FIXREG
A software for Fixed Point Cluster analysis of linear regression data

Christian Hennig
Institut für Mathematische Stochastik
Universität Hamburg
Bundesstr. 55
D-20146 Hamburg, Germany
hennig@math.uni-hamburg.de

Abstract

FIXREG is a software that performs Fixed Point Cluster (FPC) analysis for regression data. FPC analysis is a technique for cluster analysis based on the identification of outliers. In general it is introduced in Hennig (1998a), and the linear regression case is treated in Hennig (1997) and Hennig (1998b). An FPC is a subset of a dataset that does not contain any outlier, and all other points of the dataset have to be outliers w.r.t. the FPC. FPCs can overlap and need not to be exhaustive. Outliers need not to be included in any FPC and FPCs remain stable under addition or deletion of outliers. FPC analysis does not require a specification of the number of clusters.

This paper should serve as a user's manual for the software FIXREG. The software FIXMAHAL for finding multivariate Gaussian FPCs can be handled in a similar way.

1 Regression Fixed Point Clusters

FPC analysis is a tool for finding subsets (*clusters*) of a dataset which are

- homogenous in the sense that they can be adequately described by some homogenous parametrical distribution which I call *cluster reference distribution*, and that they contain no outlier from this distribution,
- separated from the rest of the data in the sense that all other data points are outliers w.r.t. to the FPC.

The concept is explained in more detail in Hennig (1997, 1998a).

A linear regression cluster of a dataset

$$\mathbf{Z} = (\mathbf{X}, \mathbf{y}),\ \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' \in (I\!R^p)^n,\ \mathbf{y} = (y_1, \ldots, y_n)' \in I\!R^n$$

is a subset of data points that is characterized by approximately the same linear relationship between a $p-$dimensional independent variable $\mathbf{x}$ and a dependent variable y. That is, it can be described by a cluster reference distribution of the form

$$\begin{gathered}\mathcal{L}(y|\mathbf{x}) = F_{(\mathbf{x},\beta,\sigma^2)},\ \text{defined by}\\ y = \mathbf{x}'\beta + \beta_{p+1} + u, \qquad \mathcal{L}(u) = \mathcal{N}_{(0,\sigma^2)},\\ (\beta, \sigma^2) \in I\!R^{p+1} \times I\!R_0^+,\end{gathered}$$

and this relationship should separate the points of the linear regression cluster from the remaining data points. The values of the independent variable $\mathbf{x}$ may be random or fixed.

Let $g \in \{0,1\}^n$ be some indicator vector and $\mathbf{Z}(g)$ be the matrix of the data points $(\mathbf{x}_i, y_i)$ indicated by $g_i = 1$. Then the data subset $\mathbf{Z}(g)$ is called *Fixed Point Cluster* (FPC) w.r.t. $\mathbf{Z}$, if g is a fixed point of

$$\begin{gathered}f:\ \{0,1\}^n \mapsto \{0,1\}^n,\\ f_i(g) = 1\left[(y_i - \mathbf{x}_i'\hat\beta(\mathbf{Z}(g)))^2 \leq c\hat\sigma^2(\mathbf{Z}(g))\right]\end{gathered}$$

with some prechosen constant c (e.g. $c = 10$). $\hat{\beta}(\mathbf{Z}(g))$ and $\hat{\sigma}^2(\mathbf{Z}(g))$ are the least squares regression and the corresponding UMVU error variance estimators based on the data subset $\mathbf{Z}(g)$.

The function f is an outlier identifier (0 for outliers): A point is considered as an outlier w.r.t. a linear regression distribution with parameters β and σ^2 if it falls into the outlier region $\{(y - \mathbf{x}'\beta)^2 > c\sigma^2\}$.

Therefore an FPC $\mathbf{Z}(g)$ has the property of being exactly the set of non-outliers in $\mathbf{Z}$ w.r.t. itself. In this sense, an FPC is homogenuous and separated from the rest and can therefore be interpreted as a cluster. Note that the FPC itself needs not to have a "'Gaussian error linear regression shape"'. The linear regression reference distribution only defines what "'outlier"' and "'separatedness"' mean.

The constant c defines the tolerance level of the outlier identification. It is related to the probability that a regular point from a homogenous linear regression distribution is identified as an outlier. This probability is called *level* in the following. $c = 10$ corresponds to a level of 0.00157, for level 0.01 we get $c = 6.635$. FPC analysis with higher c yields better separated FPCs. For $c = 6.635$ there ary usually more FPCs while for $c = 10$ the method finds only the better separated clusters.

Fixed point clusters are found by an algorithm that starts with some random or prechosen point configuration and converges towards an FPC. The algorithm must run many times to find all relevant FPCs of a dataset.

In difference to other methods for linear regression clustering (Hennig (1998b) gives an overview) FPC analysis does not require a specification of the usually unknown number of clusters and it is able to treat outliers from the whole data adequately. Clusters may overlap and the data set may not be exhausted by clusters.

The program FIXREG finds FPCs in datasets. It shows the points that form FPCs and the corresponding indicator vectors (that can be used e.g. by some visualization software), regression and scale parameter estimators. There are default settings for c and the number of algorithm runs that can be modified by the user. If there is some a-priori grouping of the data and the interest lies in the question if these groups correspond to well separated clusters in the data, the algorithm can be started with these groups. The same strategy is appropriate if the dataset has a large dimension or very many points since random search needs so many iterations in these cases that the computing time can get excessive.

You can also compute the outliers w.r.t. a given data subset with FIXREG, see section 4.4 o).

2 Loading FIXREG

FIXREG can be downloaded from my web page:

`http://www.math.uni-hamburg.de/home/hennig/`

You can get the following files:

fixreg.exe - MS-DOS executable file

fixreg.tar.gz - source code, makefile, data example, `manreg.tex`, unzip with

```
gzip -d fixreg.tar.gz
tar -xvf fixreg.tar
```

on UNIX and LINUX systems.

unzip.exe - unzip-program for MS-DOS

fixreg.zip - source code, makefile, data example, `manual.tex`, unzip with

```
unzip fixreg.zip
```

on systems where `unzip.exe` is avialable.

A `fixreg`-directory will be created where the software is unzipped.

3 Before running FIXREG

FIXREG is written in C, i.e. it can be used on every system that contains a C-compiler. Under MS-DOS you can use the `FIXREG.EXE`-File and start it with the command `fixreg` without any compilation work.

Under UNIX/LINUX you have to compile an executable file first. This also holds if the transfer of the binary file `FIXREG.EXE` did not work or if you want to change the limitation of the dataset size of $n = 700$ and $p = 4$ that is implemented in `FIXREG.EXE` due to the limited stack space of the Borland C-compiler.

For the compilation you need one of the following C-compilers on your system:

cc (Sun standard C-compiler:) Compilation by `make -f fixregcc.mak`. Afterwards the program can be started by `fixregcc`.

gcc (Gnu-compiler, installed on almost all UNIX and LINUX-systems:) Compilation by `make -f fixreggc.mak`. Afterwards the program can be started by `fixreg`.

bcc (Borland-compiler under MS-DOS; the older version `tcc` should also work well:) Compilation by `make -f fixregbc.mak`. Afterwards the program can be started by `fixregbc`.

Other C-compilers may work as well, but then you have to create your own `make`-file by plugging in your compiler/linker and the appropriate options into the lines

```
CC=
LN=

CC_OPTS=
LN_OPTS=
```

of one of the `.mak`-files from the WWW-installation.

The compilation needs the following files in the same directory (this will be done automatically if you download the software from the WWW):

`fixdefs.h`	`fixglo.h`
`vfm.h`	`vfm.c`
`lesen.h`	`lesen.c`
`regstan.h`	`regstan.c`
`lssch.h`	`lssch.c`
`standard.h`	`standard.c`
`matrix.h`	`matrix.c`

and the suitable `.mak`-file. You also will find the test data described below and the LaTeX-file of this manual.

You can modify the sizes of some variables - maximum size of the dataset, maximum filename length for in- and output etc. - using the file `fixdefs.h` that looks like follows:

```
#define NKOR [number] /* Maximum n */
#define PKOR [number]  /* Maximum p; for regression maximum (p + 1) */
#define NLEN 50 /* Maximum filename length  */
#define VLEN 10 /* Maximum variable name length */

#ifndef _vecdef_
#define _vecdef_

typedef double vektor[PKOR];
typedef double kvektor[PKOR-1];

#endif
```

Note that `PKOR` defines the maximum dimension of the regression parameter β including the intercept parameter, i.e. `PKOR 5` leads to maximum $p = 4$. Take care of any memory restrictions caused by your compiler or your system. In principle there is no limitation of the dataset size. I tried examples up to $n = 187000$, but this needs lots of memory even for the output, since casewise information is given for every cluster.

4 Running FIXREG

4.1 In case of trouble...

The software FIXREG was tested with many datasets on various systems. However, one can never be sure to avoid any remaining errors and system incompatibilities. Also the protection of the software against inadmissible input is weak. If you have any trouble, please check first if it could be a consequence of some wrong or inadmissible input. Surprising results and even program crashs can also stem from problems with the data set (e.g. if there are commata instead of decimal points). You can take option i) from section 4.4 to examine if your data was read properly. Too large values for `NKOR` and `PKOR` in `fixdefs.h` (see section 3 to modify) can result in stack overflow and memory errors. FIXREG needs memory to store alle found FPCs. If there are too much of them, the output closes with

```
WARNING! Heap memory full! Program terminated.
Enlarge minimum cluster size or scale factor to get less clusters.
```

In that case consider e) and h) in section 4.4.

In case of other problems, please contact me under hennig@math.uni-hamburg.de.

4.2 Data format

After having started the program, the first thing you have to specify is the format of your data. Your data need to be an ASCII-file. The values have to be separated by arbitrarily many spaces, carriage returns or TAB-characters. The program reads files casewise, i.e. the file has to contain all variable values of one case before giving the values of the next case. Usually it will have the form "lines are cases, columns are variables". Decimal points have to be points. Scientific notation (1.2256e-8) is possible. You have three options for the beginning of the file:

```
 1: Data starts with 1st line, manual input of n, p
 2: 1st line head,
    2nd line number of data points n, number of indep. variables p
 3: 1st line head, manual input of n, p
```

The following dataset (belcalls.dat) contains a headline and a line indicating $n = 24$ and $p = 1$ (option 2). It was taken from Rousseeuw and Leroy (1987). The first variable is the number of telephone calls from Belgium in 10 millions, the second variable is the year, the third was added for illustratory purposes and will not be used for regression calculations. We have one independent variable "year" so that $p = 1$.

```
Calls from Belgium
24 1
0.44 50  1
0.47 51  1
0.47 52  1
0.59 53  1
0.66 54  1
0.73 55  2
0.81 56  2
0.88 57  2
1.06 58  2
1.20 59  2
1.35 60  3
1.49 61  3
1.61 62  3
2.12 63  3
11.9 64  3
12.4 65  4
14.2 66  4
15.9 67  4
18.2 68  4
```

```
21.2 69  4
4.3  70  5
2.4  71  5
2.7  72  5
2.9  73  5
```

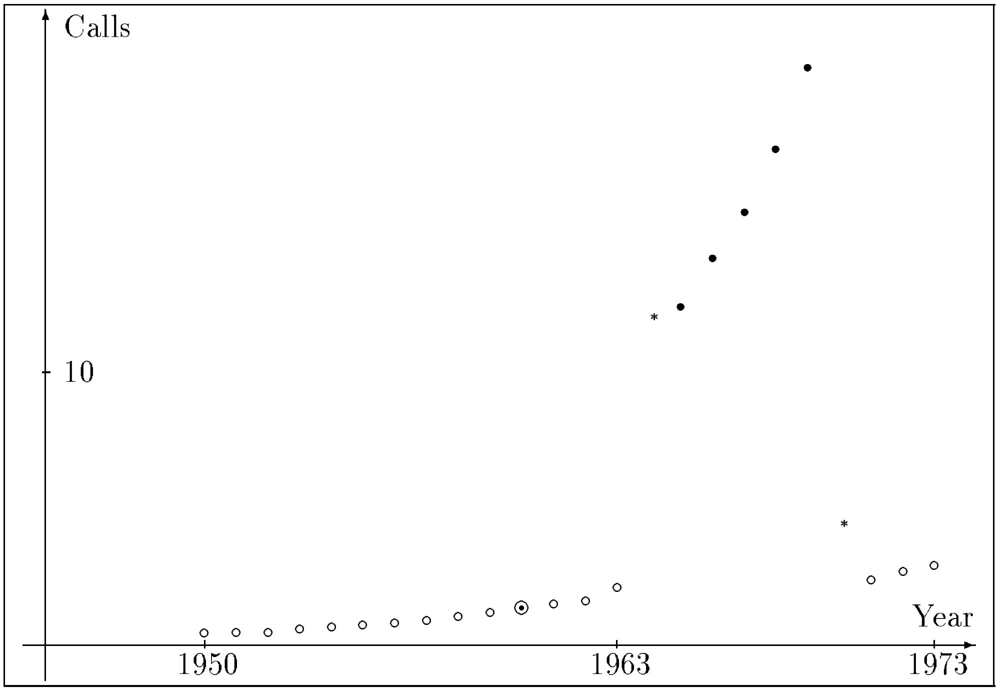

Figure 1: Calls from Belgium with FPCs 4 and 5, see section 5

The data file may contain more cases and variables than those needed for the analysis. Missing values are possible. They must be indicated by a single character that is not allowed to be a number, e.g. "*" or ".".

4.3 File specifications

Now you have to enter the filename. The next questions of the program are:

```
Read an identificator variable (value length max. 10) ? (y/n)
Use a grouping variable for iteration starts ? (y/n)
```

An identificator variable is a variable that is used to indicate the cases in the output file. If no identificator variable is given, the cases are indicated by their position number in the data file. The values of the identificator variable have to be strings without spaces. A variable used for other purposes can also serve as identificator variable, e.g. the independent variable `year` in the `belcalls`-data.

If you want to start the FPC algorithm from some prechosen point configuration, you can define a "grouping variable". The grouping variable must be non-negative integer-valued and can be contained in the data file or may be generated by use of the case numbers (see below). In the `belcalls`-example, the third variable will serve as a grouping variable. That is, there will be five iteration runs, started from the points indicated by the values 1, 2, 3, 4, 5, respectively, of the grouping variable additional to a number of iteration runs from random starting configurations.

Now you have to specify the name of the output file. It is important at least under MS-DOS that you do not give it an ending. The endings are added to the name by the program. FIXREG creates a *main output file* that gets the ending `.fix` and for each FPC an additional *case information file* getting a number as ending. For the `belcalls`-data I entered the filename `belcalls` and got output files `belcalls.fix, belcalls.1, belcalls.2` and so on. In a particular situation you might get also files with the ending `.nnumber`, i.e. `belcalls.n1` and so on, see section 4.4 o).

If you made any wrong inputs up to now, break the program and start again. The following steps can be repeated.

Next the program asks you to give the total numer of variables in the input file (3 in the example). Afterwards you can enter the variable names and positions (variables, not characters are counted). For the `belcalls`-example:

```
  Name of identificator variable    year
  Position from the left ? 2
  Name of dependent variable  calls
  Position from the left: 1
  Name of 1th independent variable  year
  Position from the left of 1th independent variable:  2
  Name of grouping variable (values must be integer >= 0!) group
  Generation of grouping variable by case number? (y/n) n
  Position from the left of grouping variable:  3
```

If you decide to generate the grouping variable by case numbers, the program asks

```
  Group [number] from case [previous input] to case [input number]
```

until the input number is greater or equal than n.

Now you have to enter the character that indicates a missing value. Enter an arbitrary character if the data does not contain missing values. All cases with missing values will be excluded from the analysis.

Answering "n" to the question `Everything correct ? (y/n)` gives rise to a repetition of the variable definition procedure.

4.4 Iteration specifications

Now the program shows the settings for the iterations (default settings are given in square brackets):

```
 Settings for cluster search:
 (*) indicates test functions for curious users; possibly unstable or useless.
     a) [100*3^p] runs with random start,
     b) (*) Increase factor of search probability for data not included in
              clusters: [1] (integer !),
     c) (*) Two levels for cluster search: [no]
     e) Standard level [0.00157] (=>Scale factor c [10])
     f) (*) Number of data points for search start: [p+2]
     g) (*) Do you want to get a question for break all 20 runs: [no]
     h) Minimum size of clusters: [2*(p+3)]
     i) Output of standardized values (median and MAD): [no]
     j) (*) Do you want to get a question if a cluster is relevant: [no]
     k) (*) Do you want to enter points with decreased search probability: [no]
     m) Input of scale factor for standard level
     n) Suppress standardization: [no]
     o) Iteration step limit: [min(n,5000)]
     q) Start of calculations
Enter a letter, if you want to change the default settings
```

During the development of the program I used some test functions that are presumably useless for the standard user. In order to that, they are not tested as good as the rest of the program. Nervertheless I decided to keep them in the program for curious users. They are not fully documentated. If you want to be as sure as possible that everything will work, feel discouraged to change the settings indicated by (*).

You can start the program immediately by entering "q". The default settings can be changed by chosing one of the other letters. If you decide to change some number, you will be asked for the new value after having entered the letter. To change a "yes-no-variable", you only have to enter the letter. The order of your changes is free. You may repeat and therefore cancel all your changes.

a) The default setting for the number of iterations with randomly chosen starting configuration is chosen so that the probability for finding an FPC of $\frac{1}{3}$ of the points of the dataset stays approximatively the same for each value of p. If the cluster would be very well separated, i.e. it would be a clear pattern of the dataset, this probability would be very high. The probability decreases with the size of the cluster relatively to n, but with the default choice you have already good chances to find clear separated FPCs down to $\frac{1}{5}$ of the dataset.

There could be two main reasons for changing the default: If n is large, the interest in finding smaller clusters could be greater so that the value should be enlarged. If on the other hand p is too large, the default setting leads to too large - with $p > 4$ presumably unacceptable - computation times. The only feasible way to overcome this problem lies in the use of deterministic starting configurations from some reasonable grouping variable based on a-priori information about the data. The number of iterations with random starting configurations can then be set to 0; a small number of random iterations can only lead by chance to reasonable findings.

b) (*) If you want to find parts of a partition of the dataset, it can be reasonable to exclude the points from the random search, which are already included in found FPCs. You can use option b) to increase the probability of points, that are not already parts of found FPCs, compared to the others. In my experience this does not lead to better results.

c) (*) If you change c), the program will always first iterate an FPC of a lower level[1] and the iteration with the *standard level* will always be started with the FPC of lower level. For the lower level clusters you will get a reduced output and so I suggest to run the program twice if you really are interested in the results using two distinct levels.

d), l) only appear if you have changed c) and correspond to e), m).

e) If you want to change the level[2] and the "tuning constant" (*scale factor*) c, you can enter the new level directly by "e". Choose "m" to change the level via c. The respective corresponding other value will be calculated by the program. Experience shows that in a small dataset often more FPCs occur. This is because there are more clear gaps between points in that case. Though $c = 10$ often appears to be a good choice, I suggest to try smaller values ($c \geq 5$) for very large datasets, if too few FPCs were found, and larger values to prevent a too large number of FPCs in small datasets, especially where $p > 3$.

f) (*) By default, the random starting configuration has the least possible size. This leads to the best chance to find many clusters. If you change f), do not choose a smaller value. If you enter a larger value, you decrease the chance to find clusters with remarkably fewer points as n.

g) (*) If you change g), you will get a question all 20 runs if the iteration should be stopped.

h) The probability is high that some small point configurations lie almost exactly on some regression hyperplane in every data set. If you allow all these constellations to come out as an FPC, you can get a very large output. This holds especially for higher p. If the data set is small, you could be interested in smaller clusters nevertheless: I changed this value to 5 for the example run with the `belcalls`-data. If n or p are large it could be reasonable to increase the value to see less FPCs. Note that in my experience the restrictions to some point group to come out as an FPC are remarkably smaller than using some usual partitioning method.

i) For decreasing the probability of numerical problems with very small or large values, FIXREG standardizes all the variables to median 0 and median absolute deviation ("MAD") 1 (see Rousseeuw and Leroy (1987) for explanations). This does not affect the results because all parameter estimators will be re-standardized and FPC analysis is regression equivariant. On the other hand, the residuals in the case information files of the output (see section 5.2) will come from the standardized computations. If you want to re-standardize these values, you have to multiply them with the MAD of the dependent variable. Choose "i" if you want to see the standardized values, the variable medians and MADs[3]. This could also be a way to test if your data was read properly.

[1] The term "level" is explained in section 1.

[2] The term "level" is explained in section 1, the term "standard level" is chosen as opposed to the lower level, see c).

[3] If some variable has MAD= 0, i.e. more than half of the values are equal, the standardization will be omitted.

j) (*) If an FPC is found, you can get a question if it is "relevant" in your opinion. If it is not, it will not appear in the output.

k) (*) You can choose "k" if you want to decrease the probability of some data points to get into the starting configuration. This means that the points are treated as points of already found FPCs in b); change b) also to get a real decrease. The input procedure is not comfortable and this option is not tested very good.

m) see e).

n) see i). If n is large, the standardization takes a lot of time and option n) becomes reasonable. Also changing n) is a way to get unstandardized residuals into the case information output files (see i)). Note that this does not prevent the word "standardized" from appearing in the output. It must be ignored if the standardization has been suppressed.

o) Though the algorithm is proven to converge in a finite number of steps, the convergence can fail because of rounding errors[4]. Here is the number of iteration steps after which the program terminates the iteration. It is large enough that I expect the iteration to find an FPC certainly even with $n > 100000$ if it does not get trapped in some loop. If this happens, you will get the information in the output. If n is very large and your computer is slow or you do not have enough time, a smaller number could be reasonable.

Since the algorithm exceeds the step limit usually only when convergence to some FPC has finished and some single points get out of and into the configuration, FIXREG produces a case information file (see section 5.2) also in the case that the step limit was exceeded. These files get the ending `.n1`, `.n2` and so on. This enables a further attractive feature of the program: If you want to compute the outliers w.r.t. to some given data subset, you can use some grouping variable for iteration starts, where the subset of interest appears as a group, and set the iteration step limit to 1. Then you get the outlier indicator vector (0 for outliers) in the corresponding case information file.

After having entered "q", the program starts the first iteration with the whole data set as starting configuration. This leads usually to some FPC that contains almost the whole data set (*WD-cluster*) and that also appears very often during the iterations with random starting configuration. If there are enough points in the complement of the WD-cluster, an iteration follows that starts with these points. Then the iterations with random starting configurations are carried out and at last the iterations with starting configurations determined by the grouping variable. The monitor shows you the state of the program.

[4] I know only one example of this phenomenon.

5 Interpreting the FIXREG-output

5.1 The main output file

The output consists of a main output file with the suffix `.fix` and case information files for each found FPC that have the endings `.[FPC-number]`. First it is advisable to consider the main output file. It is called `belcalls.fix` for the `belcalls`-example. The file starts with some informations concerning the data set and the iteration settings. You get the medians, MADs and standardized data values if you have changed setting i). The file also contains the LS-regression estimator of the whole data set. After that, the found FPCs, i.e. their points and parameters - sometimes very many - are shown in order of appearance.

I recommend to start the inspection of the output by taking a look at the end of the file for the `cluster similarity table`. This can help you considerably to understand your output. The table for the `belcalls`-data looks as follows[5]:

```
* Cluster similarity table *
Total number of standard level-clusters: 14
Number of points in the intersection of standard level-clusters no.

No.  Total    1    2    3    4    5    6    7    8    9   10
Times found 174   22   29   58    3    1    1    1    1    1

  1     24         13   16   17    6    6    5    6    7    6
  2     13              13   13    1    6    5    5    7    2
  3     16                   16    1    6    5    6    7    3
  4     17                         1    6    5    6    7    3
  5      6                              1    0    0    1    1
  6      6                                   1    1    6    0
  7      5                                        5    2    0
  8      6                                             2    0
  9      7                                                  0
 10      6
 11      6
 12      7
 13      7
 14      8
No.  Total   11   12   13   14
Times found   1    3    1    1

  1     24    6    7    7    8
  2     13    6    5    7    6
  3     16    6    7    7    8
  4     17    6    7    7    8
  5      6    0    0    1    0
```

[5] Recall that this is the result of iterations with random starting configurations. It will not necessarily be reproduced with the same data set and the same iteration settings.

```
  6      6    1    1    6    1
  7      5    5    5    2    5
  8      6    5    6    2    6
  9      7    2    2    6    2
 10      6    0    0    0    0
 11      6         5    2    6
 12      7              2    7
 13      7                   2
 14      8

0 iterations led to collinear regressors.

0 iterations did not finish fast enough.

9 iterations led to too small clusters.
```

The program found 14 FPCs. This seems to be a lot and not all are interesting. You can use the table to get some overview. The line `Times found` gives the number of times that the algorithm yielded the corresponding FPC. The column `Total` gives the number of points that belong to the FPC. The rest of the table gives the number of points in the intersection of the FPCs. The number of "Times found" is a measure for the stability of an FPC. If there are too many clusters, one should start to interpret the most stable FPCs. FPCs that occured only once are usually not of interest. Note, however, that it is more difficult to find a small cluster. The numbers of "Times found" are only adequate to compare FPCs of roughly the same size.

We see that cluster 1 that contains 24 points (i.e. all) came out 174 times of 306 runs (300 random + 1 whole dataset + 5 groups as starting configurations). Cluster 1 is always the WD-cluster. The first guideline to the interpretion is that almost always one or more clusters are found most often which contain almost the whole data set. This happens since if the iteration starts with points that cannot be fitted adequately by the same cluster reference distribution, it usually converges to the WD-cluster or some very similar constellation. The corresponding FPCs are only interesting if one wants to find outliers from the tendency of the whole data set. These will not be included at least in some of the FPCs that are similar to the whole data set.

The most stable other clusters are the clusters 2, 3, and 4. The table shows that 2 is a subset of 3 that is a subset of 4. That is, we have three variants of the same data pattern and if we are only interested in a rough description of the data set, we can concentrate on FPC 4 that came out most often (empty circles in Fig. 1). If we want a more sophisticated interpretation, we can explain the points that belong to some but not all of these FPCs as lying "in the periphery" of the pattern. The next interesting cluster is number 5 (filled circles in Fig. 1). There is only one point intersection between 4 and 5 and so FPC 5 clearly describes a distinct pattern of the data. Rousseeuw and Leroy (1987) explained that in the years 1964-1969 and in parts of 1963 and 1970 not the telephone calls but the minutes were recorded. The FPCs 3 and 5 correspond to the distinct periods, the 17th point of FPC 4 is 1963 that is roughly but not clearly assigned to the group where the calls had been recorded. The point in the intersection is 1960 (circle around small filled circle in Fig. 1) that cannot be separated from FPC 5 - the call minutes increase faster than the call numbers and presumably in 1960 the values would have been similar. 1964 does not belong to FPC 5 since it does not fit well enough into the general tendency of the "minutes"-period.

There are nine further FPCs. Most of them were found only once and correspond to parts of the "number of calls"-period that can be fitted approximately exact. You could use them for a very detailed analysis of the data, but I will not discuss them here. The number of clusters could be reduced by the choice of a larger c.

The single clusters appear in the main output file as follows:

```
Successful run.

Output of standard level-cluster no. 4:

LS-estimator
indep. variable   coeffizient    standardized coeff.
       year         0.11052887364    0.65016984495
constant           -5.26015151515

Squared scale estimator:   0.021290, bias-corrected:   0.021691

Cluster quality against whole dataset:   7.0383527
Cluster quality against  WD-cluster:   7.0383527
Corresponding outlier region: absolute residual>         0.4614;
                 absolute standardized residual>         0.3051

Cluster contains points
         50          51          52          53          54          55
         56          57          58          59          60          61
         62          63          71          72          73
Cluster contains 17 points
```

First the program shows the LS-regression estimator. The coefficient for `constant` is the intercept parameter. `standardized coeff.` indicates the regression parameter obtained with the standardized variables, see section 4.4, option i). The `squared scale estimator` is the corresponding UMVU-variance estimator $\hat{\sigma}^2$ under Gaussian errors. If the data would consist of only one cluster or there would be no overlap between clusters, the scale estimators for the single clusters would be biased because sometimes data points would be cut off by the outlier identification. A `bias-corrected` value for the scale estimator is given that would be a consistent scale estimator for a homogenous dataset. For overlapping clusters, this interpretation fails.

The values for the `cluster quality` are the logarithms (to the basis 10) of 1 divided by P where P is the probability to get a sample of the size and variance lower or equal than that of the FPC by random from a population that has the variance of the whole dataset, the WD-cluster respectively. High values of the "quality" mean that the variance is very low and the FPC is fitted much better than one would expect for a subset of homogenous data. That is, the "quality" is a kind of validity measure. Values of 5 and larger are "high" in my experience. A quality of 100 indicates that P could not be precisely computed because it was too small. However, the value of this measure is limited because the FPCs were not found "by random" but by some algorithm designed to find subsets with small variance. So at least the small FPCs have almost always "high quality", even if they stem indeed from a homogenous population.

The following lines give the value of $\sqrt{c\hat{\sigma}^2}$, i.e. points with larger residuals from the regression hyperplane defined by the FPC's parameters are "outliers w.r.t. the FPC" in the sense of section 1. The value from the standardized computation can be used to assess how far the outliers lied out and what non-outliers were "close to the edge" with help of the case information files.

If you used a grouping variable, you get also the information what FPCs came out when the iterations were started from the groups above the cluster similarity table, e.g.

```
* Iteration start with group 4: *
* Iteration led to standard level-cluster no. 5. *
```

5.2 The case information files

The case information files, e.g. `belcalls.5`, look like this:

```
Case information for cluster no. 5

year - group - residual of standardized value - cluster indicator
          50       1         13.62304  0
          51       1         12.22138  0
          52       1         10.79987  0
          53       1         9.457722  0
          54       1         8.082507  0
          55       2         6.707291  0
          56       2         5.338689  0
          57       2         3.963473  0
          58       2         2.660997  0
          59       2          1.33207  0
          60       3      0.009756375  1
          61       3         -1.31917  0
          62       3        -2.661322  0
          63       3        -3.745581  0
          64       3         1.300091  0
          65       4        0.2092201  1
          66       4      -0.02200605  1
          67       4       -0.3193587  1
          68       4       -0.2199521  1
          69       4        0.3423404  1
          70       5        -12.25455  0
          71       5        -14.93246  0
          72       5        -16.15558  0
          73       5        -17.44483  0
```

```
group - number of cases from group in cluster
   1    0 (of 5)
   2    0 (of 5)
   3    1 (of 5)
   4    5 (of 5)
   5    0 (of 4)
```

If you want to visualize the findings of the FPC analysis, I recommend to add the values as new columns[6] to your data file and to use some software that is able to show e.g. scatterplots - possibly with more than two dimensions. The cluster indicator can be used to give the points of the FPC a particular color. The standardized residuals can be used to assess the separatedness of the FPC from the rest of the data. Of course, identificators and groups appear only if you used an identificator or a group variable. The list `group - number of cases from group` helps you to evaluate the relation of your groups to the FPCs. You can see to what extent the groups correspond to clusters in the data.

6 References

Hennig, C. (1997). Fixed Point Clusters and their Relation to Stochastic Models, in: *Classification and Knowledge Organization*, Klar, R. and Opitz, O. (Eds.), Springer, 20-28.

Hennig, C. (1998a). Clustering and Outlier Identification: Fixed Point Cluster Analysis, in: *Advances in Data Science and Classification*, Rizzi, A., Vichi, M. and Bock, H.-H. (Eds.), Springer, 37-42.

Hennig, C. (1998b). Models and Methods for Clusterwise Linear Regression, in Preprint-No. 98-5, Institut für Mathematische Stochastik, Hamburg.

Rousseeuw, P. J. and. Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, Wiley.

All my papers are avialable from `http://www.math.uni-hamburg.de/home/hennig/`.

[6] Unfortunately most of the text editors that I know are not able to handle columns adequately. Under UNIX you can use the direct shell commands `cut` and `paste`, under MS-DOS I recommend the *q-editor*. With some effort, you can also do it with *EXCEL*.