

# Package ‘SpectralGEM’

July 26, 2009

**Version** 1.0

**Date** 2009-07-07

**Title** Discovering Genetic Ancestry Using Spectral Graph Theory

**Author** Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, Kathryn Roeder

**Maintainer** Jing Wu <jwu@stat.cmu.edu>

**Depends** survival, optmatch, R (>= 2.0)

**Description** Discovering Genetic Ancestry Using Spectral Graph Theory

**License** GPL-2

**Repository** CRAN

**Date/Publication** 2009-07-26 18:32:45

## R topics documented:

full_matching . . . . .	2
getExcludeFile . . . . .	2
getRecord . . . . .	3
getVersion . . . . .	3
H . . . . .	3
id.info . . . . .	4
InFile . . . . .	4
loadInputFile . . . . .	5
MMfile . . . . .	6
pc_graphs_GEMp . . . . .	7
pc_graphs_GEMpClusters . . . . .	8
plotLam . . . . .	9
SpectralGEM . . . . .	9
updateExcludeFileDstr . . . . .	11
updateInputFile . . . . .	12

<b>Index</b>	<b>13</b>
--------------	-----------

---

full_matching	<i>Matching case and control</i>
---------------	----------------------------------

---

**Description**

The function matches case and control based on distances as measured by the significant eigenvectors. Internal function.

**Usage**

```
full_matching(ext)
```

**Arguments**

ext	the file extension of the distance matrix
-----	---

**Details**

The function calls the full matching program in optmatch library.

**Value**

ma	A matrix that consists of sampleId, matched strata, and case/control status.
----	--

---

getExcludeFile	<i>getExcludeFile</i>
----------------	-----------------------

---

**Description**

Internal function

**Usage**

```
getExcludeFile(InputFile)
```

**Arguments**

InputFile	Input file name
-----------	-----------------

---

getRecord	<i>getRecord</i>
-----------	------------------

---

**Description**

Internal function

**Usage**

```
getRecord(ext)
```

**Arguments**

ext	file extension
-----	----------------

---

getVersion	<i>getVersion</i>
------------	-------------------

---

**Description**

Internal function.

**Usage**

```
getVersion(InputFile)
```

**Arguments**

InputFile	Input file name.
-----------	------------------

---

H	<i>Matrix example</i>
---	-----------------------

---

**Description**

A matrix that can be used to match individuals into homogeneous strata.

**Usage**

```
data(H)
```

**Format**

The format is: num [1:345, 1:345] 2.941 0.048 -0.047 0.023 0.055 ... - attr(\*, "dimnames")=List of 2 .. chr[1 : 345] "1" "2" "3" "4" ..... : chr [1:345] "V5" "V6" "V7" "V8" ...

---

 id.info

*Sample information matrix*


---

### Description

A four-column matrix.

### Usage

```
data(id.info)
```

### Format

A data frame with 345 observations on the following 4 variables.

**sampleId** a numeric or string vector

**sex** a numeric vector

**case.1/control.2** a numeric vector

**groupId** a numeric vector

---

 InFile

*Generate a text file*


---

### Description

The function generates a text file that contains the parameters needed in the main fortran program and saves it in the current directory.

### Usage

```
InFile(identifier="smal", stage="1", directory=".",
        MMfile="MMprime.txt", excludefile="exclude.txt",
        idlength=8, mincluster=10, logtype=0,
        mdim=-1, msnp=-1,
        outfile="matching_input.txt")
```

### Arguments

identifier	Must be 4-letters long.
stage	Must be 1-letter long.
directory	The directory where the files will be generated.
MMfile	The MMprime matrix file name.
excludefile	The name of the file where the outliers have been or will be written.

idlength	The number of letters the longest sample name has.
mincluster	The smallest cluster required when do clustering, must be at least 5.
logtype	The type of log information output to the log file: 0=limited, 2=lots.
mdim	The maximum number of eigenvalues to print to the output file. The default is set at -1, which prints out every eigenvalues.
msnp	The number of SNPs to be used in determine the threshold for the significant eigenvalues. The default is set at -1, which the program estimates the value.
outfile	The name of the output file.

### Details

The function generates a text file with one column which contains all the parameters needed. The parameters need to be in the right order in order for the main fortran program to use.

### Value

A text file is saved in the current directory.

### Author(s)

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

### References

<http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/GEM+.htm>

### See Also

[http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/matching\\_input.txt](http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/matching_input.txt)

---

loadInputFile	<i>Load the input file</i>
---------------	----------------------------

---

### Description

Internal function.

### Usage

```
loadInputFile (InputFile)
```

### Arguments

InputFile     file name

MMfile

*Produce a .txt file that contains the MM prime matrix.***Description**

The function produces a .txt file of the proper format that contains the input data and the adjacency matrix. Because a common choice for the latter is  $MM'$ , this is called the MM prime matrix by default. This file is loaded to the main fortran program to do eigenvalue decomposition. Alternatively, the output file can be created from an input file containing M by the fortran program located at [http://wpicr.wpic.pitt.edu/WPICCompGen/MMp/MMp\\_page.htm](http://wpicr.wpic.pitt.edu/WPICCompGen/MMp/MMp_page.htm).

**Usage**

```
MMfile(H = H, sampleInfo = id.info, n=dim(H) [1], ntag=ntag, outfile = "MMprime.txt")
```

**Arguments**

H	A square symmetric matrix.
sampleInfo	A 4-column matrix containing sample_id, sex, case_control status, and group_id
n	Number of individual in the H matrix
ntag	Number of tag SNPs used to generate the H matrix.
outfile	The text file to write to. The default is set as MMprime.txt.

**Details**

The first two lines in the output file contains the information about the square matrix. The first line is the number of individuals in the matrix. The second line is the number of tag SNPs. The rest of the file is a table. Each line in the table contains the sample ID, gender, case/control status, group ID, and the square matrix.

**Value**

A text file is produced in the current directory.

**Author(s)**

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

**References**

<http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/directions.pdf>

**See Also**

[http://wpicr.wpic.pitt.edu/WPICCompGen/MMp/MMp\\_page.htm](http://wpicr.wpic.pitt.edu/WPICCompGen/MMp/MMp_page.htm)

---

pc\_graphs\_GEMp      *Plot ancestral structures*

---

**Description**

The function makes .pdf and .ps plots based on the significant eigenvectors from the SpectralGEM function.

**Usage**

```
pc_graphs_GEMp(ext)
```

**Arguments**

`ext`                      the file extension generated by the identifier and the stage. The values are printed out in the R console after running SpectralGEM().

**Details**

The function plots pairs of principle components corresponding to the significant eigenvectors.

**Value**

The plots are in .pdf form and .ps form under the current directory.

**Author(s)**

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

**References**

<http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/GEM+.htm>

**Examples**

```
# pc_graphs_GEMp("small")
```

---

`pc_graphs_GEMpClusters`*Plot ancestral structures with cluster id*

---

**Description**

The function generates ancestral plots in .pdf and .ps form.

**Usage**

```
pc_graphs_GEMpClusters(ext)
```

**Arguments**

`ext`                    the identifier and stage concatenated

**Details**

The function plots pairs of principle componets corresponding to the significant eigenvectors. Use only after the clustering step.

**Value**

The plots are put in the current folder.

**Author(s)**

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

**References**

<http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/GEM+.htm>

**Examples**

```
# pc_graphs_GEMpClusters("small")
```

---

`plotLam`*Plot the eigenvalues*

---

**Description**

The function plots the eigenvalues the main fortran program produces. It helps user to determine the number of dimensions (mdim) to be used in the program.

**Usage**

```
plotLam(ext)
```

**Arguments**

`ext`            The identifier

**Details**

The function generates a plot of all the eigenvalues.

**Author(s)**

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

**References**

<http://wpcr.wpicr.pitt.edu/WPICCompGen/Spectral-GEM/GEM+.htm>

---

`SpectralGEM`*Software for Matching*

---

**Description**

SpectralGEM is designed to find the ancestry vectors and match cases and controls for association analysis.

**Usage**

```
SpectralGEM(InputFile = "matching_input.txt", CM="CM", outlier=TRUE)
```

**Arguments**

<code>InputFile</code>	The name of the file that contains the input parameters.
<code>CM</code>	options are C, M, and CM. C: for clustering only with removing outliers. M: for matching only without examining the outliers. CM: for clustering, removing outliers, and matching.
<code>outlier</code>	An option to remove outliers by checking the distributions of the distances between cases and contrls. Only applicable when CM="C"

**Value**

<code>cl</code>	a two column matrix: the first column is the sample ID, the second column is the cluster id
<code>U</code>	a matrix, the first column is the sample ID, the second columns is group id, the third column is the trivial eigenvector U0, and rest are the significant eigenvectors
<code>lambda</code>	eigenvalues corresponding to the eigenvectors
<code>d</code>	the distance between case and control

The program performs clustering and matching or matching only. The `cl` values are generated at the clustering stage. The significant eigenvectors are generated at the matching stage.

A series of files are produced in the current directory.

**Note**

The function depends on the local fortran executable. The function asks whether the user would like to download the executable before it automatically downloads the executable from the reference website.

**Author(s)**

Ann Lee, Diana Luca, Bert Klei, Bernie Devlin, and Kathryn Roeder

Maintainer: Jing Wu [jwu@stat.cmu.edu](mailto:jwu@stat.cmu.edu)

**References**

<http://wpicr.wpic.pitt.edu/WPICCompGen/Spectral-GEM/GEM+.htm>

**Examples**

```
data(H)
data(id.info)
MMfile(H=H, sampleInfo=id.info, ntag=1000, outfile="MMprime.txt")
InFile(Identifier="smal", stage=1, directory=".",
      MMfile="MMprime.txt", excludefile="exclude.txt",
      idlength=8, mincluster=10, logtype=0,
      outfile="matching_input.txt")

#not run#
```

```

#out=SpectralGEM() #first do clustering and remove outliers
                    #then do matching
#out=SpectralGEM(CM="C") # do clustering and remove
                        #outliers
#out=SpectralGEM(CM="M") # do matching without removing
                        #outliers

# For continuous response, create new id.info for the H matrix
data(H)
n=345
y=sample(c(rnorm(mean=0,172),rnorm(mean=1,173)))
cc=rep(1,345)
cc[y>median(y)]=2 #create case control
newid.info=cbind(c(1:n),rep(1,n),cc,cc)
MMfile(H=H,sampleInfo=newid.info,ntag=1000,outfile="MMprime1.txt")
InFile(identifier="smal",stage=1,directory="./",
        MMfile="MMprime1.txt",excludefile="exclude.txt",
        idlength=8,mincluster=10,logtype=0,
        outfile="input.txt")

#not run#
#out=SpectralGEM(InputFile="input.txt",CM="C")

# buildin ancestry plots
#Current version: ext= small ;
#not run
#pc_graphs_GEMpClusters("small")
#pc_graphs_GEMp("small")

#plot from the SpectralGEM output,
#significant eigenvectors start from out$U[,4] with
#significant eigenvalues start from from out$lambda[2]
#not run
#plot(sqrt(out$lambda[2])*out$U[,4],sqrt(out$lambda[3])*out$U[,5],
#col=out$U[,2],xlab="PC 1",ylab="PC 2") #for PC plots

```

---

```
updateExcludeFileDstr
```

*Exclude outlier*

---

## Description

Exclude outlier by examining the distribution of case/control distances. Internal function.

## Usage

```
updateExcludeFileDstr(excludeFile, ext)
```

## Arguments

excludeFile	the text file that the outliers will be recorded to
ext	the file extension used in running the program

**Value**

The outliers are recorded in the `excludeFile` in the current folder.

---

<code>updateInputFile</code>	<i>Update input file</i>
------------------------------	--------------------------

---

**Description**

Internal function.

**Usage**

```
updateInputFile(oldInputFile, newInputFile, excludeFile)
```

**Arguments**

`oldInputFile` old input file name  
`newInputFile` new input file name  
`excludeFile` exclude file name

# Index

- \*Topic **datasets**
  - id.info, 3
- \*Topic **data**
  - H, 3
- \*Topic **dplot**
  - pc\_graphs\_GEMp, 6
  - pc\_graphs\_GEMpClusters, 7
- \*Topic **file**
  - getExcludeFile, 2
  - getRecord, 2
  - getVersion, 3
  - InFile, 4
  - loadInputFile, 5
  - MMfile, 5
  - plotLam, 8
  - updateExcludeFileDstr, 11
  - updateInputFile, 11
- \*Topic **models**
  - full\_matching, 1
  - SpectralGEM, 9
- \*Topic **regression**
  - SpectralGEM, 9
- \*Topic **survival**
  - SpectralGEM, 9

full\_matching, 1

getExcludeFile, 2

getRecord, 2

getVersion, 3

H, 3

id.info, 3

InFile, 4

loadInputFile, 5

MMfile, 5

pc\_graphs\_GEMp, 6

pc\_graphs\_GEMpClusters, 7

plotLam, 8

SpectralGEM, 9

updateExcludeFileDstr, 11

updateInputFile, 11