Model selection in High-Dimensions: A Quadratic-risk Based Approach.

Surajit Ray and Bruce G. Lindsay, Journal of the Royal Statistical Society - Series B

October 17, 2007

This software is written in C++ and it has been tested in UNIX / LINUX / MAC OSX based platform. It calculates the quadratic risk of multivariate mixture models.

This program uses the Newmat Library and the EMMIX fortran library provided by Geoff McLachlan. The appropriate binaries are provided in the zipped file.

1 INSTALLATION

To install the package, follow these steps:

1. Download the relevant package from:
   `http://math.bu.edu/people/sray/software/`
2. Unzip and Untar the package under command shell:

   ```
   tar -zxvf quadrisk.tar.gz
   ```

   Move to the directory `quadrisk` `cd quadrisk`
   This will create a directory quadrisk which contains the executable files `quadrisk` and `emmix` and datasets

2 TEST

Under the command shell issue the command

```
./quadrisk -equal -3 -g 3 -i iris.dat -d 4 -n 150 -p iris.param
```

This command should at the end produce the risk file **iris.equal.risk** which contains the risk calculations.[Full Input and Output description is provided below]. If the program terminates please check to see if you have downloaded the correct binary.

3 PROGRAM USAGES

Description of options

```
-equal[-unequal]        Equal Variance or Unequal  variance

-[1|2|3]        1: No estimation needed, need to use parameter file
                2: Estimate Parameters using included EM steps
                3: Use Geoff McLachlan's emmix program to estimate parameters

-h [value]      User specified smoothing parameter

-i [filename]   Datafile name. Allocates the n and d dynamically

-p [filename]   Paremeter file name

-d [int]        Dimension of Input data. If missing the program calculates it.

-n [int]        Rows/Size of Input data. If missing the program calculates it.

-g [int]        Number of components to be fitted. Default 6.

-k [int]        Number of K-means start.Default 20.

-r [int]        Number of random starts.Default 3.

-s [seed]       Random seed. Default seed calculated form current time
```

Options '-k -r and -s' are optional and are not not used when using the option '-1'

Examples:
To see the menu/list of options type
`./quadrisk` or `./quadrisk --help`

To estimate the parameters using the EMMIX program and calculate the quadratic risk of the iris data use

```
./quadrisk -equal -3 -g 6 -i iris.dat -d 4 -n 150 -h 0.5 -p iris.param
```

The parameters will be stored in iris.param

To calculate the quadratic risk using estimated parameters in the iris.param file use

```
./quadrisk -equal -1 -g 6 -i iris.dat -d 4 -n 150 -p -h 0.5 iris.param
```

To estimate the parameters using the EMMIX program and calculate the quadratic risk of the iris data with specified number of kmeans and random starting values use

```
./quadrisk -equal -3 -g 6 -i iris.dat -d 4 -n 150 -p -h 0.5 iris.param -k 4 -r 8
```

4 Output Details

Besides the on screen output the program also stores the final risk calculations in datafilename.[equal—unequal].risk (e.g,. iris.equal.risk) From left to write the details for the contents of the output file are

```
No.of.Param:  Total number of parameters estimated
DISTANCE   :  The quadratic distance between the model and the data
MLF        :  Model Lack of Fit
EE         :  Estimation error
RISKA      :  QAIC
RISKB      :  QBIC
Likelihood :  Likelihood
AIC        :  AIC
BIC        :  BIC
```

5 DATASET

The zipped file dataset.zip contains the four data files one for each model described in the paper. R-codes for generation of similar files are also provided for each models.