Università degli Studi di Firenze

Dipartimento di Statistica "Giuseppe Parenti"

UniCredit Banca

ROBCLA 2006

ROBust CLAssification and Discrimination with High Dimensional Data

A Data Mining-Based Data Auditing tool for Deviation Detection

Alfredo Roccato, Direzione Marketing, Unicredit Banca

Commercial organizations are realizing that it is necessary to **dedicate resources to the Data Auditing** (or Data Profiling) activities to look over their data, stored in the Data Warehouse, for Data Quality purposes **to avoid the risk of making wrong decisions**. In this poster we describe a **SAS macro** named *OUTLIERS* which uses different **robust estimators to detect atypical observations**. The macro uses SAS/IML, a matrix language, because a lot of methods are supported in its call-routine library.

The macro is defined with keyword parameters. The arguments may be listed within parentheses in any order, separated by commas. For example:

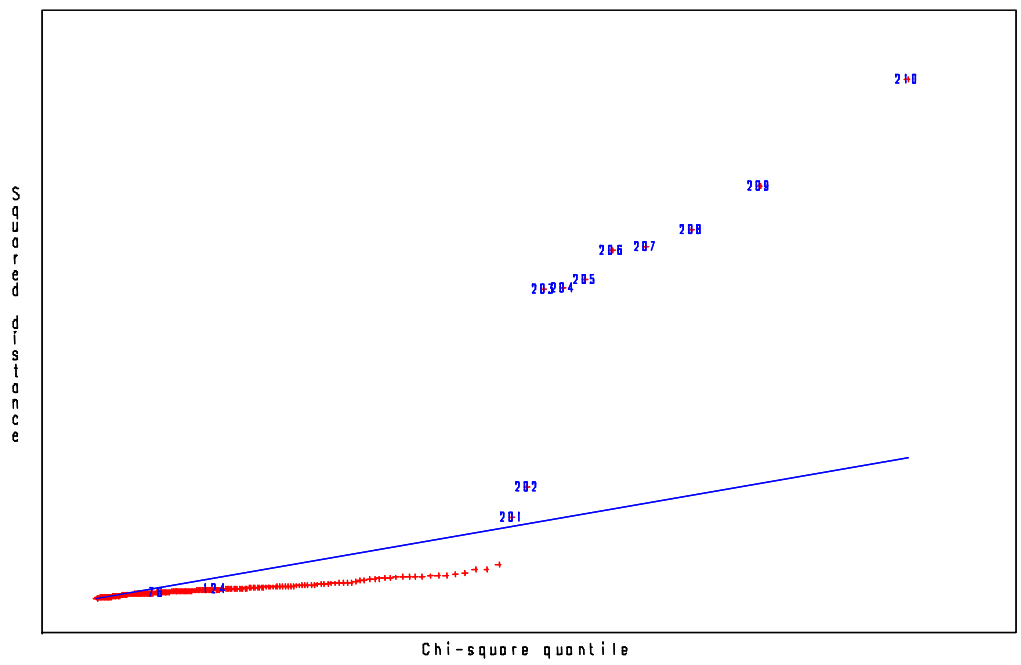
%outliers (data=test, var=x y, method=MCD, plot=y, qqplot=y, out=out);

The rules for classifying observations as potential outliers are based on the following methods: **STD** (Standard Deviation), **IQR** (Interquartile Range), **MAD** (Median Absolute Deviation), **MVE** (Minimum Volume Ellipsoid, default), **MCD** (Minimum Covariance Determinant), **MHL** (Robust Mahalanobis Distances).

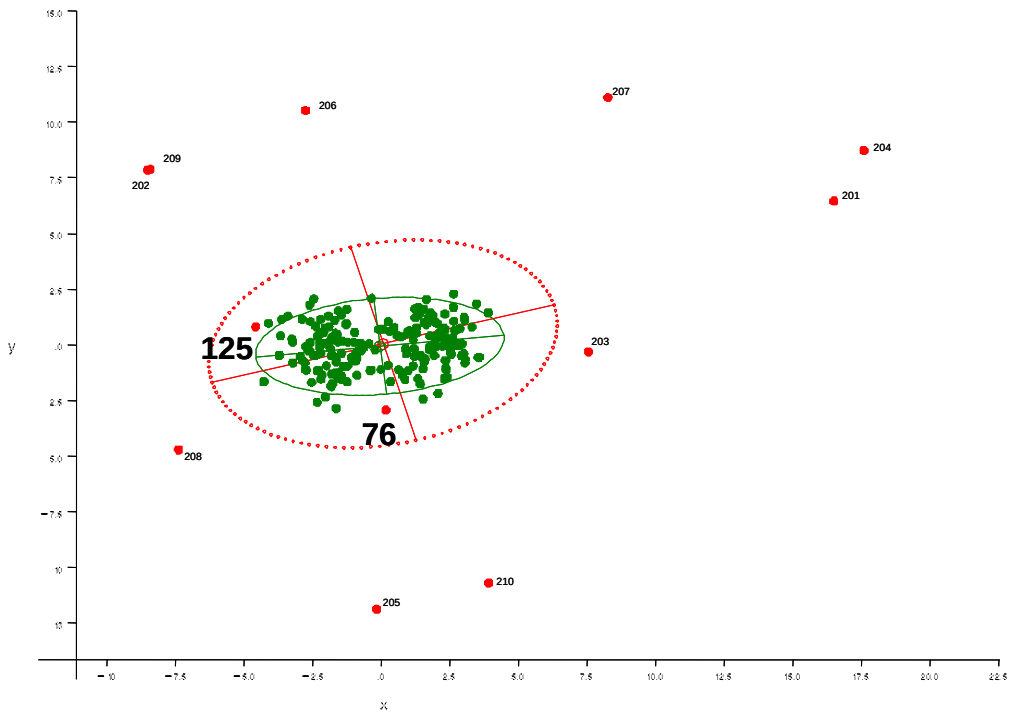
Comparison of results *obtained by all the methods*

Outlier	X	Y	STD1	STD2	IQR1	IQR2	MAD1	MAD2	MVE	MCD	MHL
27	**1,5198**	**-2,4317**								√	√
53	2,6456	2,2957									√
76	**0,1622**	**-2,9284**					√		√	√	√
82	-0,9075	-0,4825									
85	**2,0684**	**-2,1802**								√	
104	-1,7987	-0,4298									
120	-0,0870	0,6941									
124	**-1,6508**	**-2,8541**							√	√	√
125	**-4,5766**	**0,8138**									√
134	**-2,4626**	**2,0641**								√	√
141	**-2,3458**	**-2,5776**								√	√
169	-2,6072	-0,2944									
170	-1,2368	-0,9273									
175	-0,4046	-1,1466									
177	-4,2869	-1,6436									
201	16,4976	6,4640	√	√	√	√	√	√	√	√	√
202	-8,5131	7,8566	√	√	√	√	√	√	√	√	√
203	7,5461	-0,3111	√		√		√		√	√	√
204	17,5795	8,7282	√	√	√	√	√	√	√	√	√
205	-0,1681	-11,8663	√	√	√	√	√	√	√	√	√
206	-2,7676	10,5338	√	√	√	√	√	√	√	√	√
207	8,2638	11,1083	√	√	√	√	√	√	√	√	√
208	-7,3873	-4,6949	√		√		√	√	√	√	√
209	-8,4105	7,8918	√	√	√	√	√	√	√	√	√
210	3,9173	-10,6898	√	√	√	√	√	√	√	√	√

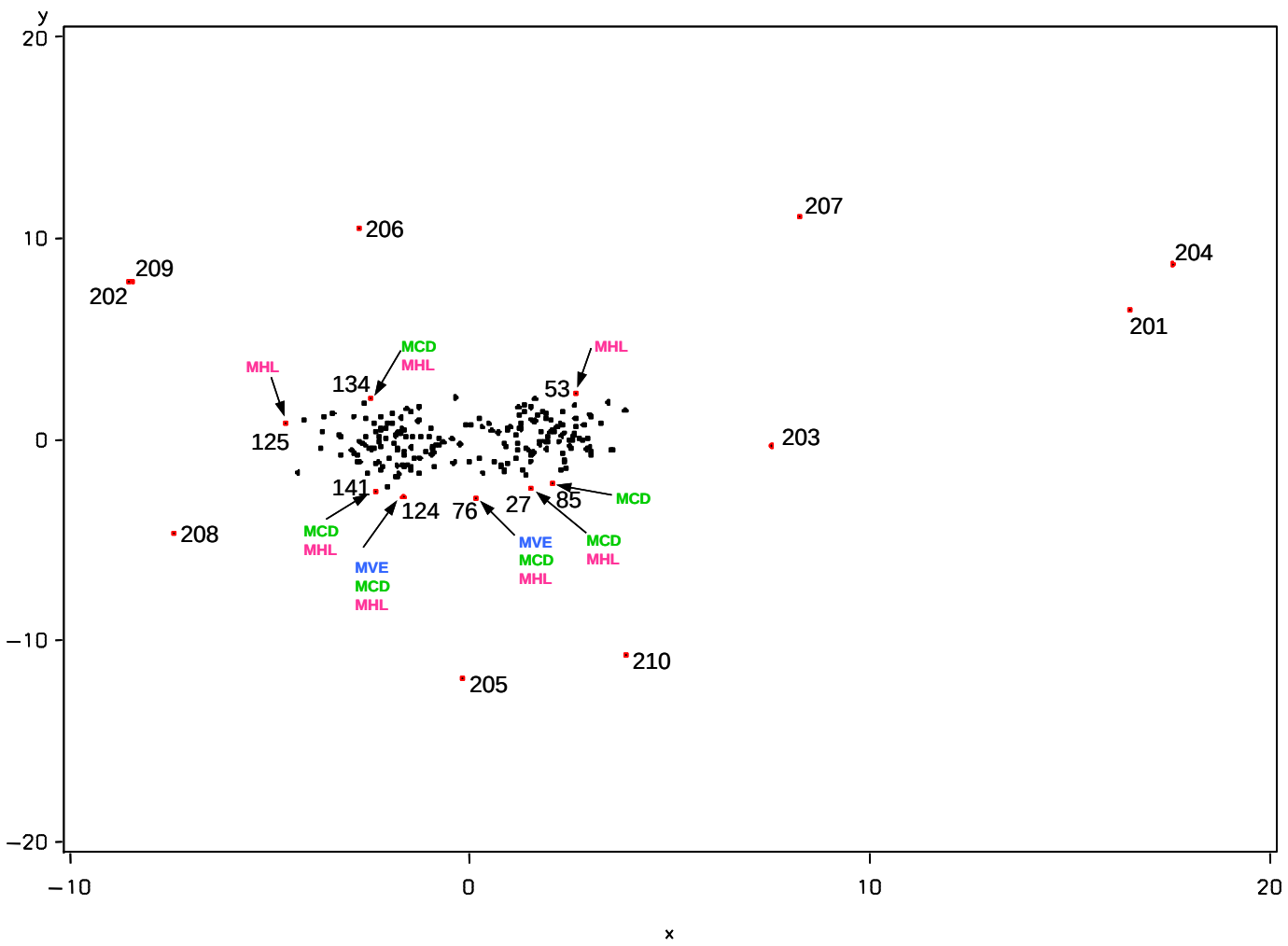
Quantile-quantile plot *for multivariate normal data (based on squared Mahalanobis distances from the centroid)*



Scatter plot of data with ***classical and robust tolerance ellipsoids***



Scatter plot of data with ***points labeled by robust outlier detection***



Performance test

The environment was an IBM eServer p5 p595 4 processors 16Gb RAM. The macro was tested running SAS 8.2 limited to a maximum RAM size of 512Mb on a large “real-life” dataset containing 12 variables and a different number of observations as showed below:

Variables	Observations	Method	Passed
8	2.000.000	***MCD(*)***	√
		MVE	√
	2.500.000	***MCD***	*Memory allocation problem*
		MVE	√
12	1.500.000	***MCD***	√
		MVE	√
	2.000.000	***MCD***	*Memory allocation problem*
		MVE	√

(*) The algorithm for the MCD subroutine is based on the FAST-MCD algorithm given by Rousseeuw and Van Driessen (1999).

A Case Study

This macro was used to detect possible outliers in a data table used in a geo-marketing analysis carried out in order to find optimal locations for new branches in Milan (Italy). The size of this table was 6 columns -representing economical profitability data of banked families- by 5800 rows (the electoral divisions).

Partial output from the OUTLIERS macro

Sezione Censimento	_2_STD	_1_5_IQR	MAD	MVE	MCD	MHL
4490						
4496						√
4504					√	
4510						√
4518					√	
4519						√
4521				√	√	√
4522						√
4533					√	√
4537						
4538						
4539						
4540						√
4541						
4544						

All the robust methods reveal the electoral division 4521 as outlier. An ex-post socio-demographic analysis reported a high percentage of old people concentrated in the biggest residential care facility for the elderly of the city. The decision was to exclude such observation from the data as not relevant for the analysis.