Robust Correlation toolbox Manual

The toolbox allows performing robust correlations analyses along with various assumption checks and data visualization.

When using the toolbox for article please cite:

Pernet, C.R., Wilcox, R. & Rousselet, G.A. (2012). Robust correlation analyses: old tools but new toolbox for psychology research. *Front. in Psychology*.

Since you are likely to use the skipped-correlation which depends on an estimation of the robust centre of the data, then also cite:

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," Journal of the American Statistical Association, Vol. 79, pp. 871-881.

Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, 41, pp. 212-223.

Verboten, S., & Hubert, M. (2005). LIBRA: a MATLAB Library for Robust Analysis, *Chemometrics and Intelligent Laboratory Systems*(75), 127–136.

If the test of multivariate normality is reported cite:

Trujillo-Ortiz, A., R. Hernandez-Walls, K. Barba-Rojo and L. Cupul-Magana. (2007). HZmvntest:Henze-Zirkler's Multivariate Normality Test. A MATLAB file.

Table of Contents

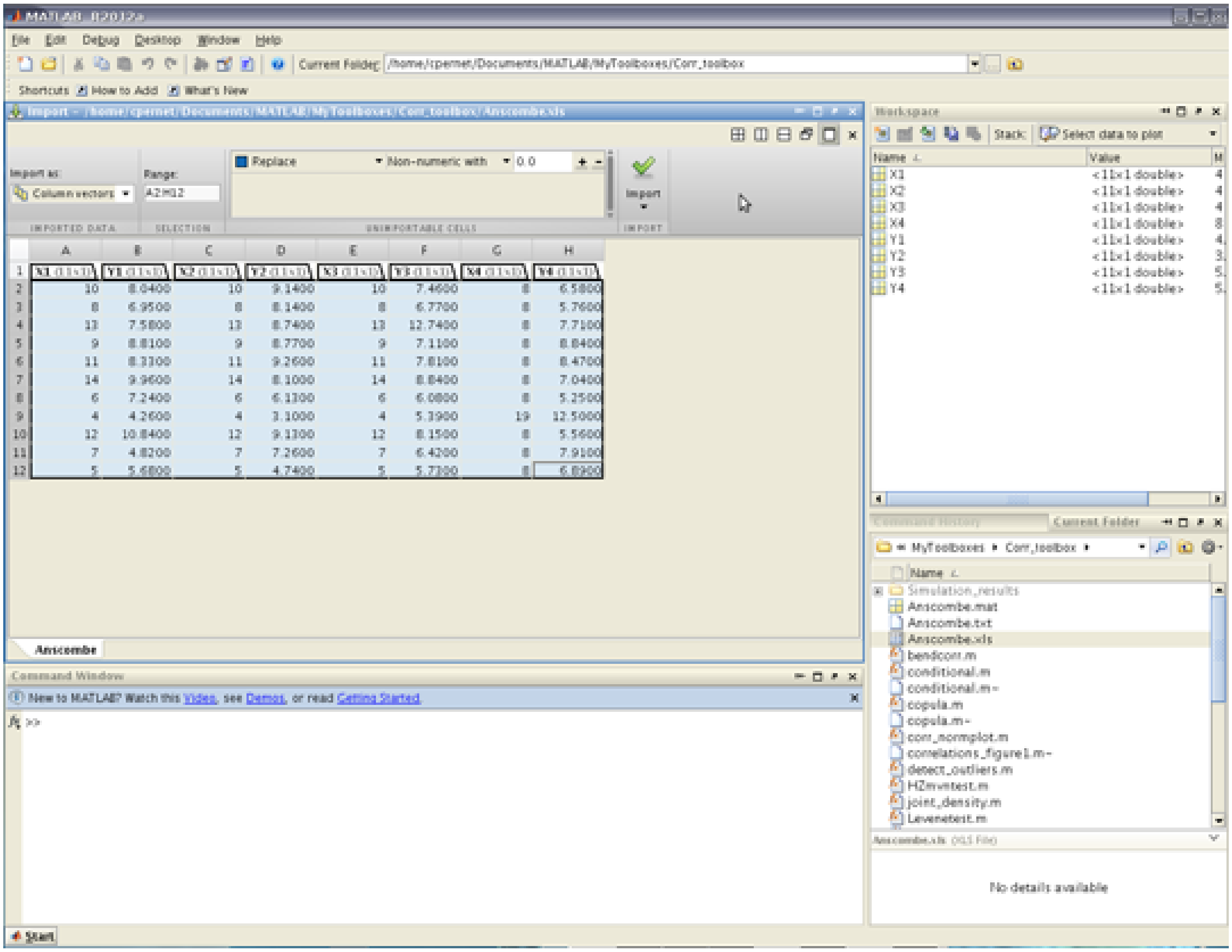
1 – Before starting

If you are not familiar with Matlab, all you need to do to get the toolbox to work is to set the path. That means that you tell Matlab where the toolbox is located. The easiest way to do so is to click at the top of the Matlab window on *File →Set Path* , then click on Add with Subfolders and select the Corr_toolbox.

2 – Importing data in Matlab

For a Matlab novice it might seem complicated to import data. This is however as easy as with any other software. With the software functions, we also saved the Anscombe's quartet data.

Import excel file data: the easiest way is, from the Malab 'Current Folder' window, double click on Anscombe.xls. The import wizard is now open and the data are already selected (figure 1). At the top left of the wizard, you have the choice about how to import, select 'Column vectors' and then click on 'Import'. You should now have in the 'Workspace' 8 variables called X1, X2, X3, X4, Y1, Y2, Y3, Y4.



MATLAB R2012a
File Edit Debug Desktop Window Help
Current Folder: /home/cpernet/Documents/MATLAB/MyToolboxes/Corr_toolbox
Shortcuts How to Add What's New
Import - /home/cpernet/Documents/MATLAB/MyToolboxes/Corr_toolbox/Anscombe.xls
Replace
Non-numeric with
0.0
Import
IMPORTED DATA
SELECTION
UNIMPORTABLE CELLS
IMPORT
Anscombe
Workspace
Stack:
Select data to plot
Command Window
Command History
Current Folder
MyToolboxes ▸ Corr_toolbox ▸
Start

Import as: Column vectors

Range: A2:H12

	A	B	C	D	E	F	G	H
1	X1 (11x1)	Y1 (11x1)	X2 (11x1)	Y2 (11x1)	X3 (11x1)	Y3 (11x1)	X4 (11x1)	Y4 (11x1)
2	10	8.0400	10	9.1400	10	7.4600	8	6.5800
3	8	6.9500	8	8.1400	8	6.7700	8	5.7600
4	13	7.5800	13	8.7400	13	12.7400	8	7.7100
5	9	8.8100	9	8.7700	9	7.1100	8	8.8400
6	11	8.3300	11	9.2600	11	7.8100	8	8.4700
7	14	9.9600	14	8.1000	14	8.8400	8	7.0400
8	6	7.2400	6	6.1300	6	6.0800	8	5.2500
9	4	4.2600	4	3.1000	4	5.3900	19	12.5000
10	12	10.8400	12	9.1300	12	8.1500	8	5.5600
11	7	4.8200	7	7.2600	7	6.4200	8	7.9100
12	5	5.6800	5	4.7400	5	5.7300	8	6.8900

Name	Value
X1	<11x1 double>
X2	<11x1 double>
X3	<11x1 double>
X4	<11x1 double>
Y1	<11x1 double>
Y2	<11x1 double>
Y3	<11x1 double>
Y4	<11x1 double>

New to MATLAB? Watch this Video, see Demos, or read Getting Started.

>>

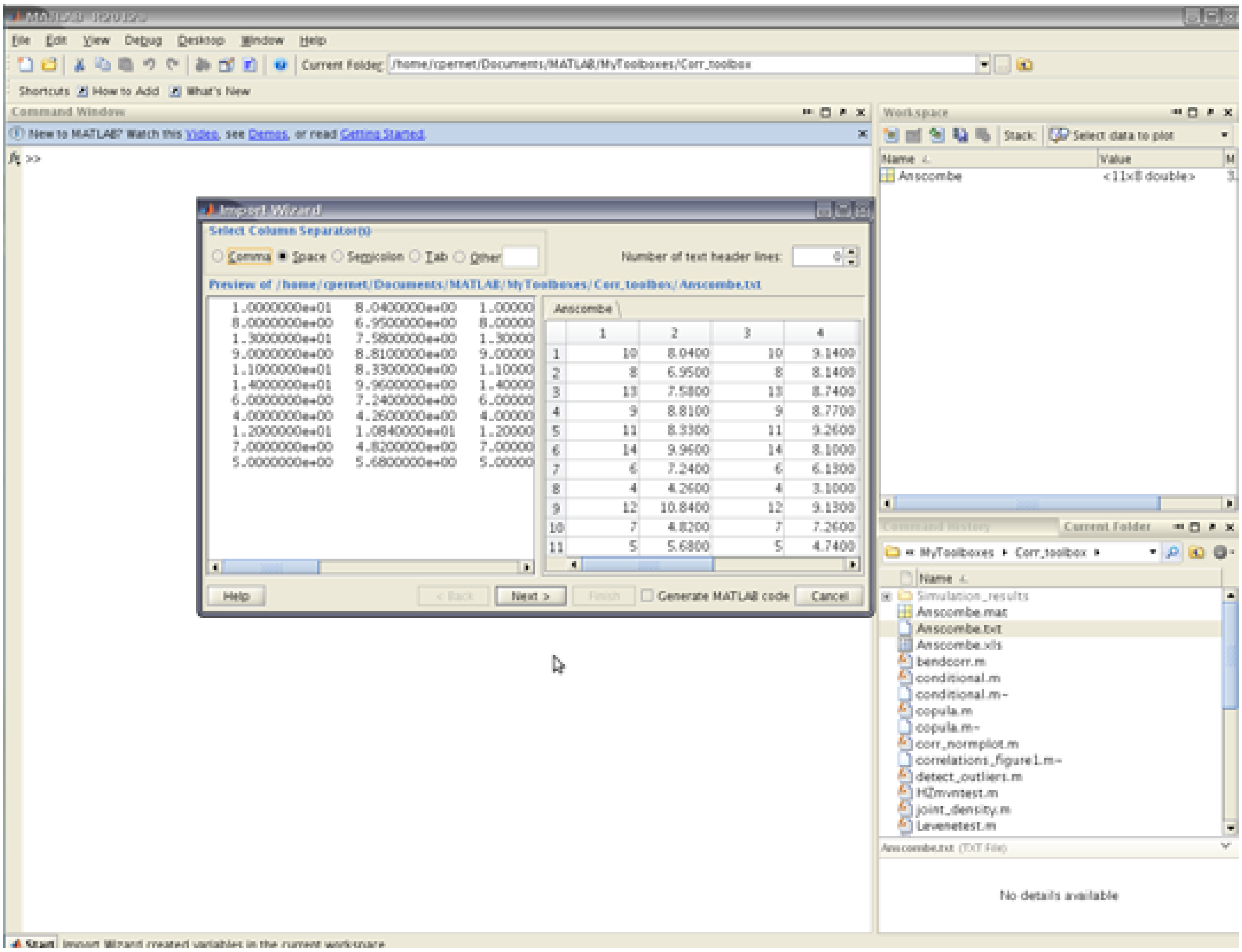
Name
Simulation_results
Anscombe.mat
Anscombe.txt
Anscombe.xls
bendcorr.m
conditional.m
conditional.m~
copula.m
copula.m~
corr_normplot.m
correlations_figure1.m~
detect_outliers.m
HZmvntest.m
joint_density.m
Levenetest.m

Anscombe.xls (XLS File)

No details available

Figure 1. Matlab import wizard for the excel file 'Anscombe.xls'.

Import text file data: the easiest way is, from the Malab 'Current Folder' window, right click on Anscombe.txt, and select Import Data. The import wizard is now open (figure 2). Simply click next and finished. You should now have all the variables loaded in the 'Workspace' as a single matrix called Anscombe.



Select Column Separator(s)

○ Comma ● Space ○ Semicolon ○ Tab ○ Other

Number of text header lines: 0

Preview of /home/cpernet/Documents/MATLAB/MyToolboxes/Corr_toolbox/Anscombe.txt

```
  1.0000000e+01   8.0400000e+00   1.00000
  8.0000000e+00   6.9500000e+00   8.00000
  1.3000000e+01   7.5800000e+00   1.30000
  9.0000000e+00   8.8100000e+00   9.00000
  1.1000000e+01   8.3300000e+00   1.10000
  1.4000000e+01   9.9600000e+00   1.40000
  6.0000000e+00   7.2400000e+00   6.00000
  4.0000000e+00   4.2600000e+00   4.00000
  1.2000000e+01   1.0840000e+01   1.20000
  7.0000000e+00   4.8200000e+00   7.00000
  5.0000000e+00   5.6800000e+00   5.00000
```

Anscombe

	1	2	3	4
1	10	8.0400	10	9.1400
2	8	6.9500	8	8.1400
3	13	7.5800	13	8.7400
4	9	8.8100	9	8.7700
5	11	8.3300	11	9.2600
6	14	9.9600	14	8.1000
7	6	7.2400	6	6.1300
8	4	4.2600	4	3.1000
9	12	10.8400	12	9.1300
10	7	4.8200	7	7.2600
11	5	5.6800	5	4.7400

☐ Generate MATLAB code

Workspace: Anscombe <11x8 double>

Import Wizard created variables in the current workspace

Figure 2. Matlab import wizard for the text file 'Anscombe.txt'.

Load data already in Matlab format: once the data have been imported, it is easier to save them in Matlab format (see help save) to be used later. Here the data can be accessed by double clicking on Anscombe.mat from the Malab 'Current Folder' window.

3 – One-step analysis between 2 variables

When only 2 variables are tested, the full range of tests and plots can be performed by simply calling the *robust_correlation.m* function. Load the data Anscombe.mat (section 2) and in the Matlab command window, type:

```
correlation_results = robust_correlation(Anscombe(:,1),Anscombe(:,2))
```

The function successively performs the following operations:

(1) Plots the data with (i) a scatter plot, (ii) the marginal (normalized) histograms with the corresponding Gaussian curves and (iii) the bivariate histogram (*corr_normplot.m*)
(2) Plots the joint density as a mesh and its isocontour (*joint_density.m*).
(3) Tests bivariate normality (*Hzmvntest.m*)
(4) Tests heteroscedasticity (*variance_honogeneity.m*)
(5) Looks for outliers (*detect_outliers.m*)
(6) Performs all types of correlations (*Pearson.m Spearman.m bendcorr.m skipped-correlation.m*)

As the different tests are performed, outputs appears in the Matlab command window and as well as graphics (figures 3 and 4). In addition, there is now a structure called correlation_results with all the results. For instance if you type correlation_results.Pearson it returns the r, t, p values, the bootstrapped confidence intervals and something telling you if it is significant or not (decision based on the bootstrapped confidence intervals).

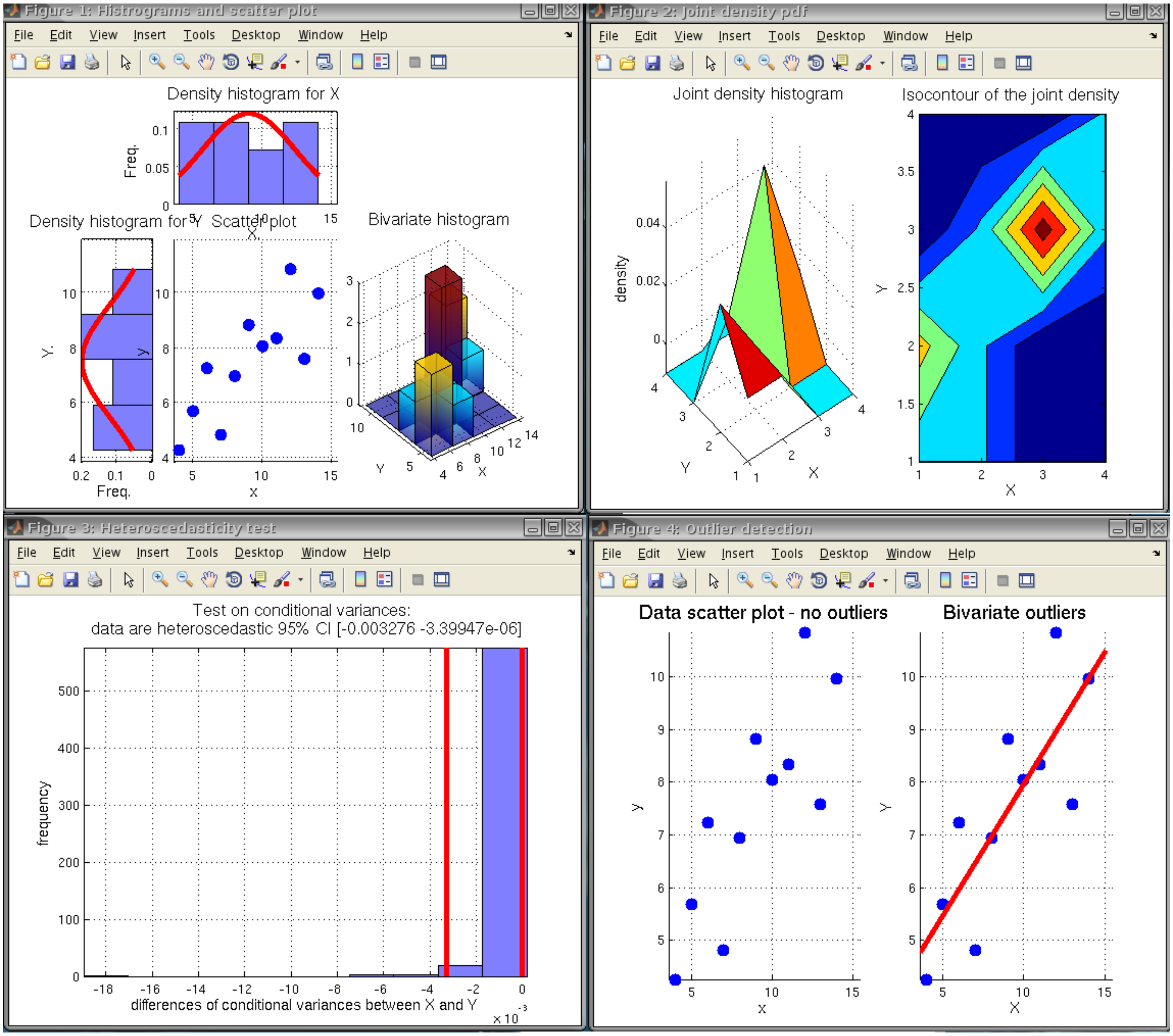


Figure 1: Histrograms and scatter plot
File Edit View Insert Tools Desktop Window Help
Density histogram for X
Freq.
0.1
0.05
0
5
10
15
X
Density histogram for Y
Scatter plot
Bivariate histogram
Y:
10
8
6
4
0.2
0.1
0
Freq.
y
x
3
2
1
0
10
Y
5
4
6
8
10
12
14
X
Figure 2: Joint density pdf
File Edit View Insert Tools Desktop Window Help
Joint density histogram
Isocontour of the joint density
density
0.04
0.02
0
4
3
2
1
Y
X
3.5
2.5
1.5
Figure 3: Heteroscedasticity test
File Edit View Insert Tools Desktop Window Help
frequency
500
400
300
200
100
0
-18
-16
-14
-12
-10
-8
-6
-4
-2
0
differences of conditional variances between X and Y
x 10
-3
Figure 4: Outlier detection
File Edit View Insert Tools Desktop Window Help
Data scatter plot - no outliers
Bivariate outliers
10
9
8
7
6
5
y
x
5
10
15

Test on conditional variances:
data are heteroscedastic 95% CI [-0.003276 -3.39947e-06]

Figure 3. Graphical outputs returned by the correlation toolbox for the 1st Anscombe's quartet: Top left (Figure 1) = scatter plots and histograms – top right (Figure 2) = mesh and isocontours of the joint density – Bottom left (Figure 3) = histogram of the differences in the conditional variances of bootstrapped data – Bottom right (Figure 4) = univariate and bivariate outliers –

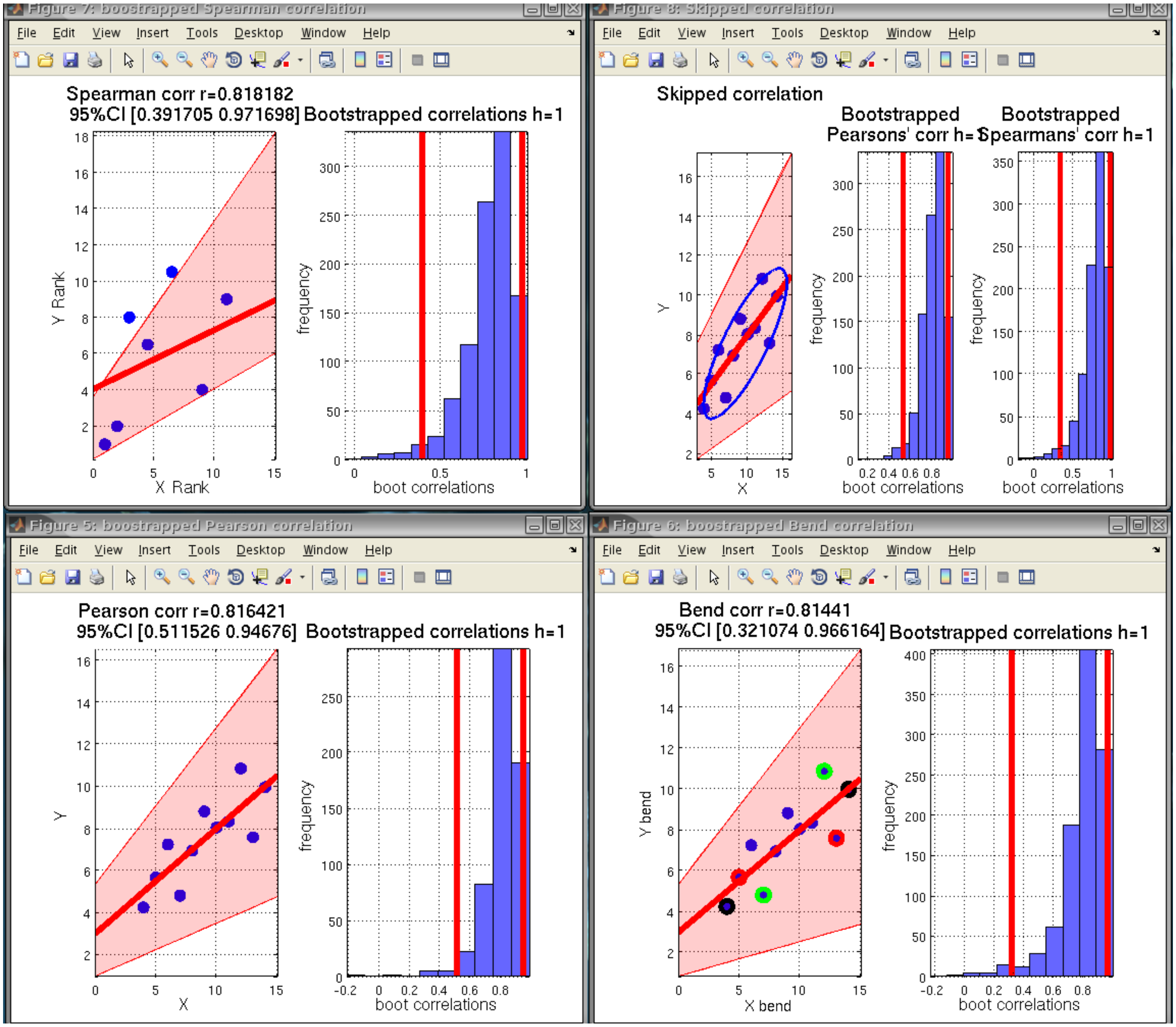


Figure 7: boostrapped Spearman correlation
File Edit View Insert Tools Desktop Window Help
Spearman corr r=0.818182
95%CI [0.391705 0.971698]
Bootstrapped correlations h=1
Y Rank
X Rank
frequency
boot correlations
Figure 8: Skipped correlation
File Edit View Insert Tools Desktop Window Help
Skipped correlation
Bootstrapped Pearsons' corr h=1
Bootstrapped Spearmans' corr h=1
Y
X
frequency
boot correlations
Figure 5: boostrapped Pearson correlation
File Edit View Insert Tools Desktop Window Help
Pearson corr r=0.816421
95%CI [0.511526 0.94676]
Bootstrapped correlations h=1
Y
X
frequency
boot correlations
Figure 6: boostrapped Bend correlation
File Edit View Insert Tools Desktop Window Help
Bend corr r=0.81441
95%CI [0.321074 0.966164]
Bootstrapped correlations h=1
Y bend
X bend
frequency
boot correlations

Figure 4. Graphical outputs returned by the correlation toolbox for the 1st Anscombe's quartet: Top-left (Figure 7) = Spearman correlation with 95% CI and the histogram of correlations for bootstrapped data – Top right (Figure 8) = Skipped Pearson correlation with 95% CI and the histograms of Pearson's and Spearman's correlations for bootstrapped data. Bottom left (Figure 5) = Pearson correlation with 95% CI and the histogram of correlations for bootstrapped data – Bottom right (Figure 6) = 20% bend correlation with 95% CI and the histogram of correlations for bootstrapped data.

4 - Calling specific function

Data visualization

To get the scatter plot with histograms type:

```
corr_normplot(Anscombe(:,1), Anscombe(:,2));
```

To get the joint density type:

```
density = joint_density(Anscombe(:,1), Anscombe(:,2));
```

It will not only plots the joint density, but also returns it as output variable (here density) such as density =

```
    0.0221    0.0221         0         0
    0.0442    0.0221         0         0
         0    0.0221    0.0663    0.0221
         0         0         0    0.0221
```

It is also possible to only plot the isocontours by typing

```
density = joint_density(Anscombe(:,1), Anscombe(:,2),0);
```

Assumption checking

To test for multivariate normality, type:

```
HZmvntest([Anscombe(:,1), Anscombe(:,2)]);
```

Note the square brackets [ ] as the functions needs a single matrix as input

Variance homogeneity can be tested simply by typing:

```
[h,CI] = variance_homogeneity(Anscombe(:,1),Anscombe(:,2));
```

If h=1 data have different variances, if h = 0 data have the same variances.
CI is the percentile bootstrap 95% confidence interval of the difference between variances.

By default the function normalizes the data and uses conditional variances (see conditional.m). It is however possible to force the function to use the original variances by typing

```
[h,CI] = variance_homogeneity(Anscombe(:,1),Anscombe(:,2),0);
```

To detect outliers using robust estimators (i.e. the normalized Median Absolute Deviation and Mid Covariance Determinant) type:

```
outliers = detect_outliers(Anscombe(:,1),Anscombe(:,2));
```

the output (outliers) is a structure with outliers.univariate.X retuning indices of univariate outliers in X, outliers.univariate.Y retuning indices of univariate outliers in Y and outliers.bivariate returning bivariate outliers (in each cases 0 means it a good data point and 1 means this is an outlier)

Correlations

Pearson

```
[r,t,p] = Pearson(Anscombe(:,1),Anscombe(:,2))
```

returns the correlation (r) value along with the t and p-values and make a plot of the data with a line representing the best linear fit

```
[r,t,p] = Pearson(Anscombe(:,1),Anscombe(:,2),0)
```

allows to compute without making a figure

```
[r,t,p,hboot,CI] = Pearson(Anscombe(:,1),Anscombe(:,2),1,10/100)
```

allows to obtain a decision about significance (hboot) based upon the bootstrapped confidence intervals (CI) at the desired type 1 level (here 10% – default is 5%). Note the difference in graphical outputs, because bootstrap was used, CI and the histogram of bootstrapped correlations are also plotted.

Spearman

```
[r,t,p] = Spearman(Anscombe(:,1),Anscombe(:,2))
```

returns the correlation (r) value along with the t and p-values and make a plot of the data with a line representing the best linear fit. Note the difference with Pearson – the scatter plot is for the ranked data.

```
[r,t,p] = Spearman(Anscombe(:,1),Anscombe(:,2),0)
```

allows to compute without making a figure

```
[r,t,p,hboot,CI] = Spearman(Anscombe(:,1),Anscombe(:,2),1,10/100)
```

allows to obtain a decision about significance (hboot) based upon the bootstrapped confidence intervals (CI) at the desired type 1 level (here 10% - default is 5%). Note the difference in graphical outputs, because bootstrap was used, CI and the histogram of bootstrapped correlations are also plotted.

Percentage bend correlation

```
[r,t,p] = bendcorr(Anscombe(:,1),Anscombe(:,2))
```

returns the correlation (r) value along with the t and p-values and make a plot of the data with a line representing the best linear fit. In the graphical output different colors are used to indicate which data points were weighted down (red for data in X, green for data in Y and black if in both X and Y). The percentage bend correlation is not an estimate of Pearson's correlation – it is however a measure of the linear relationship between X and Y.

```
[r,t,p,hboot,CI] = bendcorr(Anscombe(:,1),Anscombe(:,2),0,40)
```

allows to obtain a decision about significance (hboot) based upon the bootstrapped 95% confidence intervals (CI) – additional arguments indicate not to plot the data (0) and use 40% bending rather than the default 20%.

Skipped-correlation

```
[r,t,h] = skipped_correlation(Anscombe(:,1),Anscombe(:,2))
[r,t,h] = skipped_correlation(Anscombe(:,1),Anscombe(:,2),0)
```

returns the correlation (r) values along with the t values of Pearson and

Spearman tests performed on data after removing bivariate outliers. Note no p value is computed, but the T value is thresholded such as a decision h can be returned for a level alpha = 5%. This is performed with adjustments related to the sample size to maintain the type 1 error rate. By default, the function also makes a plot of the data with an ellipse containing non outlying data and a line representing the best linear fit to the remaining data points. It is essential to understand that outlier removal is based on normality, and Spearman is computed on the same data as Pearson.

[r,t,h,outid,hboot,CI] = skipped_correlation(Anscombe(:,1),Anscombe(:,2))
Also returns outlier data in outid (same as using outlier_detect) and add 95% bootrapped confidence intervals (CI) with the decision about significance (hboot).

<u>5 – Multiple testing solutions</u>

When performing multiple tests, we increase the chances to make a false positive error. This is only true for families, i.e when tests relate to each other. That if you do 2 tests on pairs of variables not related (at all) to each other, then it's fine. If on the other hand variables are related, errors cumulate. For instance for two tests each set at 5% of making an error, we have in fact 7.5% chances (5% +5%*5%) to make an error over the two tests. We illustrate here how to correct for this using the pairs 1, 2 and 3 of the Anscombe's quartet. For each pair X is the same, so we clearly have a family of tests.

<u>Pearson and Spearman</u>
[r,t,p,hboot,CI] = Pearson(Anscombe(:,1),Anscombe(:,[2 4 6]))
[r,t,p,hboot,CI] = Pearson(Anscombe(:,[1 3 5]),Anscombe(:,[2 4 6]))
[r,t,p,hboot,CI] = Spearman(Anscombe(:,1),Anscombe(:,[2 4 6]))
[r,t,p,hboot,CI] = Spearman(Anscombe(:,[1 3 5]),Anscombe(:,[2 4 6]))

Here either one vector and a matrix or two matrices of data were input, and correlations are always computed column-wise (the function will simply repeats the vector Anscombe(:,1) for 1 vector and a matrix as inputs)

Multiple comparison correction is simply accounted for using Bonferroni correction. That is only p values below say 1.67% (5% / 3 tests) should be considered and, accordingly, confidence intervals are adjusted to 98.33%.

<u>Percentage bend correlation</u>
[r,t,p,hboot,CI,H,pH] = bendcorr(Anscombe(:,1),Anscombe(:,[2 4 6]))
[r,t,p,hboot,CI,H,pH] = bendcorr(Anscombe(:,[1 3 5]),Anscombe(:,[2 4 6]))

As for Pearson and Spearman, a vector and a matrix or 2 matrices can be input and correlations are computed column-wise. Also, hboot and CI are adjusted using a Bonferonni correction.

In output, H and pH can also be added to compute an omnibus test of independence between all pairs, i.e. testing that the matrix of correlation is equal to the identify matrix (0 everywhere expect the diagonal being 1). H is the measure of association between all pairs (like r is the measure between two variables for 1 pair) and pH is the p value associated to this test.

Skipped-correlation

```
[r,t,h,outid,hboot,CI] = skipped_correlation(Anscombe(:,1),Anscombe(:,[2 4 6]))
```

A vector and a matrix are used as inputs, and h returns the significance with a adjustment for multiple comparisons given we have only Anscombe(:,1) against 3 other variables.

```
[r,t,h,outid,hboot,CI] = skipped_correlation(Anscombe(:,[1 3 5]),Anscombe(:,[2 4 6]))
```

Two matrices are used as inputs, and h returns the significance with a adjustment for multiple comparisons testing that all correlations are 0

In both cases, r and t are only computed for Spearman, because only Spearman provides a good type 1 error rate in the context of multiple comparisons – hboot and CI are adjusted using a Bonferroni correction.

6 - References

Henze, N. and Zirkler, B. (1990), A Class of Invariant Consistent Tests for Multivariate Normality. Communication in Statistics – Theory and Methods, 19(10): 3595-3618.

Pernet, C.R., Wilcox, R. & Rousselet, G.A. (2012). Robust correlation analyses: old tools but new toolbox for psychology research. *Front. in Psychology*.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," Journal of the American Statistical Association, Vol. 79, pp. 871-881.

Rousseeuw, P.J. and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technometrics, 41, pp. 212-223.

Rousselet, R & Pernet, C. (2012). Improving brain-behavioural correlations. *Front in Hum Neurosc, 6: 119*

Trujillo-Ortiz, A., R. Hernandez-Walls, K. Barba-Rojo and L. Cupul-Magana. (2007). HZmvntest:Henze-Zirkler's Multivariate Normality Test. A MATLAB file. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17931

Verboten, S., & Hubert, M. (2005). LIBRA: a MATLAB Library for Robust Analysis, *Chemometrics and Intelligent Laboratory Systems*(75), 127–136.

Wilcox, R. (2012). Introduction to robust estimation and hypothesis testing. 3rd Edition. Elsevier, *Academic Press, Oxford, UK.*