
[1]  5 14 37
[1]  5 14 37  9
[1] "The largest IAUC is: 0.86422989017347"
[1] "The IAUC score: && the factors:"
[1] "0.862723071961086 2.8"
[1] "0.86422989017347 3"
[1] "0.86422989017347 3"
[1] "***The best subspaces is c(5, 14, 37)"
```

try3:

result2=encore_step2(data=data, result=result1,method="spect",k=c(3), adjust= "FALSE", thread=4)

If a single value of k is provided and adjust= "FALSE", this k will be directly used for the clustering in subspaces

```
> result2=encore_step2(data=data, result=result1,method="spect",k=c(3), adjust= "FALSE", thread=4)
[1] "***The subspaces used in these step"
[1]  5 14 37  9
[1] "run cluster in subspace in spectral cluster mode"
[1] "***Used ks in this study:"
[1] 3
[1] "Start learn distance"
[1] "***spectral cluster, k="
[1] 3
[1] "***IAUC scores of ks="
[1] 0.8278007
[1] "***The best methods is spectral and k is 3"
[1] "Choose best subspaces"
[1]  5 14
[1]  5 14 37
[1]  5 14 37  9
[1] "The largest IAUC is: 0.827963101895949"
[1] "The IAUC score: && the factors:"
[1] "0.827800722835602 1.6"
[1] "0.827963101895949 1.8"
[1] "0.827963101895949 1.8"
[1] "***The best subspaces is c(5, 14, 37)"
```

mode 3: run subspace separation in ‘both’ mode (try GMM firstly and then kmeans) with fiexed k and run cluster in subspaces in ‘both’ mode (used in the study)

result1=encore_step1(data, method="both", thread=4)

```
> result1=encore_step1(data, method="both", thread=4)
[1] "running subspace separation"
fitting ...
  |                                                                      |   0%
  |======================================================================| 100%

   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
 387  584  390 3509  172  526  457  837 1056  236  368  322  461  349  420  334
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
 243  947  222  350  315  507  366  538  436  647  400  675  635  489 1440  554
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
 434  409  652  595  101  528  467  366  530  450  587  426  345  644  692  296
  49   50
 170 1542
[1] "distance caculation in subspaces"
[1] "***The subspace ordered by entropies are:"
 [1] 17 48  1 49  5 28 15 46 29 47 25  9 26 22  8 35 34 12 16 19 43 32 42 36 30
[26] 21  7 27 31  2  3  4  6 10 11 13 14 18 20 23 24 33 37 38 39 40 41 44 45 50
[1] "***The suggest k for clustering in subspaces in spectral mode is 5"
[1] "***The t-SNE plots in subspaces can be shown by type result$tsnes"
```

result2=encore_step2(data=data,result=result1,method="both",k=5,thread=4)

```
> result2=encore_step2(data=data,result=result1,method="both",k=5,thread=4
[1] "***The subspaces used in these step"
[1]  5 14 37  9
[1] "***the best method is dbscan"
[1] "Choose best subspaces"
[1]  5 14
[1]  5 14 37
[1]  5 14 37  9
[1] "The largest IAUC is: 0.854145555111113"
[1] "The IAUC score: && the factors:"
[1] "0.851437946422589 2.2"
[1] "0.851367248113571 2.2"
[1] "0.854145555111113 2.4"
[1] "***The best subspaces is c(5, 14, 37, 9)"
```

mode 4: run big data

sample sample_size cells and run subspace separation and distance calculation based on the sampled data

result1=encore_step1(data,dens=2000,thread=2,sample="TRUE",sample_size=100)

default sample_size=10000; Here, sample_size=100 is just for demonstration purposes

```
> result1=encore_step1(data,dens=2000,thread=2,sample="TRUE",sample_size=100)
[1] "Sample 100 cells from expression matrix"
[1] "running subspace separation"
[1] "running separation in kmeans mode with k= 50"

   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
 286  504  507  333  228  510  371  361   19  531 2158  604  259  293 4228  234
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
 334  512 3510  173  430  563  142  490  623  425  478  462  148  353  750  502
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
 442  409  113  251  387  381  365  146  404  539  455  738  385  370  669  415
  49   50
 299  317
[1] "distance caculation in subspaces"
[1] "***The subspace ordered by entropies are:"
 [1] 19 23  7 29 22 35  2 24 42 18 38 36 32 44 37 50 20 48 47 12 33  6 41 45 16
[26] 28  3 10 39  4 14 43  5 34  8 31 13 49 30 25 21 46  1 27 17 11  9 15 26 40
[1] "***The suggest k for clustering in subspaces in spectral mode is 3"
[1] "***The t-SNE plots in subspaces can be shown by type result$tsnes"
```

consensus clustering for the sampled data and find out the best k and best c (subspaces)

result2=encore_step2(result=result1, ac=c(23,42), k=5, sample="TRUE", thread=4)

For large data sets, it is recommended to select the appropriate subspace based on the tsnes plot rather than the inferred low-entropy subspaces from step1

```
> result2=encore_step2(result=result1, ac=c(23,42), k=5, sample="TRUE", thread=4)
[1] "***The subspaces used in these step"
[1] 23 42
[1] "run cluster in subspace in spectral cluster mode"
[1] "***Used ks in this study:"
[1] 4 5 6
[1] "Start learn distance"
[1] "***spectral cluster, k="
[1] 4 5 6
[1] "***IAUC scores of ks="
[1] 0.3229472 0.2460210 0.1830451
[1] "***The best methods is spectral and k is 4"
[1] "Choose best subspaces"
[1] 23 42
[1] "***The best subspaces is c(23, 42)"
```

predict the group labels of rest cells

result=encore_big(data=data, result=result2,result1=result1,thread=2)

```
> result=encore_big(data=data,result=result2,result1=result1,thread=2)
[1] "Run tune svm"
[1] "The error of SVM:"
[1] 0.15
[1] "Calculate distance on whole matrix"
[1] "Finish distance calculation"
[1] "complete no. 0"
[1] "Deal with 120 cells"
[1] "Got features in new dimension"
[1] "complete no. 1"
[1] "Deal with 120 cells"
[1] "Got features in new dimension"
[1] "complete no. 2"
[1] "Deal with 120 cells"
[1] "Got features in new dimension"
[1] "complete no. 3"
[1] "Deal with 120 cells"
[1] "Got features in new dimension"
[1] "complete no. 4"
[1] "Deal with 117 cells"
[1] "Got features in new dimension"
[1] "Finish cell clustering and the cluster results are save in cluster.txt"
```

visualization cluster results

ggplot(data=result,aes(x=tsne_1,y=tsne_2,color=clusters))+geom_point()

Find group markers

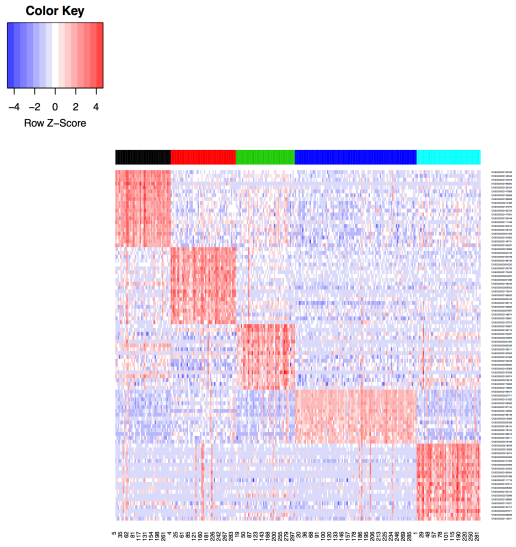
markers=find_markers(data, temp=result2$temp, thread=10)

```
> head(markers)
                             p      iauc            gene status cluster
ENSG00000132432  7.962883e-93 1.0000000 ENSG00000132432     up       1
ENSG00000146648  1.318775e-67 0.9996032 ENSG00000146648     up       1
ENSG00000154978  9.479764e-58 0.9992063 ENSG00000154978     up       1
ENSG00000196260 2.735948e-114 0.9962963 ENSG00000196260     up       1
ENSG00000132434  4.300075e-63 0.9947090 ENSG00000132434     up       1
ENSG00000147889  1.407053e-55 0.9940476 ENSG00000147889     up       1
```

Each column represents the adjusted p value, resolution, gene name, up/down expressed, cell group, respectively

The feature plots of markers will be generated automatically, stored in the working directory, named as feature_group1.pdf, feature_group2.pdf, …, feature_groupn.pdf.

Heatmap plots of markers will be plot when heatmap="TRUE"

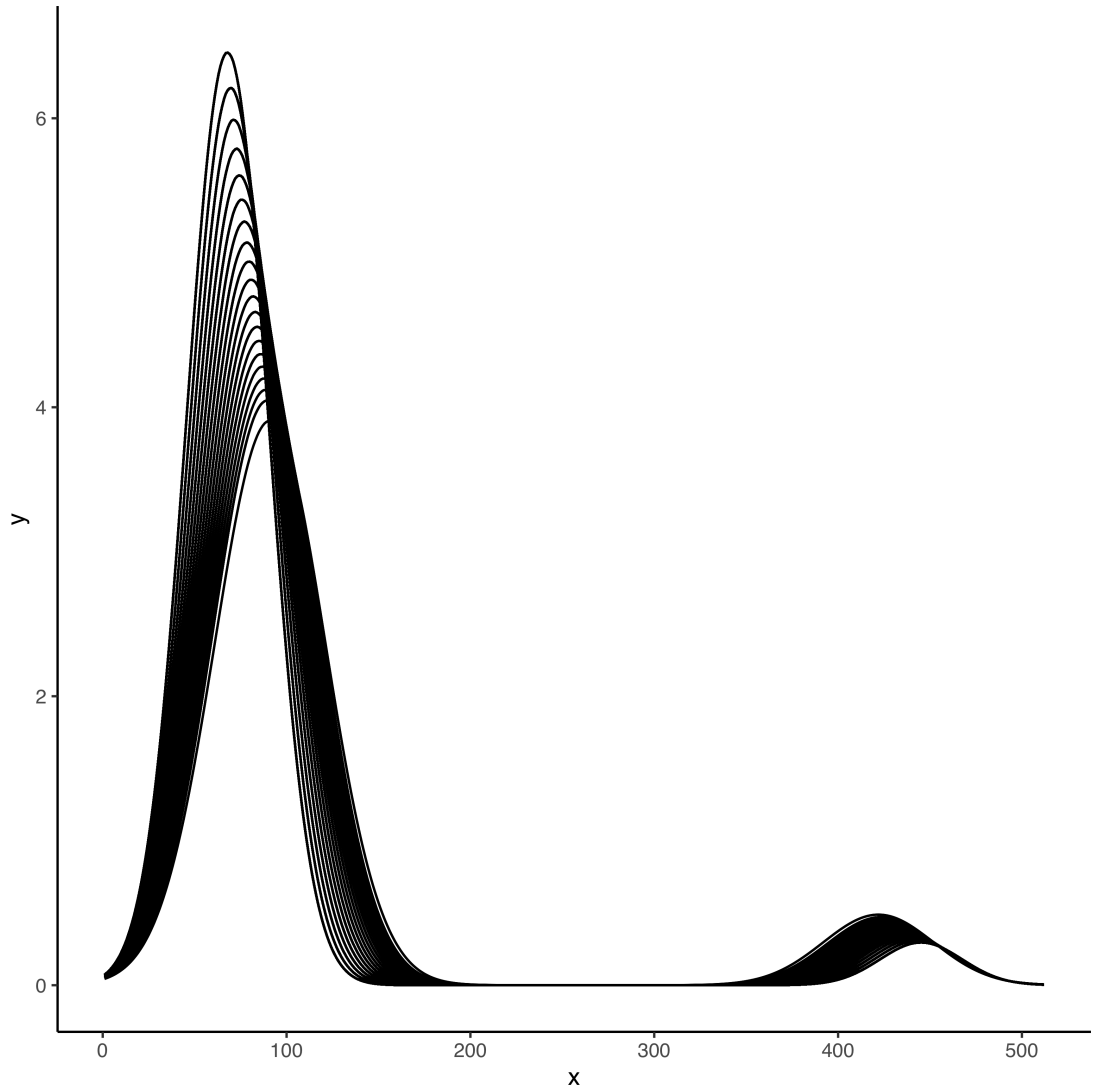


Visualization density profiles of features in each subspace

```
data=read.table("Tian_297.txt",header=T,row.names=1)
max_data=max(data)
if(max_data>=10000) data=log(data+1)/log(10) else data=log(data+1)/log(2)
if(dim(data[1])<10000) den=512 else den=2000
dy=list()
for(i in 1:dim(data)[2]){
```

```
 a=density(as.numeric(data[,i]),n=den)
 dy[[i]]=a$y
}
dy=as.data.frame(dy)
name=names(data)
names(dy)=name
library("reshape2")
yy=melt(dy)
#cluster=read.table("gcluster.txt")
#here, result1 is the result return from encore_step1
cluster=result1$gcluster
group=rep(cluster[,1],each=den)
densitys=data.frame(x=rep(1:den,dim(yy)[1]/den),y=yy$value,gene=yy$variable,group=grou
 p)
library("ggplot2")
# obtained density profiles of each subspaces by modifying the number of "1"
d=densitys[densitys$group=="1",]
ggplot(d,aes(x=x,y=y,group=gene))+geom_line()+theme_bw()+theme(panel.border=element_bl
 ank(),panel.grid.major=element_blank(),panel.grid.minor=element_blank(),axis.line=ele
 ment_line(color="black"))
```

Example:



Suggestions

- ✓ encore_step1: We recommend using method="kmeans" in subspace separation firstly, because it is much faster and work well on most data. Using method="both" if a bad result is gotten in method="kmeans".
- ✓ encore_step2:

  method: The tsnes plot given by encore_step1 can be used to determine which method to use in encore_step2: "dbscan" is recommended when the cells has a clear clustering bound on the low entropy subspace; "spect" is used in other cases; If you have clear prior information about k, you can get better results by using "both".

  k: When running "spect" and "both" modes, you need to provide k, and it is better to provide a more accurate k for "both" mode. For "spect" mode, if k is't provided, the k that inferred by step1, with its mutant (k-1,k+1) will be used and evaluated at this step. It is suggested user provide more k to get more accurate results

  ac: If this parameter is not assigned, it will automatically extract four subspaces with lowest entropy from the result of step1. In fact, this value can be better assigned through the tsnes plots from step1, because the tsnes plots show which space has better clustering information intuitively.