

# User manual for software from article

## **Algorithm xxx: The OutlierLib – a MATLAB library for outliers’ detection**

Alexander Novoselsky, Weizmann Institute of Science  
Eugene Kagan, Ariel University

This user manual contains the syntax, descriptions, examples of usage, expected results and references to origin of example data for each function of MATLAB library for outliers’ detection.

## **Contents**

Tukey test to screen data for outliers.....	2
Modified Z-score to screen data for outliers.....	4
One-sided Dixon test for single outlier.....	6
Two-sided Grubbs test for single outlier.....	8
Two-sided Generalized (extreme Studentized deviate) ESD test for one or more outliers.....	10
Two-sided Tietjen-Moore test for multiple outliers.....	13
Test for outliers in multivariate data using Mahalanobis distance and F-test.	15

## Tukey test to screen data for outliers

### Syntax

`[outliers, idx] = tukey(x, Name, Value)`

### Input arguments

`x` – Input array

### Name-Value pair arguments

Specify 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 up to 2 name and value pair arguments. Any order of the Name-Value pairs is allowed.

'whisker' – Maximum whisker length (optional)

1.5 (default) | positive numeric value

'plotBoxplot' – Boxplot is plotted (optional)

true | false (default)

### Output arguments

`outliers` – Outliers

`idx` – Indices of outliers in input array

### Description

`tukey(x, Name, Value)` screens the input array `x` for multiple outliers presence.

### Example of usage

Screen the input array inputArray for multiple outliers presence and show the boxplot.

```
[outliers, idx] = tukey(inputArray, 'plotBoxplot', true);
```

Input:

```
load carsmall
```

```
idx = strmatch('Germany', Origin);
```

```
inputArray = MPG(idx);
```

Expected results:

```
outliers      = 44
```

```
idx           = 9
```

Output to screen:

```
no output
```

Reference to origin of example data:

MATLAB sample data 'carsmall.mat'. Measurements of cars, 1970, 1976, 1982.

Miles per gallon (MPG) measurements for Germany from sample data is used.

<http://www.mathworks.com/help/stats/boxplot.html>

## Modified Z-score to screen data for outliers

### Syntax

`[outliers, idx] = mzscore(x)`

### Input arguments

`x` – Input array

### Output arguments

`outliers` – Potential outliers

`idx` – Indices of potential outliers in input array

### Description

`mzscore(x)` screens the input array `x` for multiple outliers presence. Test supposes an approximately normal distribution of input data.

### Example of usage

Screens the input array `inputArray` for multiple outliers presence.

```
[outliers, idx] = mzscore(inputArray);
```

Input:

```
inputArray = [2.0 0.1 0.8 0.2 3.0 1.9 1.0 14.6 4.8 0.4 0.9 0.1 0.3 0.3];
```

Expected results:

```
outliers      = 14.6000 4.8000
```

idx = 8 9

Output to screen:

no output

Reference to origin of example data:

United Nations Office on Drugs and Crime (UNODC). 2011.

Z-Score Report per Substance. Page 2.

Round: 2011/1. Substance: BS-2/6-Monoacetylmorphine (6-MAM).

<https://www.unodc.org/documents/scientific/ZScorePerSubstanceBS.pdf>

## One-sided Dixon test for single outlier

### Syntax

```
[outliers, idx] = dixon(x, Name, Value)
```

### Input arguments

x – Input array (size should be between 3 and 30)

### Name-Value pair arguments

Specify 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 up to 2 name and value pair arguments. Any order of the Name-Value pairs is allowed.

'alpha' – Significance level (optional)

value between 0 and 1 (if not provided, default is 0.05 for 5% significance)

'verboseOutput' – Verbose output (optional)

'on' | 'off' (default)

### Output arguments

outlier – Outlier

idx – Index of outlier in input array

### Description

dixon(x, Name, Value) tests the input array x for single outlier presence. Test supposes an approximately normal distribution of input data.

### **Example of usage**

Test the input array inputArray for outlier presence with significance level 0.10 and verbose output.

```
[outlier, idx] = dixon(inputArray, 'verboseOutput', 'on', 'alpha', 0.1);
```

Input:

```
inputArray = [568 570 570 570 572 578 584 596];
```

Expected results:

no outliers

Output to screen:

Significance level: 0.10

Test kind: right-side

Test statistic: 0.461538

Critical value: 0.478911 (based on 25000 simulations)

Reference to origin of example data:

<http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/dixon.htm>

## Two-sided Grubbs test for single outlier

### Syntax

[outliers, idx] = grubbs(x, Name, Value)

### Input arguments

x – Input array

### Name-Value pair arguments

Specify 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 up to 2 name and value pair arguments. Any order of the Name-Value pairs is allowed.

'alpha' – Significance level (optional)

value between 0 and 1 (if not provided, default is 0.05 for 5% significance)

'verboseOutput' – Verbose output (optional)

'on' | 'off' (default)

### Output arguments

outlier – Outlier

idx – Index of outlier in input array

### Description



grubbs(x, Name, Value) tests the input array x for single outlier presence. Test supposes an approximately normal distribution of input data.

### **Example of usage**

Test the input array inputArray for outlier presence with significance level 0.10 and verbose output.

```
[outlier, idx] = grubbs(inputArray, 'verboseOutput', 'on', 'alpha', 0.1);
```

Input:

```
inputArray = [199.31, 199.53, 200.19, 200.82, 201.92, 201.95, 202.18, 245.57];
```

Expected results:

```
outlier = 245.5700
```

```
idx      = 8
```

Output to screen:

```
Significance level:    0.10
```

```
Test statistic:       2.468765
```

```
Critical value:       2.031652
```

Reference to origin of example data:

<http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h1.htm>

## Two-sided Generalized (extreme Studentized deviate) ESD test for one or more outliers

### Syntax

```
[outliers, idx] = gesd(x, Name, Value)
```

### Input arguments

x – Input array

### Name-Value pair arguments

Specify 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 up to 3 name and value pair arguments. Any order of the Name-Value pairs is allowed.

'outliersNumber' – Upper bound for suspected number of outliers (mandatory)

'alpha' – Significance level (optional)

value between 0 and 1 (if not provided, default is 0.05 for 5% significance)

'verboseOutput' – Verbose output (optional)

'on' | 'off' (default)

### Output arguments

outliers – Outliers

idx – Indices of outliers in input array

## Description

`gesd(x, Name, Value)` tests the input array `x` for one or more outliers presence. Test supposes an approximately normal distribution of input data.

## Example of usage

Test the input array `inputArray` for presence of 10 outliers with verbose output.

```
[outliers, idx] = gesd(inputArray, 'outliersNumber', 10, 'verboseOutput', 'on');
```

Input:

```
inputArray = [-0.25 0.68 0.94 1.15 1.20 1.26 1.26 1.34 1.38 1.43 1.49 1.49 1.55 1.56 1.58  
1.65 1.69 1.70 1.76 1.77 1.81 1.91 1.94 1.96 1.99 2.06 2.09 2.10 2.14 2.15 2.23 2.24 2.26  
2.35 2.37 2.40 2.47 2.54 2.62 2.64 2.90 2.92 2.92 2.93 3.21 3.26 3.30 3.59 3.68 4.30 4.64  
5.34 5.42 6.01];
```

Expected results:

```
outliers      = 6.0100 5.4200 5.3400
```

```
idx           = 54 53 52
```

Output to screen:

Significance level: 0.05

Outliers: 1, Test statistic: 3.119, critical value: 3.159

Outliers: 2, Test statistic: 2.943, critical value: 3.151

Outliers: 3, Test statistic: 3.179, critical value: 3.144

Outliers: 4, Test statistic: 2.810, critical value: 3.136

Outliers: 5, Test statistic: 2.816, critical value: 3.128

Outliers: 6, Test statistic: 2.848, critical value: 3.120

Outliers: 7, Test statistic: 2.279, critical value: 3.112

Outliers: 8, Test statistic: 2.310, critical value: 3.103

Outliers: 9, Test statistic: 2.102, critical value: 3.094

Outliers: 10, Test statistic: 2.067, critical value: 3.085

Number of outliers: 3

Reference to origin of example data:

<http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm>

## Two-sided Tietjen-Moore test for multiple outliers

### Syntax

`[outliers, idx] = tietjen(x, Name, Value)`

### Input arguments

`x` – Input array

### Name-Value pair arguments

Specify 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 up to 3 name and value pair arguments. Any order of the Name-Value pairs is allowed.

`'outliersNumber'` – Number of outliers (mandatory)

`'alpha'` – Significance level (optional)

value between 0 and 1 (if not provided, default is 0.05 for 5% significance)

`'verboseOutput'` – Verbose output (optional)

`'on' | 'off'` (default)

### Output arguments

`outliers` – Outliers

`idx` – Indices of outliers in input array

## Description

tietjen(x, Name, Value) tests the input array x for multiple outliers presence. Test supposes an approximately normal distribution of input data.

## Example of usage

Test the input array inputArray for presence of 2 outliers with verbose output.

```
[outliers, idx] = tietjen(inputArray, 'outliersNumber', 2, 'verboseOutput', 'on');
```

Input:

```
inputArray = [-1.40 -0.44 -0.30 -0.24 -0.22 -0.13 -0.05 0.06 0.10 0.18 0.20 0.39 0.48 0.63  
1.01];
```

Expected results:

```
outliers      = -1.4000 1.0100
```

```
idx           = 1 15
```

Output to screen:

```
Significance level: 0.05
```

```
Test statistic: 0.291999
```

```
Critical value: 0.316691 (based on 10000 simulations)
```

Reference to origin of example data:

<http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h2.htm>

## Test for outliers in multivariate data using Mahalanobis distance and F-test

### Syntax

```
[outliers, idx] = mahdist(x, Name, Value)
```

### Input arguments

x – Array of multivariate data (samples of each variate in separate row)

### Name-Value pair arguments

Specify 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 up to 2 name and value pair arguments. Any order of the Name-Value pairs is allowed.

'alpha' – Significance level (optional)

value between 0 and 1 (if not provided, default is 0.05 for 5% significance)

'verboseOutput' – Verbose output (optional)

'on' | 'off' (default)

### Output arguments

outliers – Outliers

idx – Indices of outliers in input array

### Description

mahdist(x1, x2, Name, Value) tests the input arrays x1 and x2 for outliers presence. Test supposes an approximately normal multivariate distribution of input data.

### Example of usage

Test the input bivariate array inputArray for outliers presence with verbose output.

```
[outlier, idx] = mahdist(inputArray, 'verboseOutput', 'on');
```

Input:

```
inputArray = [154 136 91 125 133 125 93 80 132 107 142 115 114 120 141; 108 90 54 89  
93 77 43 50 125 76 96 74 79 71 90];
```

Expected results:

outlier =

132 93

125 43

idx = 9 7

Output to screen:

Significance level: 0.05

Variates: 2

Samples: 15

-----

Sample Test statistic P

1 1.52 0.258058

2 0.36 0.705487

3 1.09 0.367818



4	0.08	0.921651
5	0.17	0.844478
6	0.27	0.767123
7	1.95	0.185275
8	2.64	0.112036
9	16.38	0.000372
10	0.41	0.674741
11	0.61	0.560950
12	0.05	0.952880
13	0.10	0.904111
14	0.36	0.705578
15	0.87	0.441904

min P: 0.000372, Sample: 9 (132.0 125.0)

-----

Sample Test statistic P

1	1.56	0.253543
2	0.30	0.746988
3	0.96	0.411185
4	0.68	0.526699
5	0.43	0.662707
6	0.39	0.683226
7	5.62	0.020854
8	2.82	0.102589
9	1.36	0.296418
10	0.54	0.595561

11	0.02	0.978085
12	0.50	0.617706
13	0.71	0.511994
14	0.96	0.413978

min P: 0.020854, Sample: 7 (93.0 43.0)

-----

Sample Test statistic P

1	1.74	0.224137
2	0.26	0.779609
3	1.61	0.246982
4	0.87	0.446666
5	0.50	0.619839
6	0.85	0.454451
7	2.91	0.101114
8	1.41	0.288811
9	0.47	0.638841
10	0.09	0.913226
11	0.48	0.632789
12	1.87	0.203993
13	1.09	0.374218

min P: 0.101114, Sample: 7 (80.0 50.0)

Reference to origin of example data:

Afifi, A.A. and Azen, S.P. 1979. *Statistical Analysis: A Computer Oriented Approach* (2<sup>nd</sup> ed.). Academic Press. Chapter 5.1.