User Manual for PCCAT

for State of Michigan, Department of Environmental Quality

Zhen Zhang

Department of Statistics and Probability

Michigan State University

December 28, 2009

Abstract

PCCAT (Principle Component and Clustering Analysis Tool) is an add-on module for ***R*** designed for Michigan Department of Environmental Quality (MDEQ) clients via Center for Statistical Training and Consulting (CSTAT). It is mainly designed for multivariate environmental data analysis. Based on statistical software ***R*** functions and packages, it consists of four main steps: pre-processing, visualization, principle component analysis and clustering analysis. It has been designed to be user-friendly with graphic output and user prompts to take advantage of the many useful functions and packages coded in ***R***, combined with Excel for data analysis.

Contents

1 Introduction

PCCAT is a user-friendly platform based on a collection of ***R*** functions for exploratory data analysis, principle component analysis and clustering analysis. ***R*** [15] [11] is an open source, free statistical software package (***R*** web site CRAN: http://www.r-project.org/). It is based on the S language [17] and can be viewed as an open source, free alternative to the commercial statistical software S-Plus. Most code written for S-Plus will run in ***R*** and vice versa. ***R*** is extremely powerful because many users contribute packages of new functions, while its base functions and packages are relatively stable and reliable for data analysis. Designed for multivariate and exploratory data analysis, PCCAT consists of four main modules for environmental data analysis which are *data preprocessing*, *visualization*, *principle component analysis* and *clustering analysis* written in ***R***. Some modules use selected collections of existing ***R*** functions, but provide customized output and options, according to the requirements from the MDEQ clients. In other modules, the basic functions are modified and new features based on recent researches on statistical techniques have been added.

To better protect the integrity of the raw data likely to be used by the MDEQ staff, Rexcel, an add-on for Microsoft Excel that interacts with ***R***, is recommended for reading data directly from Excel and analyzing data in R. There is a useful video tutorial by its author Erich Neuwirth about fully using it on the web site http://www.statconn.com/. This powerful tool combining ***R*** and Excel is recommended for PCCAT users.

1.1 Installation

To begin, first download ***R*** from the CRAN web site. ***R*** for Windows operation system can be downloaded at http://cran.r-project.org/bin/windows/base/. It is suggested that you download and install the latest ***R*** version. Furthermore, PCCAT requires several add-on ***R*** packages, which will be automatically loaded or installed if not detected. In the latter case you will be provided to choose a CRAN mirror, for instance `USA(MI)`.

To make use of RExcel for interactive data analysis between ***R*** and Excel, you need to perform several more steps. First download RExcel directly from

http://rcom.univie.ac.at/download/RExcel.distro/RExcelInst.latest.exe Then you need to manually download two ***R*** packages: rcom and rscproxy. You can use the following command to install the packages and use a function in the package loaded:

```
> install.packages('rcom')
> library(rcom)
> comRegisterRegistry()
```

Rexcel may also work if you download and install the latest `"statconnDcom"` from http://rcom.univie.ac.at/download/current/statconnDCOM.latest.exe. If RExcel has been installed correctly, when you open Excel and choose "Add-Ins" you will find the add-on "RExcel".

1.2 Data importing

From Excel, click "R Start" in the RExcel to open ***R*** GUI in the background. You can also open ***R*** GUI first, then open Excel and click the "connect R" item in the RExcel menu. To import the data, highlight the part of the Excel data set you would like to input as an R data frame, then click the "Put R Var" and "Dataframe" option in the sub-menu as shown in Figure 1.1. Click "with rownames" if the first column consists of the names of the samples or sites. Make sure the box "make active in RCommander" is not clicked if you do not have the ***R*** package `RCommander` installed.

Note that in multivariate data analysis, the input data-frame is usually expected to have one row per sample collected, with data values about that sample represented in various columns. Therefore samples(or sites) correspond to *rows*, while group information, spatial coordinates, and all attributes (measurements or, variables) pertaining to those samples will be contained in the *columns* of the data frame.

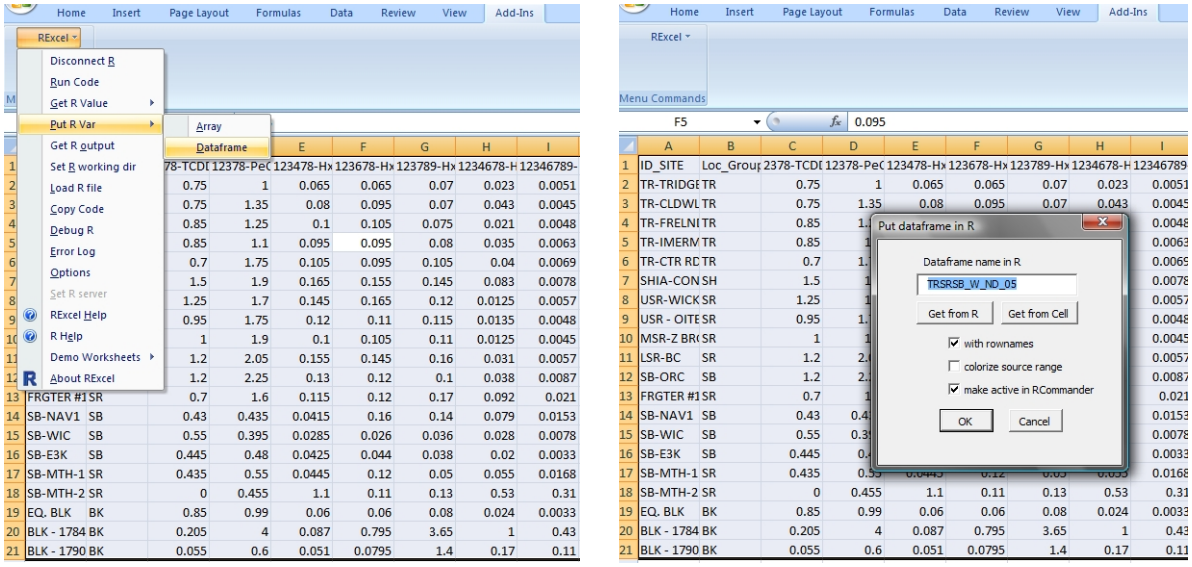

figure 1.1: Import data in Excel sheet into R data frame using RExcel

After you import the data frame from Excel into ***R***, you can go to ***R***GUI and type the following command into the ***R*** console to list the objects now available for use in the ***R*** workspace:

```
> ls()
```

If the import was successful, the name you have just typed for your data frame will be in the list returned by the `ls()`command. If it returns `character(0)`, then it means there are no objects in the current workspace and the data frame has not been successfully imported yet. If you see the name of your data frame in the list, then the next step is to set the working directory in R to the folder where you placed the PCCAT script files.

You can do that by clicking the File menu and selecting "Change dir". That will bring up a dialog box you can use to select the folder where you put the script files.

After getting the data frame into ***R*** and setting the working directory, you are ready to use PCCAT . To do that, you need to run an ***R*** script file. In ***R*** GUI, click "File" and then "Source R code". In the resulting dialog box, choose the file "pccat.r", then click the Open button. The result should look like this:

```
 Wed Nov 11 09:31:47 2009

 ***************************
      Welcome to PCCAT!
 ***************************

 Please press Enter key to continue...
```

Once you press enter, you will get a list of the objects in the current workspace, one of which should be the data frame object you imported previously with RExcel. Although RExcel is highly recommended for users because it preserves the data form, PCCAT will accept three ways of importing data:

```
The objects in the current workspace are
[1] "TRSRSB_W_ND_05"
 what data set do you want to use?
 1. I want to type the name of the object
 2. I want to use the example data set in the tutorial
 3. I want to import from csv/txt file
```

If you want to use the data set, for instance "TRSRSB_W_ND_05.xls" imported from RExcel, you can type `TRSRSB_W_ND_05`; PCCAT can also provide 5 training data sets, if you choose 2:

```
which example data set do you want to use?
1.TRSRSB_W_ND_0        2.TRSRSB_W_ND_05    3.T_BASF_Adjusted
4.T_Trenton_Adjusted   5.T_Trenton_Given
```

These are provided as examples of the types of data formats that MDEQ clients will most commonly bring to PCCAT for multivariate data analysis.

PCCAT will then show you a list of the column names (variables and attributes) in the data frame and row names (names of the samples or sites) from your data frame. Examine these lists to confirm the data set was imported correctly.

2 Data Preprocessing

Next, PCCAT asks you to specify potential grouping and geographical information to be extracted. It is expected the remaining data matrix consists only variables(or chemical parameters). Cells with letters may be treated as censored data.

```
1. I want to specify the grouping information
2. I want to specify the spatial coordinates
3. I want to select/drop some columns
4. I want to continue
```

You can select these options repeatedly until you choose to continue. PCCAT provides a customized input style for users, which will be quite useful for the following procedures. To select more than one number, the input is expected to be separated by spaces rather than other characters like commas. For instance, if you want to choose columns 2 and 3, you may simply type "2 3"; you can also type "1:20" to choose from 1 to 20. You can furthermore type:"1:10 15 20:25" and any such combinations.

The data preprocessing step in PCCAT mainly consists of two steps. It first extracts the matrix from the imported data set, appropriately handling things like censored, non-detected or missing data, etc. and then at the second step it can apply popular and useful transformations for data for different objects.

2.1 Pre-preprocessing

Generally the data types are summarized as nominal, ordinal, ratio scale, symmetric or asymmetric binary data [13] [2]:

- **binary:** categorial data with values of 0 or 1. For instance, "male" and "female". Symmetry means that the outcomes will be treated as equally important, whereas asymmetry means that one outcome is more important than the other(eg,live versus dead).
- **nominal:** integer representation of more than two values of categorial data. For instance if there are five groups for different locations, we can convert them into integers "1,2,...,5".
- **ordinal:** alike nominal, but the order of integers assigned matters. A typical example is in sampling survey: strongly disagree = 1, disagree = 2, ... , strongly agree = 5. The position or order has specific meanings.
- **interval scale:** typically when the unit of data becomes not important, we treat for instance 100 versus 30 the same as 10 versus 3.
- **ratio scale:** the most common data type with strongest numerical meaning, where the intervals keep the same importance throughout the scale.

Although data types with strong numerical meanings are recommended, all these data types above can be handled either by selecting appropriate dissimilarity measure or by data transformations.

In an environmental data set, typically there will be truncated or censored, missing and notational data types. Dealing with such data types greatly depends on the data set itself and particular purpose or criterion. You need to extract a patter matrix with strong numeric properties. Therefore these data types must be handled properly in the preprocessing steps.

2.1.1 Missing Data

In a data frame, cells containing no numeric information will be detected as missing data in PCCAT , such as "NAs", blanks and " $< R$", etc. Although there are many discussions about handling missing data, PCCAT applies the simplest way: excluding either rows or columns containing the missing data. However it will sometimes cause significant loss of information. A potential alternative to make full use of such information is data imputation, i.e. by treating the missing counts as censored data with a very large detection limit(DL) and then applying the techniques in the following context.

2.1.2 Flagged Data

The data sets to be brought to PCCAT are most likely censored, since it is a common problem in environmental data and geochemical data. One such example in our training data set in PCCAT is "T_Trenton_Given". Table 3(See Appendix) shows a small fraction of that data, which contains a certain portion of left censored data with multiple detection limits. Millard[20] also provides a table for commonly used data qualifiers:

- for J, the result is of limited use due to dicrepancies in holding times, blank analysis, duplicate analysis, spike analysis, or laboratory contamination problems;
- for L, the result is of limited use because it is between the instrument/method detection limit and the contract detection/quantitation limit.
- N means the result is probably acceptable but it is just outside the calibration range or the recovery is just outside the specification range.
- R means the results are unusable due to discrepancies in analytical technique/protocol, improper calibration, outside calibration range, outside specified recovery windows, or blunder.
- U indicates that reading was below instrument detection limit , or method detection limit.

For the data sets provided by MDEQ, only "U" and $<$ will be treated as flags for censored data; flag "R" will be discarded and the rest will be treated as estimated data for direct use. Garrett et al[5] lists the following options as potential solutions for the detection limit problem:

- Delete the whole variable(or chemical parameter) or all samples with values "<DL" from data analysis;
- Mark all observations "<DL" as missing;
- Model a distribution in the interval [0,DL] and assign an arbitrary chosen value from this distribution to each sample;
- Predict a value for this variable in each sample via multiple regression (imputation) techniques using all other analytical results; or
- Set all values marked "<DL" to an arbitrary lower number.

As [5] mentions, none of these solutions always has the best performance. For instance, the last item with the lower number set to half(or $1/\sqrt{2}$ sometimes) of the DL is often suggested because it is the simplest imputation. Nevertheless EnvironmentalStat module for Splus[20] mainly assumes specific distributions(normal or log-normal) with parameters estimated or compared to the ECDF(empirical cumulative distribution function). [7] has a full discussion and provides a detailed table of recommended methods for estimation of summary statistics for different types of data. Aiming at multivariate analysis for the following procedure, it is recommended to fill or impute the cells to enlarge the pattern matrix. Therefore, PCCAT provides the following procedure to handle censored data: At the first step, PCCAT detects and makes a summary of cells with characters other than numbers; then for each chemical parameter the percentage of censoring data will be displayed. You can then specify a threshold to filter those with high censoring rate. Next, there are three options for data imputation:

```
for data imputation:
1. I want to assign the half of the detection limit(DL)
2. I want to fit a normal/lognormal distribution with ML estimation
3. I want to fit the ECDF with Kaplan-Meier method
```

Method 2 first obtains the *maximal likelihood(ML)* estimates of the parameters for assumed model(whether it assumes a normal or log-normal distribution is decided by a normality test with a threshold p-value of 0.1); then for each censored cell a random sample is drawn from the fitted distribution until it is less than the DL. Such trial will be done at most 100 times, after which the half of the DL will be used instead and it will return a warning message. This approach may become inaccurate when the size of uncensored counts is small. The result for T_Trenton_Given Data Set is shown in Figure

2.2, where the line represents the fitted normal distribution with the maximum likelihood using uncensored data in each variable.

Method 3 is a robust nonparametric approach. It estimates the empirical CDF(culmulative distribution function) of the chemical parameter using the *Kaplan-Meier method* and then similarly draws the sample using an Inverse CDF Method but restricted below the DL. The red points in Figure 2.3 represent the data after imputation.

2.2 Transformation

After the preprocessing, the extracted pattern matrix should be numerical for transformation perspective. Data transformation becomes necessary when you are pursuing either of the following two objectives:

- non-statistically, it removes the difference of magnitude or unit especially when the data type is ratio scale. This will be important for obtaining beautiful plots under principle coordinate plane and meaningful principle component analysis results, especially when the chemical parameters or concentrations have quite different orders of magnitude but can be treated as ratio scale data.

- statistically, it aims to enhance the normality of the variables which assumes merits like symmetry, especially when most variable have log-normal distributions which are highly skewed.

After extracting the $n \times p$ numeric data matrix with n samples as rows and p attributes or variables $X_1, X_2, ..., X_p$ as columns from the first step, PCCAT provides several common data transformations:

- **standardization:**

$$Y_j = \frac{X_j - mean(X_j)}{std(X_j)}, \qquad j = 1, 2, ..., p$$

  known as Z-score of the p variables with each subtracting its mean and divided by its standard deviation. A full standardization scales the selected variables to have mean=0 and variance=1, thus reducing the effects of variables with larger variance and getting rid of units. However this is not always necessary, since it may sometimes remove the desired patterns [2] [18]. If there is constant variable, standardization will not apply since the denominator will be zero, which leaves the Z-score undefined.

- **logarithm transformation:**

$$Y_j = \log X_j, \qquad j = 1, 2, ..., p$$

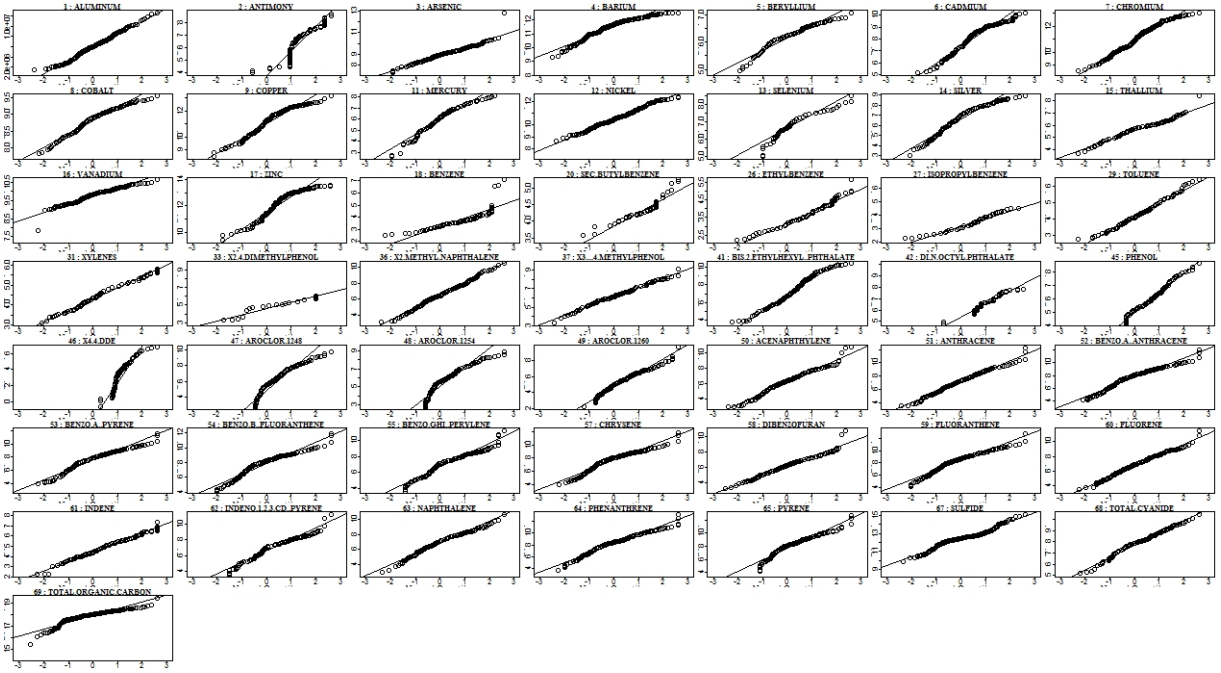

figure 2.2: Probability plot of uncensored data for T_Trenton_Given Data Set

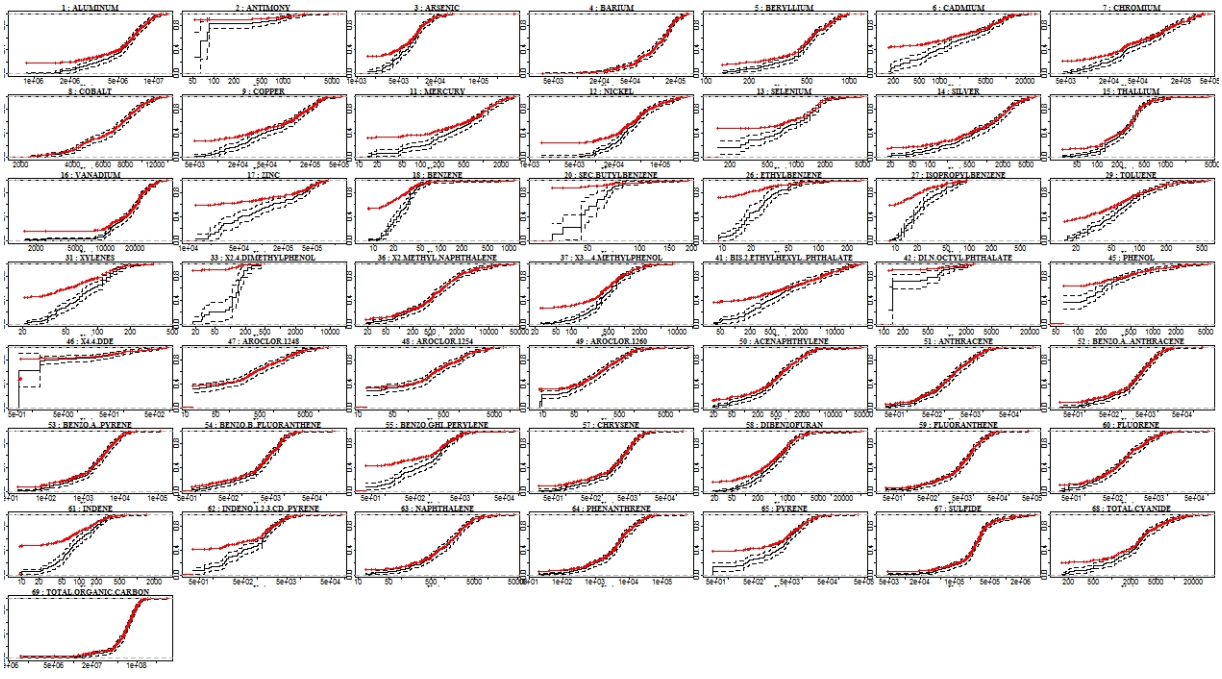

figure 2.3: K-M estimate of ECDF with imputed data for T_Trenton_Given Data Set

applies to data with positive values. it removes different orders of magnitude and hence places the observations on similar scales, which will be quite useful for visualization purposes. However if there are non-positive values in the data frame, this transformation will not apply. For applied geochemical data this is often not the case since concentrations are typically positive and highly skewed(e.g. log-normal), hence can better be transformed to have higher normality.

- **square-root transformation:**

$$Y_j = \sqrt{X_j}, \qquad j = 1, 2, ..., p$$

has an effect similar to a *log-transformation* in that it can make data at different orders of magnitude more comparable, however also applying to zeros. However if there are negative values in the data frame, this transformation will not apply.

- **power transformation:**

$$Y_j = \begin{cases} \dfrac{X_j^{\lambda} - 1}{\lambda(\mathbf{GM}(X_j))^{\lambda-1}}, & if \lambda \neq 0 \\ \\ \mathbf{GM}(X_j) \cdot \log X_j, & if \lambda = 0 \end{cases} \qquad j = 1, 2, ..., p$$

where $\mathbf{GM(X_j)}$ the geometric mean of X_j, i.e. $\mathbf{GM(X_j)} = (X_{1j}X_{2j} \cdots X_{nj})^{1/n}$ is used since it takes advantage of right-skewed(e.g. log-normal) distributions. It is also known as *Box-Cox transformation*. If selected, the power λ will be roughly determined such that the transformed data has approximately strongest normality(measured by Shapiro-Wilks normality test). Again it applies to data with positive values.

- **scale transformation:** The variable selected will be divided by a number specified by the user. It is helpful when the variables are ratio scale and it is desirable to get rid of effects of unit.

To determine a subset of variables for transformations, the empirical culmulative distribution functions for all variables are plotted, which can be compared with the true ECDF of normal and log-normal distributions as shown in Figure 2.5. The associated p-value for the $Shapiro - Wilks$ normality test is also attached. The p-values for the normality test(default threshold: 0.1) with green colors in the plot indicate a strong normality for those variables, while those with red color indicates non-normality, or high skewness, where an appropriate transformation may becomes necessary. For instance, the plot of T_BASF_Adjusted Data Set after Box-Cox transformation of variables that fail to pass normality test is shown in Figure 2.4. You can take different transformations for different subsets since this process will be repeatable.

For one more example, we import the "TRSRSB_W_ND_0" data set without dropping any attributes(variables). The first several columns in the raw data set are shown in Table 1(see Appendix).

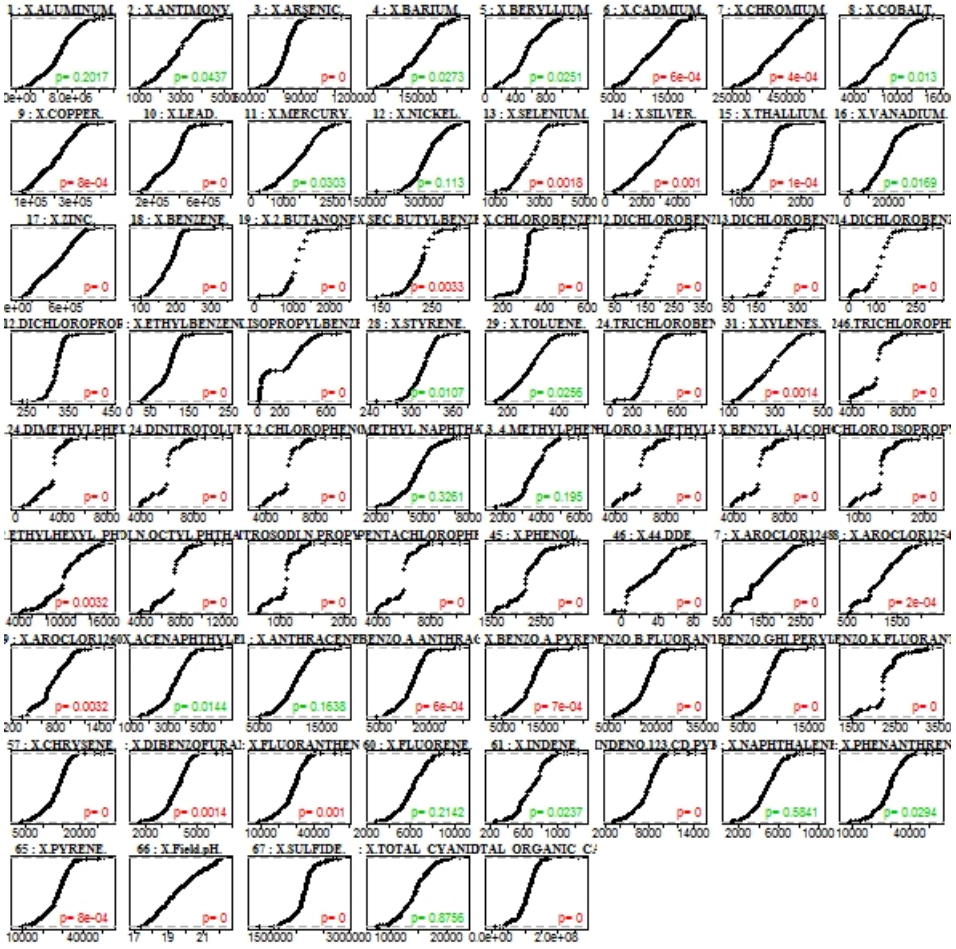

figure 2.4: ECDF for T_BASF_Adjusted Data Set after Box-Cox transformation

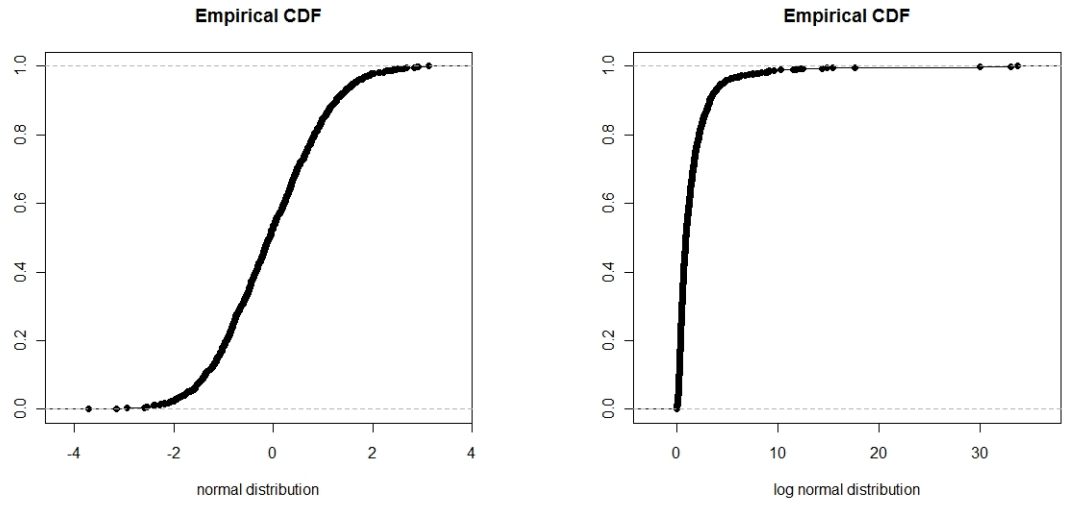

figure 2.5: Example of ECDF for samples from normal and log-normal distribution

```
Choose the type of data transformation:
 1.none  2.standardization  3.logarithm  4.square root  5.box-cox  6.scale
1: 4

 Data preprocessing step is done:
 The location groups are encoded as:
      [,1] [,2] [,3] [,4] [,5]
 [1,] "1"  "2"  "3"  "4"  "5"
 [2,] "TR" "SH" "SR" "SB" "BK"

  The final pattern matrix involve 20 samples and 17 attributes
  Do you want to display it?  y/n
```

By entering "y" you can display and check the extracted pattern matrix for the following multivariate data analysis.

3 Visualization

3.1 An Overview

PCCAT provides several popular and useful visualization techniques for exploratory data analysis. These are scatterplot matrix associated with correlation coefficients and histograms for individual variables, boxplots with outliers detected, segment plots for comparison between samples and variables, and 3D scatter plots for selected variables.

3.2 An example

For instance, we take the "TRSRSB_W_ND_05" data set with first several columns in the raw data set shown in Table 2.

Notice that only the second column (or first variable) contains an entry "0", we drop this column so that we can take logarithm transformation. Then in ***R*** GUI, click "File" and then "Source R code". Choose the directory where PCCAT modules locate and select the file "visualization.r", PCCAT will first print the detected attributes, or variables. Then similarly it asks users to select or drop some or keep all of them. It is usually recommended to select a few of the variables to make it clear for scatterplot matrix, particularly when the data frame is fairly large. Here we drop the first and third variables for example. For correlation detection, PCCAT asks users to specify two thresholds for low and high correlation between each two variables in the scatterplot matrix. We choose "0.15" and "0.9" for example.

The scatterplot matrix and how to usually save a graphic object in ***R*** are shown in Figure 3.6. In the symmetric pair plots the samples are colored with given group information. The diagonal shows the name of each variable with and histogram. Those

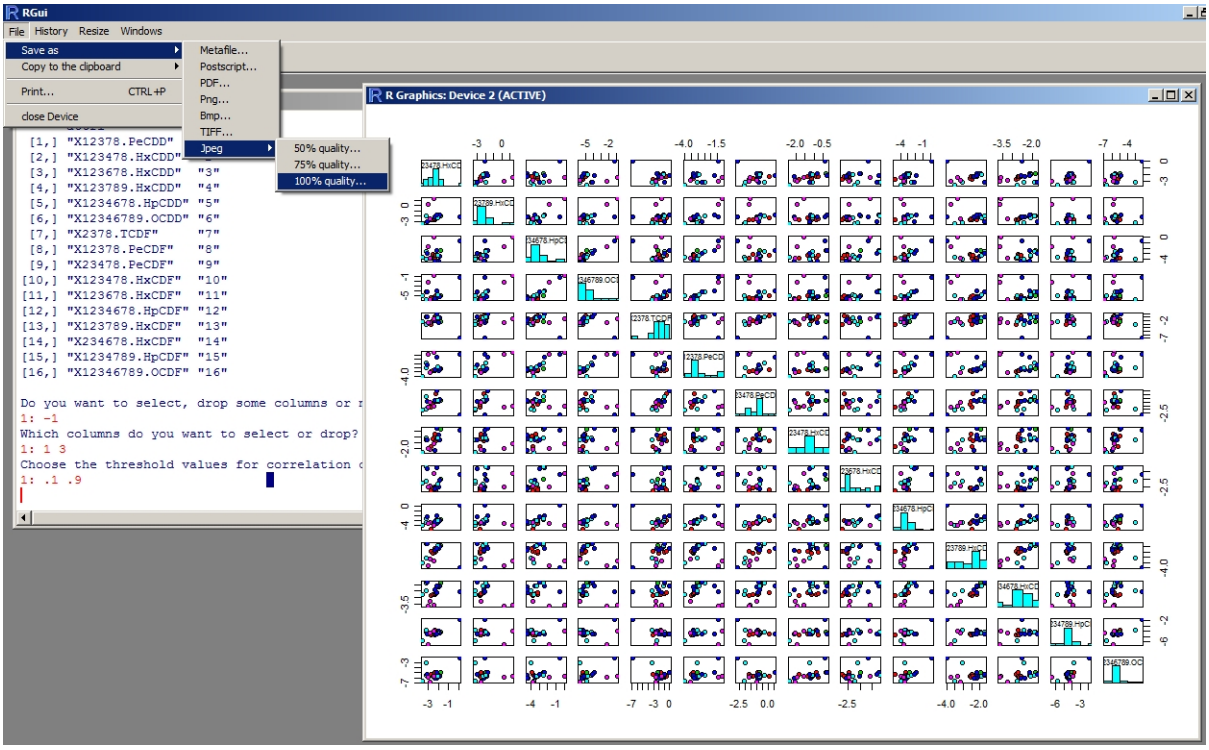

figure 3.6: How to save a graphic output in ***R***

with histogram like bell shapes may come from normal distributions, while those with highly skewed histogram are suspected from other distributions, like the log-normal. We can plot their empirical cumulative distribution and test the assumption as [20] typically does. The scatterplot matrix with both histogram for individual variable and correlation coefficients between variables is shown in Figure 3.8. According to the thresholds proposed, those with correlation coefficients less than 0.15 are colored in blue and those higher than 0.9 are colored in red. For instance, the forth and fifth variable in this plot has strong correlation with coefficient 0.94. Correspondingly, the scatter plot with the two has an obvious linear trend.

Figure 3.9 shows the Box plot for selected variables. With the lower quartile (Q1), median (Q2), upper quartile (Q3), most extreme observations within a distance of 1.5 of the upper and lower quartiles manifested for each variable, Box plot can be fairly helpful for summarizing individual variable and detecting outliers. The dots beyond the whiskers denote outliers and the maximal outlier values are shown. PCCAT also specifies those samples as outliers with respect to individual variables.

```
Outlier detection:

 X123478.HxCDD :              TR-IMERMAN_PK
 X123789.HxCDD :              BLK_17841        BLK_17905
 X1234678.HpCDD :             BLK_17841
```

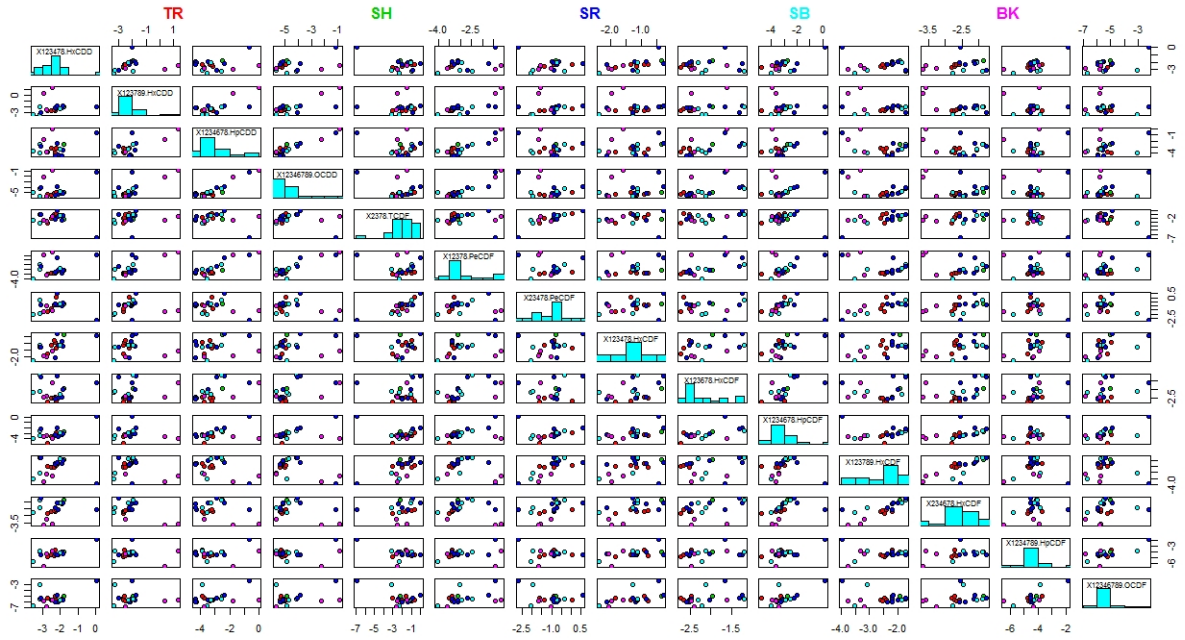

figure 3.7: Scatterplot with histogram and groups information

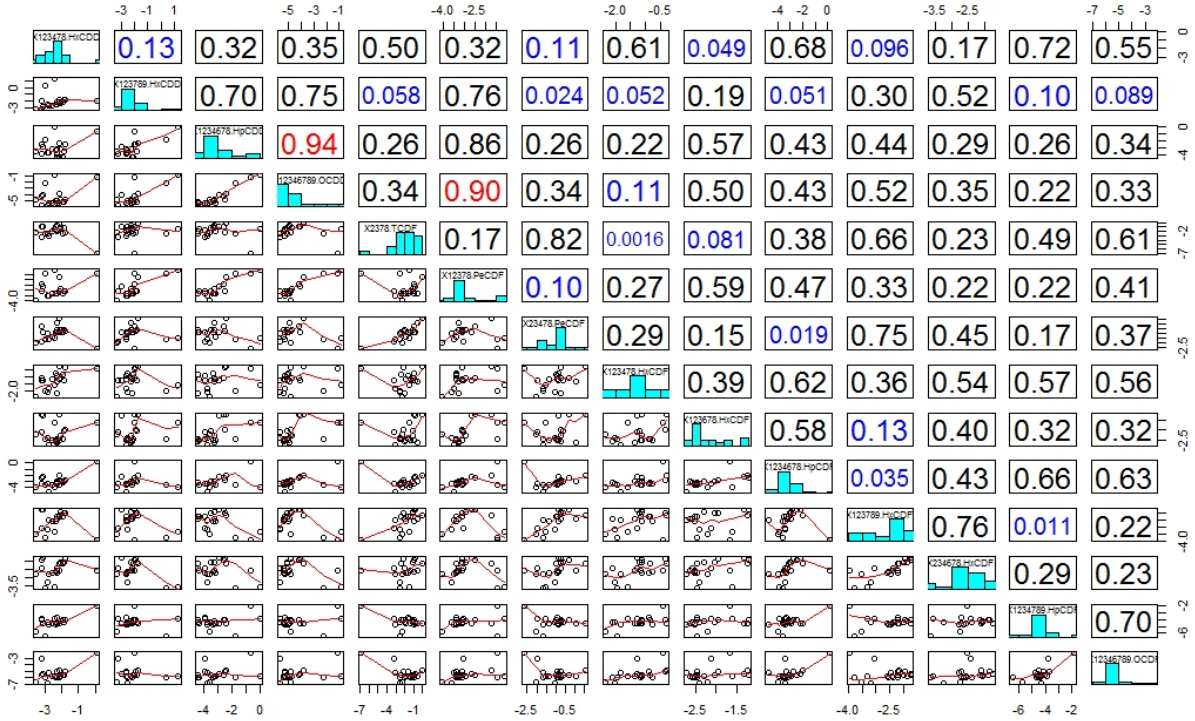

figure 3.8: Scatterplot with correlation coefficients and smoother

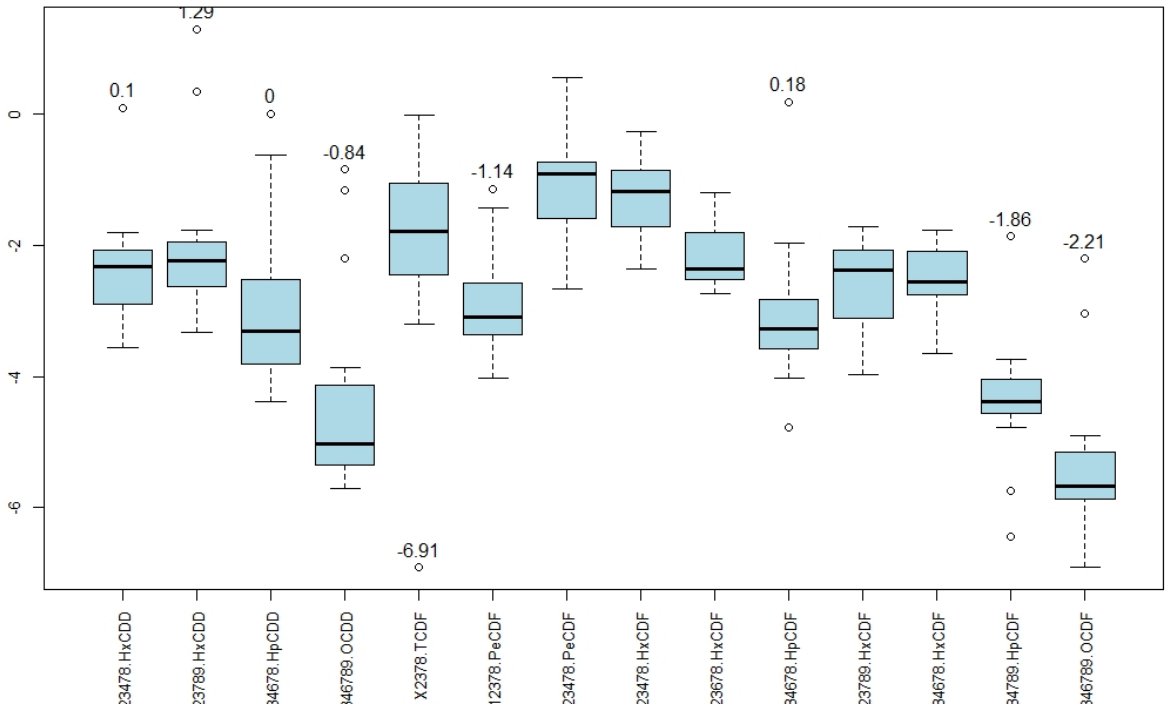

figure 3.9: Boxplot with maximal outlier values imposed on

X12346789.OCDD :	SB-MTH-2	BLK_17841	BLK_17905
X2378.TCDF :	SB-MTH-2		
X12378.PeCDF :	BLK_17841		
X1234678.HpCDF :	TR-TRIDGE	SB-MTH-2	
X1234789.HpCDF :	SB-WIC	SB-MTH-2	BLK_17905
X12346789.OCDF :	SB-E3K	SB-MTH-2	

A segment plot, on the other hand, gives a graphic description with respect to individual sample, as shown in Figure 3.10 for our example. Each segment diagram represents one row (site) of the input x. Variables (columns) start on the right and wind counterclockwise around the circle. The size of the column is shown by the distance from the center to the point on the star or the radius of the segment representing the variable.

Figure 3.11 shows the 3D visualization for selected three attributes (For instance we choose variables 2,3,4, namely " X123478.HxCDD ", " X123678.HxCDD " and " X123789.HxCDD "). As 2D pair plot, each dot denote a site colored with given group information. You can drag to change angle of view, and scroll the graph to zoom in or zoom out. PCCAT also allows users to display the site names on the dots.

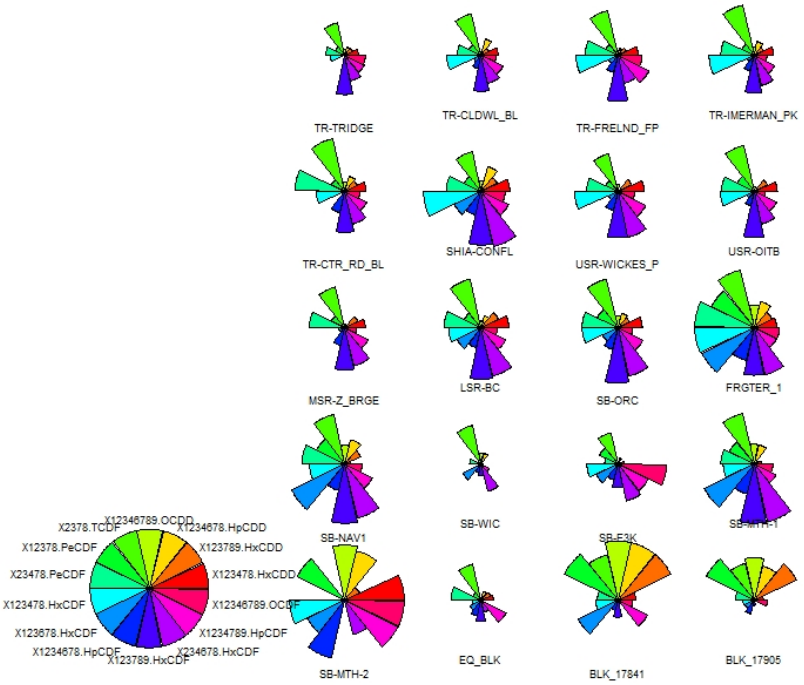

figure 3.10: Segment plot

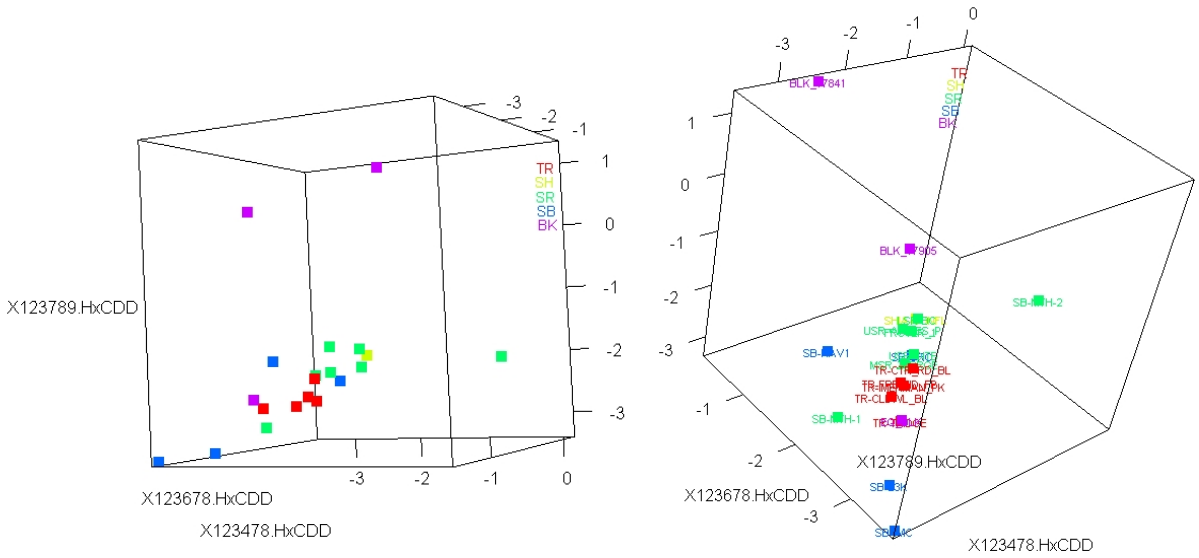

figure 3.11: 3D scatter plot for selected variables

4 Principle Component Analysis

4.1 An Overview

In the visualization step, PCCAT provides 2D or 3D plots to visualize the samples with respect to selected variables. However usually the variables are correlated, and it is not quite clear which combination of variables (as there are so many) best capture the patterns underlying the data. We naturally consider transformation of variables to get rid of their correlations while utilizing as much information as possible. This is exactly what the Principle Component Analysis (PCA) is designed to do. A strict but friendly mathematical definition of principle components as well as some procedures involved in PCCAT is introduced in [21].

4.2 An example

To better interpret PCA procedures, we use the same example in the visualization step. As before, in ***R*** GUI, click "File" and then "Source R code". Choose the directory where PCCAT modules locate and select the "principle component analysis.r" module.

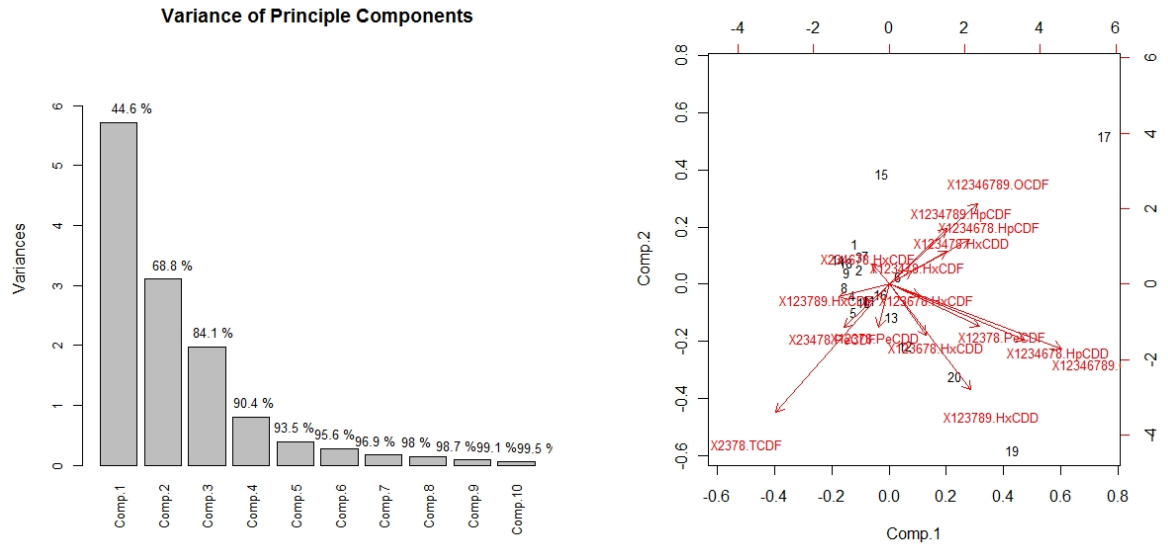

figure 4.12: Bar plot and Biplot of Principle components

As shown in Figure 4.12, the bar plots in the left plane are the variances for each Principle Component or PC (did not show all of them since the rest have very small variances compared to the first several PCs), i.e. the squared value of the first row in the table shown in ***R*** GUI. The *cumulative* portions of variance are displayed on the bars. For instance for Component 4 the number is 90.4%, which means that the variances of the first four components sum to 90.4% of the total variance, or they can be interpreted

as representing 90.4% of the information in all the variables. PCCAT will correspondingly print the importance matrix of components in the ***R*** GUI.

The plot shown in the right plane is using the first two PCs as coordinates, with all samples denoted as numbers and more importantly, the projections of all variables on this plane. Those with longer arrows and more parallel to one of the two PCs are expected to have higher contribution, or loading for that PC. Accordingly PCCAT allows sorting all variables' *absolute* loadings by user-specified component. In our example we choose PC one since it contains the most information:

```
 sort
the variables by its loading on which principle component? 1: 1
 variable        PC1    PC2    PC3    PC4
 X12346789.OCDD 0.533  -0.271 -0.094 0.163
 X1234678.HpCDD 0.421  -0.234 0.002  0.168
 X2378.TCDF     -0.351 -0.539 0.365  0.229
 X12378.PeCDF   0.279  -0.178 0.066  0.124
 X12346789.OCDF 0.273  0.34   0.211  0.082
 X123789.HxCDD  0.252  -0.444 -0.116 -0.294
```

As we shall see, the first several variables, such as X12346789.OCDD(name not shown entirely), X1234678.HpCDD and X2378.TCDF, have the longest projection in the horizonal line (PC 1) as shown in Figure 4.12.

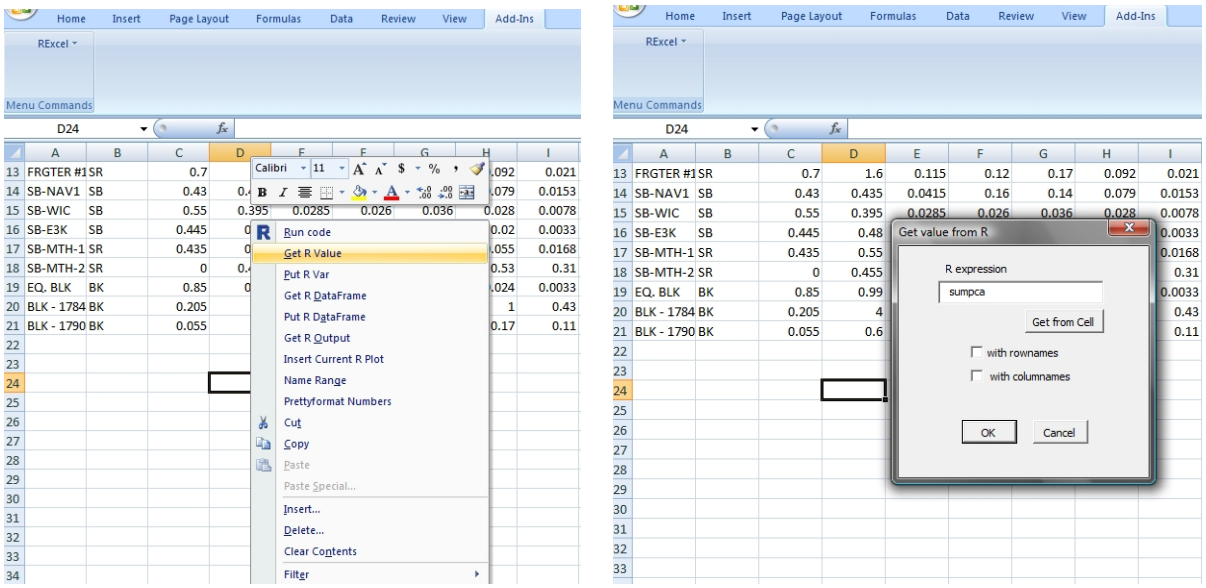

figure 4.13: Directly get R values in Excel with RExcel

To output the Principle Component Analysis results, PCCAT stores some variables in the workspace for users to directly obtain the table form using RExcel as shown in Figure 4.13. Type the R expression "`sumpca`" to obtain the summary of all PCs with information about

variances, proportions; and type "`loadMat4`" to get the sorted loading matrix for the first 4 PCs by each variable.

We can also use the same strategies for visualization of Principle Components as shown in Figure 4.14 and 4.15. PCCAT provides flexible manipulations for users to have different trials. Since PCs have nicer properties than the raw variables, patterns underlying the data are clearer with PCA results. Principle coordinates will also be quite useful for visualizing the clustering or discriminant analysis.

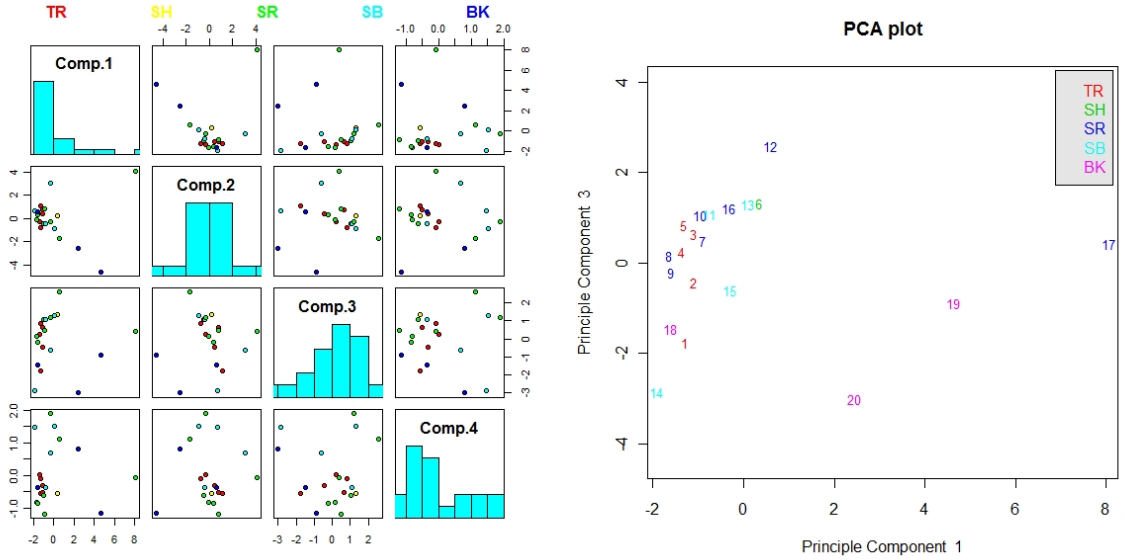

figure 4.14: Selected pairs plot with sample information for Principle Components

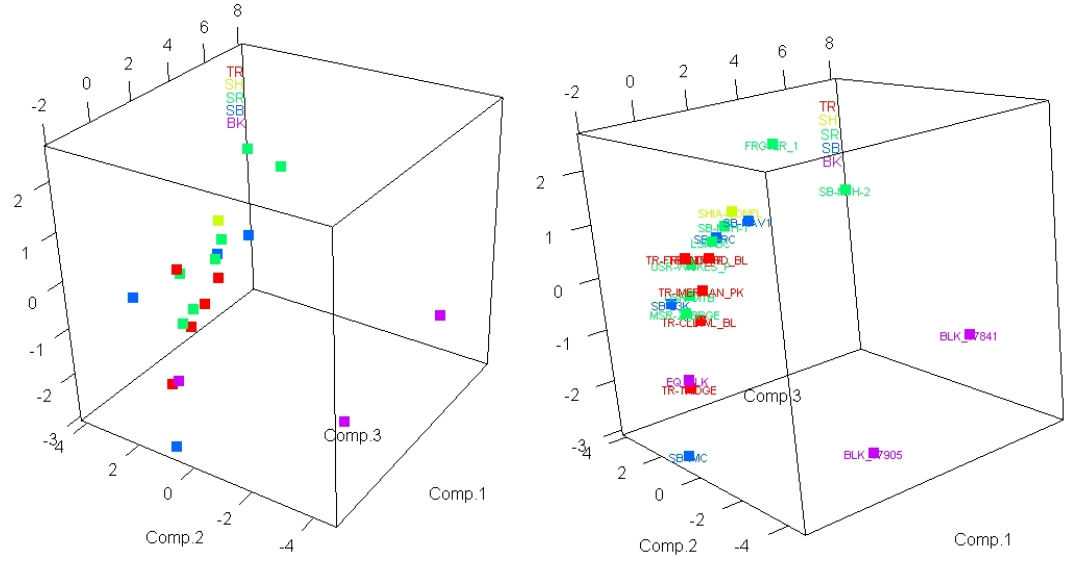

figure 4.15: 3D visualization with Principle Component Coordinates

5 Clustering Analysis

5.1 An Overview

Clustering, which is related to unsupervised learning in the pattern recognition literature [19], is used to group or label those samples (or sites) which are most similar to each other to capture patterns in the data. The statistical software ***R*** has many built-in functions and packages to implement clustering analysis. [13] and [12] are the main references for the basic package `cluster`, while some useful summaries and tutorials can be found in [14] [9] [16]. The most common and popular methods in practice are hierarchical and partitioning clustering methods. However model-based clustering [3] is becoming more popular according to the dramatic development of Bayesian models and methods. Fuzzy clustering, essentially a variation on k-Means clustering [13], nevertheless produces "fuzzy" rather than "hard" clusters, bearing somewhat different characteristics. Detailed algorithms of these methods can be found in [4].

5.1.1 Selection of Methods

If there is no particular purpose for determining specific clustering methods, it is strongly suggested to adopt several algorithms for comparison and conclusion. To achieve this, PCCAT is designed to allow replications of options at each step. However for specific purpose, it is worth mentioning the objects or characteristics for each clustering methods:

1. **Hierarchical clustering** can be implemented without specifying the number of clusters. This can be useful if you are not sure about how many cluster to be achieved. It will also produce an ordering of the objects and a dendrograms which is informative for data display. Notice that at some hierarch, if two samples are grouped, then they will not be separated at the following stage. Sometimes this trait causes so-called chaining effect [8], though it can also be judged to be right the desired pattern [13].

2. **Partitioning clustering** achieves a partition of all samples into k clusters where the number is determined ahead of its implementation. At each stage there are k clusters, with each object belonging to exact one cluster. The partitioning algorithms achieve reasonable k clusters by flexibly changing the k centroid to minimize typically the within-cluster sum of squares.

3. **Fuzzy clustering** assigns k probabilities, which sum to 1, to each sample to evaluate the membership belonging to the k clusters. That is, each sample has some probability of belonging to each cluster. The probabilities sets are achieved by minimizing some objective function. It can also incorporate some spatial information with the expression of objective functions [6].

4. **Model-based clustering** assumes that the samples are observations from finite distributions(typically normal). Based on the finite mixture model assumption, it implements the famous EM algorithm to estimates the parameters and allocates the samples with Bayesian decision rule. The number of clusters and models are simultaneously determined according to Bayesian Information Criterion [3] [1].

Generally, hierarchical or partitioning clustering methods are popular and you need select an appropriate distance measure. Fuzzy and Model-based clustering methods provide the probabilities in addition but are two totally different stories. but they are similarly better for smaller number of variables. passing a large number of variables, the two methods may result in trivial outcome.

5.1.2 Selection of Distance measure

Various measure of distance need consideration when you are using hierarchical or partitioning method. Like the selection of clustering methods, the selection of measures of similarity or dissimilarities between samples can also cause significantly different clustering results. There are numerous such measures [8], according to different types of data. Distance is one such measure for dissimilarity. Other useful measures for similarities can be the correlation coefficients. For particular use, PCCAT has chosen following list of distance measures:

- **euclidean:**
$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$
most commonly used measure for distance. ranging from 0 to infinity. Sometimes it is not in favor of if some are too large, causing the rest to huddle together.

- **maximum:**
$$d_{ij} = \max_{k=1,2,\ldots,p} |x_{ik} - x_{jk}|$$
or Chebyshev distance, which is the maximum distance between two components of feature vector for the pair of samples.

- **manhattan:**
$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

- **canberra:**
$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$$
Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing . It examines the sum of series of a fraction differences

between coordinates of the pair of samples. This measure is bounded since each term of fraction difference has value between 0 and 1. If one of coordinate is zero, the term become unity regardless the other value, thus the distance will not be affected. However this distance is very sensitive to a small change when both coordinate near to zero.

- **minkowski:**

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^p \right)^{1/p}$$

This is more general p-norm distance. Notice that the former three, *manhattan*, *euclidean* and *maximum* are all special cases of this distance, with $p = 1, 2$ and ∞ respectively. If this distance is chosen, p need to be specified. The unit circles ($d = 1$) with distinct p are depicted in Figure 5.16 (From Wiki):

figure 5.16: unit circles with various values of p

- **gower:**

$$d_{ij} = \frac{1}{p^\star} \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{\max x_k - \min x_k}$$

where $p^\star$ is the number of columns excluding missing values and $R_k = \max x_k - \min x_k$ is the range of kth variable with the extreme value over all non-missing samples. This is originally proposed by Gower [10] and later developed [13] to deal with mixed variables with different data types.

- **bray:**

$$d_{ij} = \frac{\sum_{k=1}^{p} |x_{ik} - x_{jk}|}{\sum_{k=1}^{p} (x_{ik} + x_{jk})} = 1 - \frac{2 \sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} (x_{ik} - x_{jk})}$$

The Bray-Curtis dissimilarity coefficient has the range $[0, 1]$ where $d_{ij} = 0$ indicates the maximum similarity between the pair of sites. To use this coefficient, the input data matrix should be non-negative. Thus it will not apply to standardized data. Also for a single column of data matrix there should never be more than one zeros. Otherwise the denominator will become 0 and it is undefined.

5.2 Hierarchical Clustering

PCCAT provides most popular agglomerative hierarchical clustering methods. In hierarchical clustering procedure, given the selected distance measure, many algorithms, known as different linkage criteria, for merging groups at each step are famous and popular, such

as `ward, single/complete linkage, UPGMA`, etc. Basic ideas about the linkage criteria are table below:

algorithm	name	mathematical definition
`single`	Minimum or single-linkage clustering	$\min\{d(a,b) : a \in A, b \in B\}$
`complete`	Maximum or complete linkage clustering	$\max\{d(a,b) : a \in A, b \in B\}$
`average`	Mean or average linkage clustering, or `UPGMA`	$\frac{1}{\lvert A\rvert \lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d(a,b)$

`ward` however is seeking for a minimal error sum of squares at each stage merging two groups, known as sum-of-square criterion.

To illustrate, we follow our former example of "TRSRSB_W_ND_05" data set after dropping the first variable and taking logarithm transformation. For instance we choose "`gower`" at the distance selection step and "`ward`" at the algorithm selection step. PCCAT provides two methods to visualize the hierarchical clustering results, Dendrogram and Principle Component Coordinates plot, as shown in Figure 5.17. In the Dendrogram PCCAT also imposes the cophenetic correlation coefficient which measures how the clustering results distorts the input data information [8].

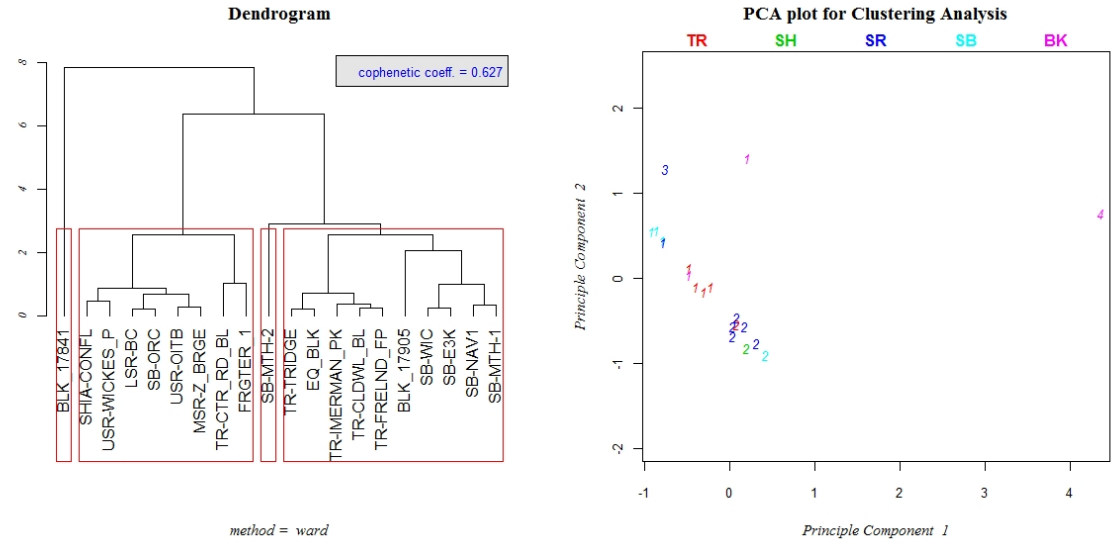

figure 5.17: Dendrogram with hierarchical clustering

Since agglomerative hierarchical clustering methods are sensitive to the selection of distance measure and linkage criterion, PCCAT provides a flexible combination for users to choose and repeat the procedure. Also for Principle Component Coordinate plot, the flexible and repeatable choices can be convenient to seek for the best angle of view for the results associated with the given group information.

5.3 Partitioning Clustering

To implement partitioning clustering methods, the number of clusters to achieve k should be specified. PCCAT provides a plot of the within-cluster sum of squares for users to make tentative determinations. To illustrate, using the previously extracted data(Table3), we select "`euclidean`", the most common distance measure, to implement the well-known K-means algorithm, which aims to partition the points into k groups such that the sum of squares from points to the assigned cluster centres is minimized. As shown in the left panel of Figure 5.18, there is a considerable cut-down of *within sum squares* if we choose $k = 5$. The plot also shows the reduction percentage we achieve and the slope of the line down to this number. We then specify $k = 5$ and implement the K-means partitioning clustering algorithm. It is relatively flexible in that the centroid for each group can be changed at every stage so that "incorrectly" grouped objects may still be separated. As an example, Figure 5.18 shows the process of selecting the number of clusters for partitioning and corresponding results for log-transformed T_BASF_Adjusted data set(Table 5).

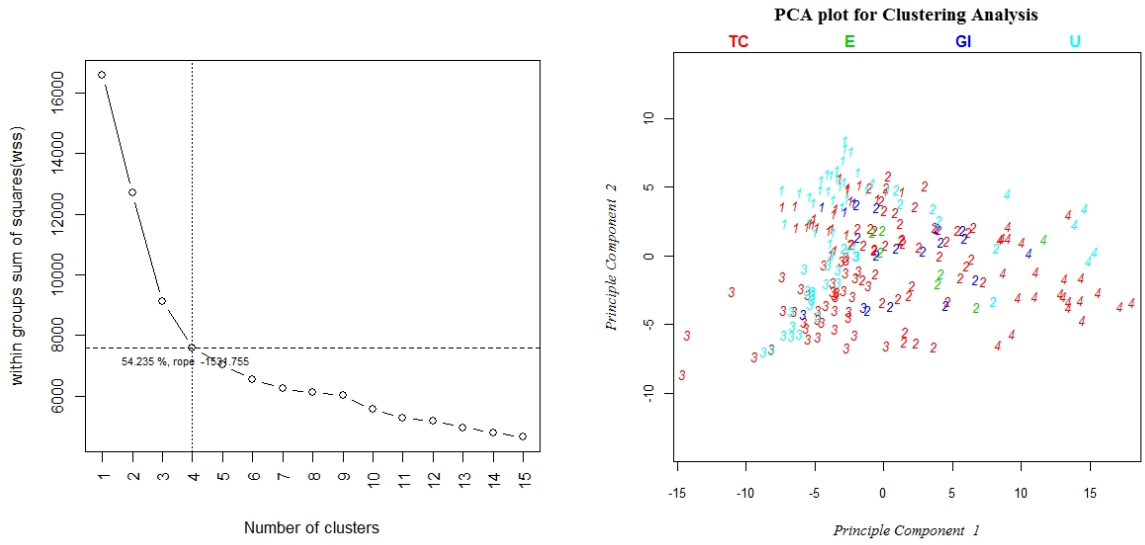

figure 5.18: Selection of number of clusters and PCA plot of results

5.4 Fuzzy Clustering

As a generalization of partitioning clustering, fuzzy clustering bears many features like the specification of number of clusters. For instance we use "T_BASF_Adjusted" (Table 5) data set for fuzzy analysis. To achieve non-trivial results we choose the first 12 variables for interpretation after dropping those with missing data and take log-transformation. The result is shown in Figure 5.19. Rather than a simple number, a segment containing information of membership to each cluster is superimposed. In addition to the clustering results(the largest component in the segment), we see clearly more information about the

trends of cluster patterns. The results will becomes "fuzzier" if larger number of clusters is specified.

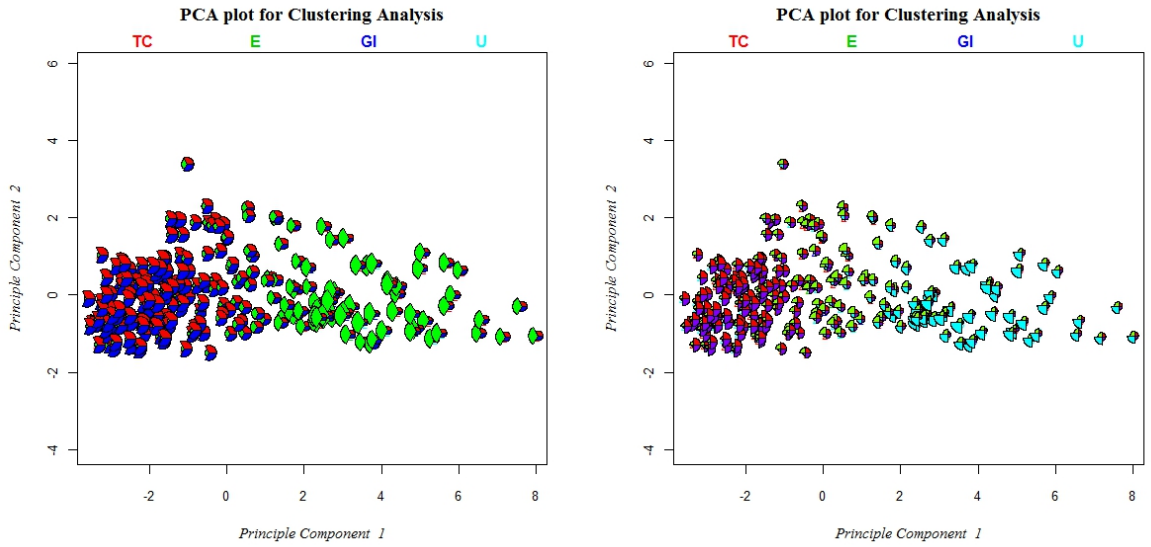

figure 5.19: PCA plot of fuzzy clustering results(*Left*:k=3 *Right*:k=4)

5.5 Model-based Clustering

PCCAT provides a foreground for model-based clustering analysis based on *R* package `mclust`. We take the same example we used for fuzzy clustering analysis for illustration. It informs the best model is "elliposidal, equal variance with 9 components". A detailed interpretation of candidate models can be found in [3]. The a plot window is activated for users to click. Consequently the BIC(Bayesian Information Criterion) for all candidate models, classification and uncertainty results will be plotted. As shown in Figure 5.20, the "EEE" which corresponds to the best model reported with 9 components (or clusters) attains the maximal BIC and therefore is selected. Notice that the number of components considered is ranging from 1 to 9 and usually the number of components for the best model attaining its maximal before 9 will be more meaningful. The clustering result is obtained from the maximal probability evaluated at each site, as shown in the principle component plane.

5.6 Assessment of clustering results

All clustering methods eventually assign labels to each sample. We therefore can assess and compare the results with common measures like the famous *Sum-of-Squared-Error* criterion where we expect a smaller within-sum-square and a larger between-sum-square for a sound clustering result. The sum-of-square-errors is equivalently the sum of the

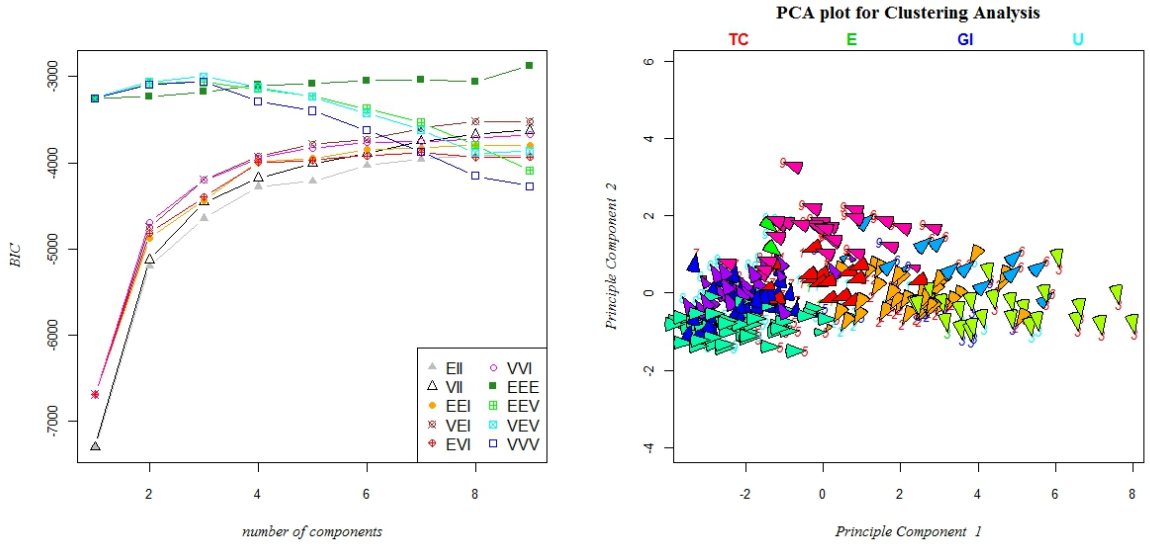

figure 5.20: summary of BIC and PCA plot of model-based clustering results

trace of within-cluster scatter matrix S_W and there are several more criteria about S_W and between-cluster scatter matrix S_B[19]. PCCAT will provide the within-sum square or trace of S_W(small), between-sum square or trace of S_B(large), $|S_W|$(small), trace of $(S_B+S_W)^{-1}S_W$(large) and $\frac{|S_W|}{|S_W+S_B|}$(large), where $|\cdot|$ indicates the determinant of a matrix, and the case in the parenthesis is in favor of the clustering results. These measures will be summarized whenever a clustering result is achieved in PCCAT .

Appendices

A Acknowledges

The author is sincerely thankful to Steven Piece, who invited me to participate in this project via CSTAT so that I was able to work on PCCAT and have such a great hands-on experience. He indeed helped me in writing this manual and shared many appealing ideas in both languages of statistics and English. I also highly appreciate the MDEQ clients. Working with them turned out to be an interesting and rewarding experience; Arthur named the PCCAT tool and found the powerful tool RExcel; Dale helped me further understand environmental data with specific features. They all gave very good advice for the coding work and design of PCCAT. The meetings and discussions with them greatly prompted me to step further into related topics and bring new ideas, useful tools and solutions to this module.

Furthermore, I am greatly indebted to the faculty members in the Department of Statistics and Probability, MSU, particularly Dr.Dass, Dr.Lim, Dr.Finley and Dr.Maiti for sharing techniques and knowledge on statistical computing and Bayesian inference which indeed become the best treasure and milestone to my future efforts. Working with them become most rewarding experience ever. Some of the modules are direct applications from the course that Dr.Dass taught me. He also encouraged me to read several books in the bibliography of this manual and gave me some ideas about EDA(Exploratory data analysis) and Bayesian Inference. Dr.Lim has also greatly helped me in coding with ***R*** and generously shared her work that greatly benefited me. I also owe many thanks to Prof.Finley for helping me become familiar with ***R*** programming with LINUX settings and write C++ codes with R interface. It was a rewarding experience of working with him and building the ***R*** package nndiag, where I learned so much for coding PCCAT.

I am also deeply grateful to Dr.Gilliland, Dr.Roy, Dr.Meerschaert, Dr.Wang and Dr.Cui for their kindness and inspiration in my core statistical courses. It was their work that made me greatly interested in statistical methods with applications. The chat and discussion with them become a unforgettable part of my graduate student life and their encouragement will always inspire me to learn more fantastic things for future achievements.

B Figures & Tables

List of Figures

List of Tables

C Example Data Sets

- TRSRSB_W_ND_0 data set consists of 20 samples and 18 variables as shown in Table 1

ID_SITE	Loc_Group	2378-TCDD	12378-PeCDD	123478-HxCDD	123678-HxCDD	123789-HxCDD
TR-TRIDGE	TR	0	0	0	0	0
TR-CLDWL_BL	TR	0	0	0	0	0
TR-FRELND_FP	TR	0	0	0	0	0
TR-IMERMAN_PK	TR	0	0	0	0	0
TR-CTR_RD_BL	TR	0	0	0	0	0
SHIA-CONFL.	SH	0	0	0	0	0
USR-WICKES_P	SR	0	0	0	0	0
USR-OITB	SR	0	0	0	0	0
MSR-Z_BRGE	SR	0	0	0	0	0
LSR-BC	SR	0	0	0	0	0
SB-ORC	SB	0	0	0	0	0
FRGTER.1	SR	0	0	0	0	0
SB-NAV1	SB	0	0	0	0.16	0.14
SB-WIC	SB	0	0	0	0	0
SB-E3K	SB	0	0	0	0	0
SB-MTH-1	SR	0	0	0	0.12	0
SB-MTH-2	SR	0	0	1.1	0.11	0.13
EQ.BLK	BK	0	0.99	0	0	0
BLK-17841	BK	0	0	0	0	0
BLK-17905	BK	0	0	0.051	0	1.4

Table 1: a fraction of TRSRSB_W_ND_0 data set

- TRSRSB_W_ND_05 data set consists of 20 samples and 18 variables as shown in Table 2

ID_SITE	Loc_Group	2378-TCDD	12378-PeCDD	123478-HxCDD	123678-HxCDD	123789-HxCDD
TR-TRIDGE	TR	0.75	1	0.065	0.065	0.07
TR-CLDWL BL	TR	0.75	1.35	0.08	0.095	0.07
TR-FRELND FP	TR	0.85	1.25	0.1	0.105	0.075
TR-IMERMAN PK	TR	0.85	1.1	0.095	0.095	0.08
TR-CTR RD BL	TR	0.7	1.75	0.105	0.095	0.105
SHIA-CONFL.	SH	1.5	1.9	0.165	0.155	0.145
USR-WICKES P	SR	1.25	1.7	0.145	0.165	0.12
USR - OITB	SR	0.95	1.75	0.12	0.11	0.115
MSR-Z BRGE	SR	1	1.9	0.1	0.105	0.11
LSR-BC	SR	1.2	2.05	0.155	0.145	0.16
SB-ORC	SB	1.2	2.25	0.13	0.12	0.1
FRGTER #1	SR	0.7	1.6	0.115	0.12	0.17
SB-NAV1	SB	0.43	0.435	0.0415	0.16	0.14
SB-WIC	SB	0.55	0.395	0.0285	0.026	0.036
SB-E3K	SB	0.445	0.48	0.0425	0.044	0.038
SB-MTH-1	SR	0.435	0.55	0.0445	0.12	0.05
SB-MTH-2	SR	0	0.455	1.1	0.11	0.13
EQ. BLK	BK	0.85	0.99	0.06	0.06	0.08
BLK - 17841	BK	0.205	4	0.087	0.795	3.65
BLK - 17905	BK	0.055	0.6	0.051	0.0795	1.4

Table 2: a fraction of TRSRSB_W_ND_05 Data Set

- T_Trenton_Given data set consists of 255 samples and 69 variables with considerable proportion of censoring data. A small fraction of the data set is shown in Table 3:

Sample ID	ALUMINUM	ANTIMONY	ARSENIC	BARIUM	BERYLLIUM	···
SD-T19-25-0-0.5	5000000	940 J	7700 J	106000	440	···
SD-T19-25-0.5-1	3670000	930 J	6800	84000	270 J	···
SD-T19-25-1-2.1	3390000	680 J	6300	96700	240 J	···
SD-T19-50-0-0.5	3040000	590 J	6300 J	56400	400	···
SD-T19-50-0.5-1	3080000	580 J	5400 J	56900	300	···
SD-T19-50-1-3	2550000	470 J	5200 J	36000	250	···
SD-T19-75-0-0.5	2450000 J	310 J	6600 J	60200	300	···
SD-T19-75-0.5-1	2040000 J	150 J	5900 J	22300	170	···
SD-T19-75-1-1.6	1690000	88 J	9600 J	15400 J	130	···
SD-T19-100-0-0.5	3770000	630 J	6900 J	61000	440	···
SD-T19-100-0.5-1	4590000	460 J	7800 J	64200	560	···
SD-T19-100-1-3	2700000	160 J	3000 J	32100	360	···
SD-T19-100-3-3.9	1980000	140 J	3000 J	25100	240	···
SD-T20-25-0-0.5	6300000	1900 J	**10100**	185000	420 J	···
SD-T20-25-0.5-1	7210000	1700 J	**10100**	207000	460 J	···
SD-T20-25-1-3	6210000	1300 J	**9900**	161000	480 J	···
SD-DUP-02	7350000	1300 J	**10200**	199000	490 J	···
SD-T20-25-3-5	5730000	1200 J	9500	155000	400 J	···
SD-T20-25-5-7	2340000	230 J	4500	22700	150 J	···

Table 3: A small fraction of T_Trenton_Given data set (Part1)

PHENANTHRENE	PYRENE	Field pH	SULFIDE	TOTAL CYANIDE	TOTAL ORGANIC CARBON
18000	**11000**	7.37	239000	**1400**	64400000
5200	**2200**	7.51	160000	**1600**	69900000
21000	**19000**	8.21	165000	**1100**	57700000
24000	**19000**	9.1	198000	**1300**	71500000
26000	**20000 J**	9.02	195000	**1100**	45200000
5400	**2900**	8.63	81400	**460**	37800000
270000 D	**220000 D**	9.51	88200	**2400**	59400000
4100	**1300**	9.77	19300	< 600	15600000
1100	**340**	9.47	< 34200	< 570	5030000
7200	**7800**	10.51	522000	**7400**	73800000
850	**1000 J**	10.69	178000	**18100**	87900000 J
210	**280 J**	11.21	54200	**13900**	61500000
110	140 J	11.01	35300	**15200**	38700000 J
16000	**19000**	7.92	318000	**2300**	93000000
9600	**9600**	8.21	324000	**3100**	67200000
30000 D	**17000**	8.42	298000	**4200**	72300000
24000	**15000**		237000	< 6300 UB	68300000
32000	**24000**	8.85	365000	**3000**	1.03E+08
430	**260**	9.09	84000	< 660 UB	20000000

Table 4: A small fraction of T_Trenton_Given data set (Part2)

- T_BASF_Adjusted data set consists of 232 samples and 73 variables with grouping variables and spatial coordinates. A fraction of the data set is shown in Table 5:

'Sample_ID'	'Groupings'	'Profiles'	'Latitude'	'Longitude'	'ALUMINUM'	'ANTIMONY'	'ARSENIC'
'SD-T19-25-0-0.5'	TC	U	42.23487	-83.1489	5000000	940	7700
'SD-T19-25-0.5-1'	TC	D	42.23487	-83.1489	3670000	930	6800
'SD-T19-25-1-2.1'	TC	D	42.23487	-83.1489	3390000	680	6300
'SD-T19-50-0-0.5'	TC	U	42.23472	-83.1505	3040000	590	6300
'SD-T19-50-0.5-1'	TC	D	42.23472	-83.1505	3080000	580	5400
'SD-T19-50-1-3'	TC	D	42.23472	-83.1505	2550000	470	5200
'SD-T19-75-0-0.5'	TC	U	42.23472	-83.1505	2450000	310	6600
'SD-T19-75-0.5-1'	TC	D	42.22537	-83.1369	2040000	150	5900
'SD-T19-75-1-1.6'	TC	D	42.22537	-83.1369	1690000	88	9600
'SD-T19-100-0-0.5'	TC	U	42.22537	-83.1369	3770000	630	6900
'SD-T19-100-0.5-1'	TC	D	42.22107	-83.1362	4590000	460	7800
'SD-T19-100-1-3'	TC	D	42.22107	-83.1362	2700000	160	3000
'SD-T19-100-3-3.9'	TC	D	42.22107	-83.1362	1980000	140	3000
'SD-T20-25-0-0.5'	TC	U	42.21845	-83.1364	6300000	1900	10100
'SD-T20-25-0.5-1'	TC	D	42.21845	-83.1364	7210000	1700	10100
'SD-T20-25-1-3'	TC	D	42.21845	-83.1364	6210000	1300	9900
'SD-T20-25-3-5'	TC	D	42.2177	-83.1359	5730000	1200	9500
'SD-T20-25-5-7'	TC	D	42.2177	-83.1359	2340000	230	4500
'SD-T20-50-0-0.5'	TC	U	42.2177	-83.1359	2580000	710	5000
'SD-T20-50-0.5-1'	TC	D	42.2177	-83.1359	4440000	1300	8200
'SD-T20-50-1-3'	TC	D	42.2184	-83.1344	6770000	1200	10800
'SD-T20-50-3-5'	TC	D	42.2184	-83.1344	2210000	100	6200
'SD-T20-50-5-5.5'	TC	D	42.2184	-83.1344	2240000	74	7500
'SD-T20-100-0-0.5'	TC	U	42.21767	-83.1347	6550000	220	3400
'SD-T20-100-0.5-1'	TC	D	42.21767	-83.1347	5680000	280	4900
'SD-T20-100-1-3'	TC	D	42.21767	-83.1347	9030000	220	6800
'SD-T20-100-3-3.6'	TC	D	42.21767	-83.1347	6740000	230	8300
'SD-T21-25-0-0.5'	TC	U	42.22572	-83.146	850000	2400	11000
'SD-T21-25-0.5-1'	TC	D	42.22572	-83.146	8150000	2500	11300
'SD-T21-25-1-3'	TC	D	42.22572	-83.146	6210000	1200	9000
'SD-T21-25-3-5'	TC	D	42.21648	-83.1407	5350000	820	17500
'SD-T21-25-5-7'	TC	D	42.21648	-83.1407	7330000	1700	19700
'SD-T21-25-7-7.7'	TC	D	42.21648	-83.1407	6540000	1800	27300
'SD-T21-50-0-0.5'	TC	U	42.21648	-83.1407	4100000	1000	4600
'SD-T21-50-0.5-1'	TC	D	42.21653	-83.1409	7220000	1200	7000
'SD-T21-50-1-3'	TC	D	42.21653	-83.1409	7900000	900	9400
'SD-T21-50-3-5'	TC	D	42.21653	-83.1409	7560000	1500	15900
'SD-T21-50-5-7'	TC	D	42.2165	-83.1409	7080000	1000	8500
'SD-T21-50-7-7.7'	TC	D	42.2165	-83.1409	2700000	84	3300
'SD-T21-75-0-0.5'	TC	U	42.2165	-83.1409	4730000	540	7300
'SD-T21-75-0.5-1'	TC	D	42.2165	-83.1408	5180000	490	8100
'SD-T21-75-1-3'	TC	D	42.2165	-83.1408	6020000	280	7400
'SD-T21-75-3-5'	TC	D	42.2165	-83.1408	4110000	190	5800
'SD-T21-75-5-6'	TC	D	42.21594	-83.1408	2230000	80	6000
'SD-T21-100-0-0.5'	TC	U	42.21594	-83.1408	2850000	390	5100
'SD-T21-100-0.5-1'	TC	D	42.21594	-83.1408	2810000	290	4500
'SD-T21-100-1-3.1'	TC	D	42.21594	-83.1408	4600000	290	7000
'SD-T23-25-0-0.5'	TC	U	42.24218	-83.1404	6330000	600	9200
'SD-T23-25-0.5-1'	TC	D	42.24218	-83.1404	5310000	450	8100
'SD-T23-25-1-1.5'	TC	D	42.24218	-83.1404	5450000	360	17200
'SD-T23-50-0-0.5'	TC	U	42.24218	-83.1404	5280000	350	5200
'SD-T23-50-0.5-1'	TC	D	42.24218	-83.1404	5040000	330	5300
'SD-T23-50-1-3'	TC	D	42.21596	-83.141	3930000	250	3900
'SD-T23-50-3-4'	TC	D	42.21596	-83.141	6550000	310	4800
'SD-T23-75-0-0.5'	TC	U	42.21596	-83.141	5730000	240	3900
'SD-T23-75-0.5-1'	TC	D	42.21596	-83.141	4630000	300	3400
'SD-T23-75-1-2'	TC	D	42.21596	-83.141	5680000	410	3800

Table 5: A fraction of T_BASF_Adjusted data set

D Bibliography

References

[1] A.Dasgupta and A.E.Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441), March 1998. 22

[2] A.K.Jain. *Algorithms for Clustering Data*. Prentive-Hall Inc., 1 edition, 1988. 6, 9

[3] C.Fraley and A.E.Raftery. Model-based clustering, discriminant analysis, and density esimation. *Journal of the American Statistical Association*, 97(458):611, Jun 2002. 21, 22, 26

[4] Guojun Gan Chaoqun Ma and Jianhong Wu. *Data Clustering:Theory, Algorithms and Applications*. Sociity for Industrial and Applied Mathematics & American Statistical Association, 2007. 21

[5] R.G.Garrett C.Reimann, P.Filzmoser and R.Dutter. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. John Wiley & Sons, Ltd., 2008. 8

[6] D.L.Pham. Spatial models for fuzzy clustering. *Computer Vision and Image Understanding*, 84(285-297), 2001. 21

[7] D.R.Helsel. *Nondetects and Data Analysis: Statistics for censored environmental data*. John Wiley & Sons, Inc., 2005. 8

[8] H.C.Romesburg. *Cluster Analysis for researchers*. Robert E.Krieger Publishing Company, Malabar, Florida, 2 edition, 1990. 21, 22, 24

[9] Robert I.Kabacoff. *Quick-R: Cluster Analysis*. http://www.statmethods.net/advstats/cluster.html. 21

[10] J.C.Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), December 1971. 23

[11] J.M.Chambers and T.J.Hastie. *Statistical Models in S*. AT&T Bell Laboratories, 1992. 3

[12] J.T.Kent K.V.Mardia and J.M.Bibby. *Multivariate Analysis*. Academic Press, 1979. 21

[13] L.Kaufman and P.J.Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, Inc., 1990. 6, 21, 23

[14] Jari Oksanen. *Cluster Analysis: Tutorial with R.* http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf. 21

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0. 3

[16] R-project. *CRAN task view: Cluster Analysis & Finite Mixture Models.* http://cran.r-project.org/web/views/Cluster.html. 21

[17] J.M.Chambers R.A.Becker and A.R.Wilks. *The New S Language: A programming Environment for Data Analsysi and Graphics*. Wadsworth & Brooks/Cole Advanced Books & Software, 1988. 3

[18] R.O.Duda and P.E.Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 1 edition, 1973. 9

[19] P.E.Hart R.O.Duda and D.G.Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2 edition, 2001. 21, 27

[20] S.P.Millard. *EnvironmentalStats for S-Plus: user's manual for Windows and UNIX*. Springer, 1998. 7, 8, 14

[21] W.N.Venables and B.D.Ripley. *Mordern Statistics with S*. Springer, 4 edition, 2002. 18