Strumenti di clustering robusto a dati di commercio internazionale e applicazioni anti-frode

Domenico Perrotta and Francesca Torti (EC, JRC)

Marco Riani and Andrea Cerioli (University of Parma)

MATLAB EXPO 2016 Italia, 22 Giugno, Milano
Outline

• Application context

• Data and statistical problems

• Clustering methods in FSDA

• Assessing clustering methods in FSDA for MATLAB

• Application to international trade datasets

• The FSDA toolbox for MATLAB
Application context

Analysis of trade data in order to detect:

- customs frauds (e.g. under-valuation of import duties);
- trade related infringements (e.g. money laundering);
- cases of circumvention of EU trade regulations (e.g. anti-dumping and countervailing measures);
- situations which may lead to distorted competition or potentially unfair trading condition involving Member States (state monopolies, dominant positions, deflection of trade).
Legal basis


Data source 1: COMEXT

• A Eurostat database on official EU trade statistics

• Data are monthly aggregates of trade quantities and values for each Product, Origin and Destination

• 1-year COMEXT data contain about 6.000.000 records
Data source 2: customs declarations

1-year Italian custom data contains about 6.500.000 declarations

<table>
<thead>
<tr>
<th>P</th>
<th>O</th>
<th>D</th>
<th>DISPATCH</th>
<th>COMPANY_ID</th>
<th>DECLARATION</th>
<th>PREFERENCE</th>
<th>PROCED</th>
<th>DATE</th>
<th>VALUE</th>
<th>WEIGHT</th>
<th>SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>712903000</td>
<td>TR</td>
<td>IT</td>
<td>TR</td>
<td>1704</td>
<td>1</td>
<td>300</td>
<td>40</td>
<td>4/1/2011</td>
<td>36300</td>
<td>9800</td>
<td>0</td>
</tr>
<tr>
<td>8803300090</td>
<td>US</td>
<td>IT</td>
<td>US</td>
<td>100049</td>
<td>2</td>
<td>100</td>
<td>40</td>
<td>8/1/2011</td>
<td>12607.17</td>
<td>34.16</td>
<td>0</td>
</tr>
<tr>
<td>3208909190</td>
<td>US</td>
<td>IT</td>
<td>IT</td>
<td>100049</td>
<td>3</td>
<td>100</td>
<td>51</td>
<td>11/1/2011</td>
<td>5074.39</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>3102101000</td>
<td>RU</td>
<td>IT</td>
<td>IT</td>
<td>16305</td>
<td>4</td>
<td>100</td>
<td>40</td>
<td>8/1/2011</td>
<td>1113913</td>
<td>3600994</td>
<td>1656457</td>
</tr>
<tr>
<td>3923900000</td>
<td>IT</td>
<td>IT</td>
<td>AL</td>
<td>2068</td>
<td>5</td>
<td>100</td>
<td>40</td>
<td>3/1/2011</td>
<td>40</td>
<td>240</td>
<td>0</td>
</tr>
<tr>
<td>2710111100</td>
<td>LY</td>
<td>IT</td>
<td>IT</td>
<td>46840</td>
<td>6</td>
<td>140</td>
<td>40</td>
<td>4/1/2011</td>
<td>15027037</td>
<td>23175981</td>
<td>0</td>
</tr>
<tr>
<td>8411910090</td>
<td>US</td>
<td>IT</td>
<td>US</td>
<td>113981</td>
<td>7</td>
<td>100</td>
<td>51</td>
<td>11/1/2011</td>
<td>3160.02</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>8411990090</td>
<td>US</td>
<td>IT</td>
<td>US</td>
<td>113981</td>
<td>8</td>
<td>100</td>
<td>51</td>
<td>10/1/2011</td>
<td>62251.66</td>
<td>473</td>
<td>0</td>
</tr>
<tr>
<td>6204623190</td>
<td>AL</td>
<td>IT</td>
<td>AL</td>
<td>60426</td>
<td>9</td>
<td>300</td>
<td>61</td>
<td>3/1/2011</td>
<td>9060.2</td>
<td>618.84</td>
<td>831</td>
</tr>
<tr>
<td>6204623190</td>
<td>AL</td>
<td>IT</td>
<td>AL</td>
<td>60426</td>
<td>10</td>
<td>100</td>
<td>61</td>
<td>3/1/2011</td>
<td>36739.55</td>
<td>1988.54</td>
<td>3470</td>
</tr>
</tbody>
</table>
• **Value** and **Quantity** of a given product imported into a EU Member State from non-EU countries/traders in a fixed time period

• **Analysis to be repeated** for each EU Member State and many types of goods (more than 10,000 in the Combined Nomenclature)

• **Systematic** behavior in trade may originate from:
  - different traders and/or countries
  - different product quality
  
  ...  

- but also fraudulent transactions:  
  
  systematic under/over pricing, money laundering, VAT evasion ...

• **Occasional** frauds or infringements may also occur
Often a regression structure is appropriate for trade data.

This is POD44072925CMNL (final $R^2=0.99297$, initial $R^2=0.93055$)
Heteroscedasticity

Perfect fit
Sometimes elliptical contoured structures become more suitable, after transforming the data.
Purposes of statistical analysis

- **Classification**: assign each transaction to one of G groups
  - Exploratory point of view: no a priori information about possible frauds
  - Concomitant information (e.g. country; company; etc.) can be introduced in subsequent confirmative stages
**Purposes of statistical analysis**

- **Classification**: assign each transaction to one of $G$ groups
  - Exploratory point of view: no a priori information about possible frauds
  - Concomitant information (e.g. country; company; etc.) can be introduced in subsequent confirmative stages

- **Estimation** of the “fair price” for each good
Purposes of statistical analysis

- **Classification**: assign each transaction to one of G groups
  - Exploratory point of view: no a priori information about possible frauds
  - Concomitant information (e.g. country; company; etc.) can be introduced in subsequent confirmative stages

- **Estimation of the “fair price”** for each good

- **Identification of major departures from the fair price**: systematic fraudulent behavior
Purposes of statistical analysis

• **Classification**: assign each transaction to one of G groups
  — Exploratory point of view: no a priori information about possible frauds
  — Concomitant information (e.g. country; company; etc.) can be introduced in subsequent confirmative stages

• **Estimation of the “fair price”** for each good

• **Identification of major departures from the fair price**: systematic fraudulent behavior

• **Identification of major outliers**: occasional fraudulent behavior
Need to treat also large datasets with hundreds of thousands units.
We need to apply robust methods in a variety of contexts:

- regression
- clustering
- classification
- data transformation
The Forward Search: structure

The FS orders the data by closeness to the assumed model. 
Ingredients:

1. **Start** with a small (outlier-free) subset of $m_0$ observations 
2. **Move Forward** by increasing the number of observations $m \geq m_0$ used for fitting the model.
3. **Continue** until $m = n$
4. Outliers and other observations not following the general structure enter at the end: they can be clearly identified by **diagnostic monitoring**

The structure is very general and can be tailored to many problems through appropriate choices at steps 1–4.
Multivariate outlier detection: monitoring the minimum Mahalanobis distance among observations outside the subset.

load sixty_eighty;
Y = sixty_eighty.data;
FSM(Y);
Outlier detection in regression: monitoring the minimum deletion residual among observations outside the subset

```matlab
load fishery;
X = fishery.data;
x = X(:,1);
y = X(:,2);
FSR(y(ii),x(ii),'bonflev',0.9999);
```

% ii = indices of a subset of the fishery dataset.
Outlier detection in **heteroskedastic linear regression**

**Example of Heteroskedastic trade dataset:**

*fish fillets and other fish meat from MV to IT*

```matlab
XX = load('tradeH.txt');
y = XX(:,2);
X = XX(:,1);
X = X ./ max(X);
plot(X, y, 'o')
```
Outlier detection in [heteroskedastic linear regression](#).

Outliers detected with a robust (FS) but homescedastic model.

```matlab
FSR(y, X, 'init', round(length(y)/2), ...
    'plots', 1, 'ylim', [1.6 3]);
```
Outlier detection in **heteroskedastic linear regression**


```matlab
FSRH(y,X,'init',round(length(y)/2),... 'plots',1,'ylim',[1.6 3]);
```
Finding groups in regression data

load fishery;
X=fishery.data;
tclustreg(X,3,5,0.01,0.01,'intercept',0);
Finding groups in multivariate data
Compare different algorithms:

\[ Y = \text{load('mixture100.txt')}; \]
\[ \text{tclust}(Y(:,1:2),3,0.05,1000,...} \]
\[ \text{'refsteps',20,'plots',1}); \]
Finding groups in multivariate data
Compare different algorithms:

TCLUST with mixture models
(selection of untrimmed units according to densities weighted by estimates of the probability of the components)

```
tclust(Y(:,1:2),3,0.05,1000,...
    'refsteps',20,'plots',1,'mixt',2)
```
Finding groups in trade data
Italian imports – Example 1

CN-94052099: **Electric table, desk, bedside or floor-standing lamps**
Origin: **China** – Destination: **Italy**; \( n = 1271 \)

**Bivariate observations** from several populations, with outliers.

**Outliers** may occur in \( y \) and/or in \( x \) (bad leverage points), or may be observations that do not belong to any group (classification outliers).

**Small trade area**, close to the origin, of no operational interest.

**Purposes for anti-fraud**: Identify the groups (regression lines), especially the most extreme ones. But also identify the outliers and the groups with less points, that in trade data carry relevant information.
Finding groups in trade data
Italian imports – Example 2

CN-62043390: Jackets and blazers of synthetic fibres;
Origin: China – Destination: Italy; \( n = 1248 \)

Similar features as in Example 1 (and many other customs data sets).

Again: outliers in both \( y \) and \( x \), and classification outliers.

The dense small trade area close to the origin is still present: not important for anti-fraud analysis.
Finding groups in trade data
Italian imports – Example 3

CN-6212122000: **Brassières, girdles, corsets, braces, etc.**;
Origin: **Israel** – Destination: **Austria**; \( n = 186 \)

*Apparantly simpler data set:* small sample and clear regression structures without strong contamination.

Some concentration of trade close to the origin: **negligible effect?**
How we propose to treat these complex data

- Robust clustering methods to find groups;

- A regression structure for each group;

- A denoising procedure to mitigate the effect of the dense area: Thinning.
Customs Example 1 (Electric lamps) revisited

Robust clustering after thinning: realization with $n^* = 242$ and $G = 4$

On the thinned data set the robust fit is good:

- The clusters identified are reasonable
- The uninteresting signals are eliminated
- The choice of the number of clusters is not crucial

Basis for identifying anomalous behaviour of traders
Customs Ex. 2 (Jackets and blazers) revisited

Robust clustering after thinning: realization with $n^* = 222$ and $G = 4$

Again, the robust fit is good on the thinned data set
Customs Ex. 3 (Brassières, etc.) revisited

Robust clustering after thinning: realization with $n^* = 65$ and $G = 3$

Again, the robust fit is good on the thinned data set

But, what is the effect of thinning?
Customs Ex. 3 (Brassières, etc.) revisited

Robust clustering **without** thinning. Now, original data with $n^* = 186$ and $G = 3$

Even if the structure is apparently simple, the presence of the dense area distorts the robust groups
Simulating mixtures of regression data with pre-specified (maximum/average) level of overlapping

\[ Q = \text{MixSimReg}(k,p, 'BarOmega', \text{BarOmega}); \]
\[ [y,X,id] = \text{simdatasetReg}(n, Q.Pi, Q.Beta, Q.S, Q.Xdistrib); \]
\[ \text{spmplot}([y X],id); \]

**Small overlap** BarOmega=0.01

**Strong overlap** BarOmega=0.2
Simulating mixtures of multivariate data with pre-specified overlap and restrictions

Constraints on eigenvalues of covariance matrices: allows to simulate data as in the TCLUST model

- \( \text{restrfactor: } \in [1, \infty] \), specifies the maximum ratio to allow between the largest eigenvalue and the smallest eigenvalue of the \( k \) covariance matrices generated:

\[
\frac{\max_{i=1, \ldots, v} \max_{j=1, \ldots, k} \lambda_i(\hat{\Sigma}_j)}{\min_{i=1, \ldots, v} \min_{j=1, \ldots, k} \lambda_i(\hat{\Sigma}_j)} \leq \text{restrfactor}
\]

- Controls departures from sphericity and relative cluster sizes (ratio of the length of ellipsoids axes).
Simulating mixtures of multivariate data with pre-specified overlap and restrictions

Constraints on eigenvalues of covariance matrices

out = MixSim(3,5,'BarOmega',0.1, 'MaxOmega',0.2, 'restrfactor',1.1);
out1 = MixSim(3,5,'BarOmega',0.1, 'MaxOmega',0.2);

[X ,id] = simdataset(n, out.Pi , out.Mu , out.S);
[X1,id1]= simdataset(n, out1.Pi, out1.Mu, out1.S);

Without constraints on eigenvalues of covariance matrices
Some References

First Journal article on FSDA

Includes other computational improvements (efficient subsampling)

Includes important computational improvements for the FS estimator

Journal of Statistical Software

The Forward Search for Very Large Datasets

Marco Riani
Università di Parma

Domenico Perrotta
European Commission
Joint Research Centre

Andrea Cerioli
Università di Parma


The main page of FSDA

http://www.rian.i.it/MATLAB  http://fsda.jrc.ec.europa.eu

Documentation

Flexible Statistics and Data Analysis Toolbox

Analyze complex data using robust statistics estimators

Flexible Statistics and Data Analysis Toolbox ™ extends MATLAB® and statistics toolbox® to support a robust and efficient analysis of complex data sets affected by different sources of heterogeneity. The toolbox contains three categories of tools:

- Robust Regression Analysis routines (including transformations)
- Robust Multivariate Analysis routines (including transformations).
- Robust Cluster Analysis routines (regression and multivariate)

Code for any function inside the toolbox is open and extensible. Use the MATLAB Editor to review, copy, and edit M-file code for any function. Extend the toolbox by copying code to new M-files or by writing M-files that call toolbox functions.

Getting Started
Learn the basics of Flexible Statistics and Data Analysis Toolbox
More than 20 tutorials

Tutorials

Introduction to robust statistics

Introduction
Technical introduction to robust statistics in regression
Technical introduction to robust statistics in multivariate analysis
Introduction to the forward search philosophy of data analysis

Dynamic Statistical Visualization

Introduction to dynamic visualization
Dynamic visualization in the context of the forward search
Dynamic visualization in the index plot of residuals
Dynamic visualization in the monitoring residuals plot
Dynamic visualization in the minimum deletion residuals plot
Dynamic visualization in the fan plot
Dynamic visualization in the yXplot
Dynamic visualization in the candlestickplot

Robust regression analysis

Introduction to robust estimators in linear regression
Robust linear regression using LMS and LTS estimators
Robust linear regression using S and MM estimators
Robust forward linear regression with exploratory purposes
Robust forward linear regression with automatic outlier detection procedure
Tutorial example

Introduction to Robust Statistics

Awareness of outliers in some form or another has existed for at least 2000 years. Thucydides, in his third book about the Peloponnesian War (III 20, 3-4), describes how the Plataeans used in 428 BC concepts of robust statistics in order to estimate the height of the ladder which was needed to overcome the fortifications built by the Peloponnesians and the Boeotians who were besieging their city.

"The same winter the Plataeans, who were still being besieged by the Peloponnesians and Boeotians, distressed by the failure of their provisions, and seeing no hope of relief from Athens, nor any other means of safety, formed a scheme with the Athenians besieged with them for escaping, if possible, by forcing their way over the enemy’s walls; the attempt having been suggested by Theaenetus, son of Tolmides, a soothsayer, and Eupompides, son of Daimachus, one of their generals. At first all were to join: afterwards, half hung back, thinking the risk great; about two hundred and twenty, however, voluntarily persevered in the attempt, which was carried out in the following way. Ladders were made to match the height of the enemy’s wall, which they measured by the layers of bricks, the side turned towards them not being thoroughly whitewashed. These were counted by many persons at once; and though some might miss the right calculation, most would hit upon it, particularly as they counted over and over again, and were no great way from the wall, but could see it easily enough for their purpose. The length required for the ladders was thus obtained, being calculated from the breadth of the brick".
**Tutorials**

**Introduction to robust statistics**
- Introduction
- Technical introduction to robust statistics in regression
- Technical introduction to robust statistics in multivariate analysis
- Introduction to the forward search philosophy of data analysis

**Dynamic Statistical Visualization**
- Introduction to dynamic visualization
- Dynamic visualization in the context of the forward search
- Dynamic visualization in the index plot of residuals
- Dynamic visualization in the monitoring residuals plot
- Dynamic visualization in the minimum deletion residuals plot
- Dynamic visualization in the fan plot
- Dynamic visualization in the yXplot
- Dynamic visualization in the candlestickplot

**Robust regression analysis**
- Introduction to robust estimators in linear regression
- Robust linear regression using LMS and LTS estimators
- Robust linear regression using S and MM estimators
- Robust forward linear regression with exploratory purposes
- Robust forward linear regression with automatic outlier detection procedure

**Transformations**
- Introduction to robust transformations in linear regression
- Score test for transformation
- Forward score test

**Model selection**
- Introduction to variable selection
- Variable selection using forward added-t-test
- Robust model selection using Cp
### Flexible Statistics and Data Analysis Toolbox Functions

#### Robust regression analysis and transformations

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>addt</td>
<td>Produces the t test for an additional expl. variable</td>
</tr>
<tr>
<td>FSR</td>
<td>Gives an automatic outlier detection procedure in linear regression</td>
</tr>
<tr>
<td>FSRbbab</td>
<td>Returns the units belonging to the subset in each step of the Bayesian forward search</td>
</tr>
<tr>
<td>FSRbab</td>
<td>Returns the units belonging to the subset in each step of the forward search</td>
</tr>
<tr>
<td>FSRfda</td>
<td>Enables to monitor several quantities in each step of the forward search</td>
</tr>
<tr>
<td>FSRinvmdr</td>
<td>Computes the theoretical envelopes of Minimum Deletion Residual outside subset during the search</td>
</tr>
<tr>
<td>FSRinvmdrs</td>
<td>Computes values of minimum deletion residual into confidence levels</td>
</tr>
<tr>
<td>FSRminid</td>
<td>Computes minimum deletion residual and other basic linear regression quantities in each step</td>
</tr>
<tr>
<td>FSRr</td>
<td>Forward search in linear regression reweighted</td>
</tr>
<tr>
<td>LXS</td>
<td>Computes the Least Median of Squares (LMS) or Least Trimmed Squares (LTS) estimators</td>
</tr>
<tr>
<td>MMreg</td>
<td>Computes MM estimator of regression coefficients</td>
</tr>
<tr>
<td>MMregcore</td>
<td>Computes MM regression estimators for a selected fixed scale</td>
</tr>
<tr>
<td>RobCov</td>
<td>Computes covariance matrix of robust regression coefficients</td>
</tr>
</tbody>
</table>
## Alphabetical list of functions (Produced automatically)

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>tabulateFs</code></td>
<td>Create frequency table of unique values of x, excluding possible 0 counts</td>
</tr>
<tr>
<td><code>TauReg</code></td>
<td>Computes Tau estimators in linear regression</td>
</tr>
<tr>
<td><code>Tbbdp</code></td>
<td>Finds the constant c associated to the supplied breakdown point for Tukey's biweight</td>
</tr>
<tr>
<td><code>Tbc</code></td>
<td>Computes breakdown point and efficiency associated with constant c for Tukey's biweight</td>
</tr>
<tr>
<td><code>TbEff</code></td>
<td>Finds the constant c which is associated to the requested efficiency for Tukey biweight estimator</td>
</tr>
<tr>
<td><code>Tbsol</code></td>
<td>Computes psi function (derivative of rho function) for Tukey's biweight</td>
</tr>
<tr>
<td><code>Tbsolder</code></td>
<td>Computes derivative of psi function (second derivative of rho function) for Tukey's biweight</td>
</tr>
<tr>
<td><code>Tbsolx</code></td>
<td>Computes psi function (derivative of rho function) times x for Tukey's biweight</td>
</tr>
<tr>
<td><code>Tbsolho</code></td>
<td>Computes (rho) function for Tukey biweight</td>
</tr>
<tr>
<td><code>Tbsolh1</code></td>
<td>Computes weight function psi(u)/u for Tukey biweight</td>
</tr>
<tr>
<td><code>tclustIC</code></td>
<td>Computes tclust for different number of groups k and restriction factors c</td>
</tr>
<tr>
<td><code>tclustICplot</code></td>
<td>Plots information criterion as a function of c and k</td>
</tr>
<tr>
<td><code>tclustICsel</code></td>
<td>Extracts a set of best relevant solutions</td>
</tr>
<tr>
<td><code>tclustreg</code></td>
<td>Performs robust linear grouping analysis</td>
</tr>
<tr>
<td><code>tkmeans</code></td>
<td>Computes trimmed k-means</td>
</tr>
<tr>
<td><code>triu2vec</code></td>
<td>Extracts in a vector the linear indexes or the elements on and above the k-th diagonal of a square</td>
</tr>
<tr>
<td><code>unibiv</code></td>
<td>Has the purpose of detecting univariate and bivariate outliers</td>
</tr>
<tr>
<td><code>upperfracpos</code></td>
<td>Positions two figures on the upper part of the screen</td>
</tr>
<tr>
<td><code>winsor</code></td>
<td>Returns a winsorized copy of input</td>
</tr>
</tbody>
</table>
Examples, GUIs and didactic movies

**Flexible Statistics and Data Analysis Toolbox Examples**

### Robust Regression

<table>
<thead>
<tr>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Display GUI" /></td>
<td>Displays a GUI where it is possible to brush steps from the monitoring residuals plot</td>
</tr>
<tr>
<td><img src="image2" alt="Display GUI" /></td>
<td>Displays a GUI where it is possible to brush units from the index plot of residuals</td>
</tr>
<tr>
<td><img src="image3" alt="Examples" /></td>
<td>Examples of Robust Regression Using Robust Estimators</td>
</tr>
</tbody>
</table>

### Robust Multivariate Analysis

<table>
<thead>
<tr>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="Display GUI" /></td>
<td>Displays a GUI where it is possible to brush steps from the monitoring distances plot</td>
</tr>
</tbody>
</table>
Example of GUI on robust regression

To explore how dynamic brushing works, select one of the radio buttons on the left.

The monitoring of the residuals plot together with the minimum deletion residual plot will appear automatically.

The points associated with the trajectories you select in the plot of scaled residuals will also be automatically highlighted in the plots of minimum deletion residual and in the scatter plot matrix.

To monitor the residuals and the minimum deletion residual with your own data, type:

[out]=LXS(y,X); to find a robust initial subset
[out]=FSReda(y,X,out.bs); To run the forward search
mdrplot(out); To plot the minimum deletion residual
resfwdplot(out); To plot the scaled squared residuals

See the help files of the above functions for more information

eexample_regression.m
Rich repertoire of Datasets

Data Sets

The Flexible Statistics Data Analysis Toolbox™ software includes a number of sample data sets provided in tab separated data file (.txt file). The simplest way to load a data set into the MATLAB workspace is to type:

```matlab
load filename.txt
```

where `filename` is one of the files listed in the table. This will load the original txt file into a standard data matrix. Note that this straightforward loading option will not work for datasets mixing numerical and categorical variables, like `fishery.txt`.

The datasets have been also organized in a structure complementing the data matrix with the variable and observation names and data specific notes. Once the structure is loaded in the workspace by typing

```matlab
load filename.mat
```

the four structure fields can be retrieved with:

- `data = filename.data;`
- `rownames = filename.rownames;`
Datasets fully described

Regression Data Sets

<table>
<thead>
<tr>
<th>File</th>
<th>Description of Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>algae</td>
<td>The algae dataset (from Hettich and Bay, 1999) contains 90 measurements at a river in some place in Europe. There are 11 predictors. The first three are categorical: the season of the year, river size (small, medium and large) and fluid velocity (low, medium and high). The other eight are the concentrations of several chemical substances. The response is the logarithm of the abundance of a certain class of algae. The normal Q-Q plots of the residuals corresponding to the LS estimate gives the impression of short-tailed residuals, while the residuals from a robust fit indicate the existence of at least two outliers, i.e. observations 36 and 77.</td>
</tr>
<tr>
<td>aircraft</td>
<td>The aircraft dataset (Rousseeuw and Leroy, 1987, page 154, table 22) contains 23 single-engine aircraft built over the years 1947-1979, from Office of Naval Research. The dependent variable is cost in units of $100,000 (last column) and the explanatory variables are aspect ratio, lift-to-drag ratio, weight of plane (in pounds) and maximal thrust. Based on MCD without correction factor, 4 observayions, included observation 15, are detected as outliers. Based on the corrected MCD, observation 15 is no longer detected as outlier.</td>
</tr>
<tr>
<td>credit_card</td>
<td>Credit card data, introduced by Riani (2011 forthcoming), are formed by 1,000</td>
</tr>
</tbody>
</table>
Example of Documentation page (automatically produced from publishFS.m file)

Documentation

Contents

Flexible Statistics and Data Analysis (FSDA)

`tclust` computes trimmed clustering with restrictions on the eigenvalues

Syntax

```matlab
out = tclust(Y, k, alpha, restrfactor)

out = tclust(Y, k, alpha, restrfactor, Name, Value)

[out, varargout] = tclust(__)
```

Description

tclust partitions the points in the n-by-v data matrix Y into k clusters. This partition minimizes the trimmed sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of Y correspond to points, columns correspond to variables. tclust returns inside structure out an n-by-1 vector idx containing the cluster indices of each point. By default, tclust uses (squared) possibly constrained, Mahalanobis distances.

```matlab
out = tclust(Y, k, alpha, restrfactor) tclust of geyser data using k=3, alpha=0.1 and restrfactor=1000.

out = tclust(Y, k, alpha, restrfactor, Name, Value) tclust of geosser with classification plot.
```
Characteristics of .m files

- The documentation inside the .m file can be easily accessed from the command prompt typing help and the name of the function.

```matlab
>> help MixSim
MixSim generates k clusters in v dimensions with given overlap

Required input arguments:

- `k`: number of groups (components). Scalar. Desired number of groups.
  Data Types - int16|int32|int64|single|double
- `v`: number of dimensions (variables). Scalar. Desired number of variables.
  Data Types - int16|int32|int64|single|double

Optional input arguments:

- `BarOmega`: Requested average overlap. Scalar. Value of desired average overlap. The default value is ''
  Example - 'BarOmega',0.05
  Data Types - double
- `MaxOmega`: Requested maximum overlap. Scalar. Value of desired maximum
```
From .m to .html file (one to one correspondence)
Options arguments fully documented

Name-Value Pair Arguments
Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside single quotes (' '). You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example: 'nsamp',1000 , 'refsteps',10 , 'reftol',1e-05 , 'equalweights',true , 'mixt',1 , 'plots',1 , 'msg',1 , 'nocheck',10 , 'starttv1',1 , 'restr','deter' , 'Ysave',1

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsamp</td>
<td>number of subsamples to extract</td>
<td>scalar</td>
</tr>
<tr>
<td>refsteps</td>
<td>Number of refining iterations</td>
<td>scalar</td>
</tr>
<tr>
<td>reftol</td>
<td>tolerance for the refining steps</td>
<td>scalar</td>
</tr>
<tr>
<td>equalweights</td>
<td>cluster weights in the concentration and assignment steps</td>
<td>logical</td>
</tr>
</tbody>
</table>
Every input/argument is extensively described

Output:

out: structure which contains the following fields

out.muopt = k-by-v matrix containing cluster centroid locations. Robust estimate of final centroids of the groups.

out.sigmaopt = v-by-v-by-k array containing estimated constrained covariance for the k groups.

out.idx = n-by-1 vector containing assignment of each unit to each of the k groups. Cluster names are integer numbers from 1 to k. 0 indicates trimmed observations.

out.siz = matrix of size k-by-3
1st col = sequence from 0 to k
2nd col = number of observations in each cluster
Every input/output which is a structure is automatically converted in a html table.
.m file contains a series of examples readily executables inside %{  %}

```matlab
% tclust of geyser with classification plot.
Y=load('geyser2.txt');
out=tclust(Y,3,0.1,10000,'plots',1);
%

% tclust of geyser with varargout.
Y=load('geyser2.txt');
nsamp=20;
[out,MatrixContSubsets]=tclust(Y,3,0.1,10000,'nsamp',nsamp);
% MatrixContSubsets is a matrix containing in the rows the indexes of
% the nsamp subsets which have been extracted
```
Description

tclust partitions the points in the n-by-v data matrix Y into k clusters. This partition minimizes the trimmed sum of squares. The columns correspond to variables. tclust returns inside structure out an n-by-1 vector idx containing the cluster of each observation.

\[ \text{out} = \text{tclust}(Y, k, \alpha, \text{restrfactor}) \]

This function is useful for data with outliers.

\[ \text{out} = \text{tclust}(Y, k, \alpha, \text{restrfactor}, \text{Name}, \text{Value}) \]

tclust of geyser data using k=3, alpha=0.1 and restrfactor=1

\[ [\text{out}, \text{varargout}] = \text{tclust}(...) \]

tclust of geyser with varargout.

Examples

- tclust of geyser data using k=3, alpha=0.1 and restrfactor=10000.
- tclust of geyser with classification plot.
- tclust of geyser with varargout.

Related Examples

- tclust of geyser data (output comparison).
- tclust applied to the M5data.
- tclust in presence of structured noise.
- tclust applied to mixture100 data.
- tclust applied to mixture100 data, comparison of different options.
Every HTML documentation contains a series of **Examples** and Related Examples

- The icon 📚 at the beginning of the line, indicates that the associated example has been executed and its output has been captured inside the HTML file.

## Related Examples

- 📚 tclust of geyser data (output comparison).

- tclust applied to the M5data.
Output after clicking on

- tclust of geyser data (output comparison).

We compare the output using three different values of restriction factor.

```matlab
close all
Y=load('geyser2.txt');
restrfactor=10000;
% nsamp = number of subsamples which will be extracted
nsamp=500;
out=tclust(Y,3,0.1,restrfactor,'nsamp',nsamp,'plots',1);
title(['Restriction factor = ' num2str(restrfactor)])
restrfactor=10;
out=tclust(Y,3,0.1,restrfactor,'nsamp',nsamp,'refsteps',10,'plots',1);
title(['Restriction factor = ' num2str(restrfactor)])
% trimmed k-means solution restrfactor=1
restrfactor=1;
out=tclust(Y,3,0.1,restrfactor,'nsamp',nsamp,'refsteps',10,'plots',1);
title(['Restriction factor = ' num2str(restrfactor) '. Trimmed k-means solution'])
cascade
```

Total estimated time to complete tclust:  4.16 seconds
Code output
Sometimes inside the .m file (especially in the section “More About:”) we provide statistical background and we add a number of formulae in latex language (see screenshot below)

More About:

\textbf{MixSim}(k,v)$ generates k groups in v dimensions. It is possible to control the average and maximum or standard deviation of overlapping.

Given two generic clusters $i$ and $j$ with $i \neq j = 1, ..., k$, indexed by $\phi(x; \mu_i, \Sigma_i)$ and $\phi(x; \mu_j, \Sigma_j)$ with probabilities of occurrence $\pi_i$ and $\pi_j$, the misclassification probability with respect to cluster $i$ (that is conditionally on $x$ belonging to cluster $i$), which is called $w_{j|i}$ is defined as $Pr[ \pi_i \phi(x; \mu_i, \Sigma_i) < \pi_j \phi(x; \mu_j, \Sigma_j)]$. The matrix containing the misclassification probabilities $w_{j|i}$ is called OmegaMap

The probability of overlapping between groups $i$ and $j$ is given by:

\[
\begin{align*}
    w_{j|i} + w_{i|j} & \quad \text{quad} \quad i,j=1,2, \ldots, k
\end{align*}
\]

The diagonal elements of OmegaMap are equal to 1.

The average overlap (which in the code is called below BarOmega) is
More About section of each HTML file (Latex formulae automatically processed)

More About

Additional Details

MixSim(k,v) generates k groups in v dimensions. It is possible to control the average and maximum or standard deviation of overlapping.

Given two generic clusters i and j with $i \neq j = 1, \ldots, k$, indexed by $\phi(x; \mu_i, \Sigma_i)$ and $\phi(x; \mu_j, \Sigma_j)$ with probabilities of occurrence $\pi_i$ cluster i (that is conditionally on x belonging to cluster i, which is called $w_{ji|i}$) is defined as $Pr[\pi_i \phi(x; \mu_i, \Sigma_i) < \pi_j \phi(x; \mu_j, \Sigma_j)]$.

The matrix containing the misclassification probabilities $w_{ji|i}$ is called OmegaMap The probability of overlapping between groups i and j is

$$w_{ji|i} + w_{ij|i} \quad i, j = 1, 2, \ldots, k$$

The diagonal elements of OmegaMap are equal to 1.

The average overlap (which in the code is called below BarOmega) is defined as the sum of the off diagonal elements of OmegaMap (each) The maximum overlap (which in the code is called MaxOmega) is defined as $\max(w_{ji|i} + w_{ij|i}, i \neq j = 1, 2, \ldots, k)$.
FSDA html documentation files and MATLAB search engine

- Matlab2015a-2016b: full integration
Example: Output of the search string «concentration step»
FSDA html documentation files and MATLAB search engine

- Matlab2012b/2014b: it is necessary to use the old MATLAB search engine inside supplemental software
FSDA html documentation files and MATLAB search engine

• Matlab <=2012a: there was no distinction between MATLAB toolboxes and third parties toolboxes (as concerns the documentation)
Rich repertoire of robust methods and graphical tools for exploratory data analysis and inference

- Brushing, linking data tooltip etc.
- Automatic data transformations
- Robust regression analysis and transformations
- Robust multivariate analysis and transformations
- Robust clustering
- Dynamic visualization
- Guis
- Utilities

Robust clustering methods (TKMEANS, TCLUST, RLGA etc.)
Unique features of FSDA
Robust cluster analysis routines
Automatic choice of the number of groups
Enhanced scatter plot matrix
Clickable multilegend
Personalized datatooltip

Y(:,2) value equal to: 0.7914
Y(:,1) value equal to: 2.35
Unit: row 70
Unit entered in step m=74
Brushing and...
Brushing and logical linking
Brushing the index plot of residuals

Select a region to brush in the index plot of residuals

Brushed units, yvalue and x values

<table>
<thead>
<tr>
<th>yvalue</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.0000</td>
<td>-1.4192</td>
<td>-0.6787</td>
<td>-1.9986</td>
<td>-4.9259</td>
</tr>
<tr>
<td>31.0000</td>
<td>-3.9751</td>
<td>-1.8665</td>
<td>-1.7153</td>
<td>-6.5110</td>
</tr>
<tr>
<td>38.0000</td>
<td>-5.7308</td>
<td>-1.7083</td>
<td>-2.0538</td>
<td>-6.9914</td>
</tr>
<tr>
<td>47.0000</td>
<td>-4.9137</td>
<td>-1.9520</td>
<td>-1.5694</td>
<td>-7.0712</td>
</tr>
</tbody>
</table>

Highlight the index plot of residuals then: click on it to continue brushing or press a key
Brushing the BIC criteria to find optimal number of groups
Robust bivariate contours
Tools for data transformation with new interactive plots (e.g. the fan plot)
Loyalty cards: yX plot

- FREQUENCY
- AGECUST
- FAMILIYMEM

SALES
Loyalty cards: before and after transformation

- **Number of visits**
- **Age**
- **Number of persons in the family**

- **Sales**
  - Normal units
  - Outliers

- **Number of visits**
- **Age**
- **Number of persons in the family**
Effect of the outliers on the transformation

Dynamic link from the fan plot to the yXplot
Candle stick plots for robust model selection
Example files with complete analysis of many reference datasets

- examples_regression.m
- examples_multivariate.m
- examples_MixSim.m

- Also in new Matlab .mlx format
And much more

- More than 150 statistical functions
- Parallel processing for intensive tools
- Routines for calling R within MATLAB
- ...

- Please contact us at: FSDA@unipr.it
Special thanks to Mathworks: Francesca Perino, Giovanna Galliano and Dave Bergstein for the support

Authors and Developers

Authors who initiated the FSDA project and work on the software are:

Marco Riani (software development and design) and Andrea Cerioli (software design)
Section of Statistics
Department of Economics
University of Parma
mariani@unipr.it, andrea.cerioli@unipr.it

Domenico Perrotta (software development and design) and Francesca Torti (software development and design)
Global security and crisis management Unit
Institute for the Protection and Security of the Citizen
Joint Research Centre
European Commission
domenico.perrotta@ec.europa.eu, francesca.torti@jrc.ec.europa.eu

Current active developers also include:

Aldo Corbellini (software development and design, from 2013)
Section of Statistics
Department of Economics
University of Parma
aldo.corbellini@unipr.it

Software packaging and testing

Patrizia Calcaterra (from 2011)
Serco Belgium S.A.
seconded at EC Joint Research Centre
patrizia.calcaterra@ext.jrc.ec.europa.eu

Emmanuele Sordini (from 2014)
Institute for the Protection and Security of the Citizen
Joint Research Centre
European Commission
emmanuele.sordini@jrc.ec.europa.eu

Daniele Palermo (from 2014)
freelance consultant
seconded at EC Joint Research Centre
daniele.palermo@ext.jrc.ec.europa.eu