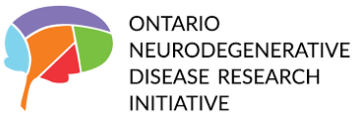


Outliers App 0.10.2.9000 - Reference Guide

On behalf of the Neuroinformatics and Biostatistics (NIBS) Platform

Updated as of September 28th, 2020

This reference guide helps you install and use the ONDRI Outliers App, with instructions on how to run the app, and some of the common errors while using the app.

Contents

1 Overview

The purpose of this app is to flag suspicious observations that may require follow up by data curators to confirm their validity. We use a combination of bootstrapping, minimum covariance determinants(MCD), principle component analysis/correspondence analysis (PCA/CA), and CorrMax plots.

For more details about the method, see Beaton et al. (2019).

2 Installation

2.1 Installing the OuRS package

To use the Outliers App, the Outliers and Robust Structures (OuRS) package is required. We recommend using the devtools package for the installation, which can be installed with:

```
install.packages(devtools)
library(devtools)
```

Then to install OuRS:

```
devtools::install("derekbeaton/GSVD")
devtools::install("derekbeaton/OuRS", subdir = "/OuRS")
```

2.2 Installing Outliers App

```
install_version("shiny", version = "1.2.0")
install_version("rmarkdown", version = "1.10")
install_version("markdown", version = "0.8")
install_version("RCurl", version = "1.95-4.11")
install_version("knitr", version = "1.20")
install_version("pheatmap", version = "1.0.10")
install_version("corrplot", version = "0.84")
install_version("pander", version = "0.6.2")
install_version("abind", version = "1.4-5")

install_version("shinyjs", version = "1.1")
install_version("plotly", version = "4.9.2.1")
install_version("stringr", version = "1.4.0")
install_version("viridis", version = "0.3.0")
```

Then you can clone the Outliers repository and run app.R to launch the app. Note, you can attempt to use other versions of these packages if you already have them installed (which in many cases should be fine), but we do not guarantee that the app will function correctly if you do so.

3 Running App

You may either run the app in your browser, or in a separate window. Choose where you will run the app in the RStudio editor, then click the Run App button, or press Ctrl-Alt-R to run the app.

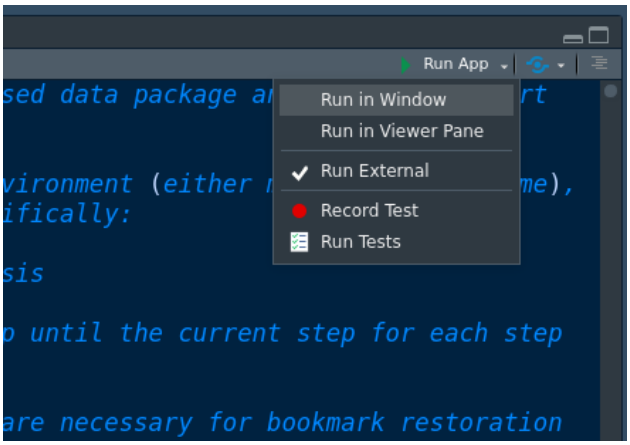


Figure 1: Run external will run the app in your default browser. You can also run the app in a seperate window

You may also refresh the app while it is running by pressing the refresh button on the top left corner of the app, or press the run hotkey Ctrl-Alt-R.

4 How to Use

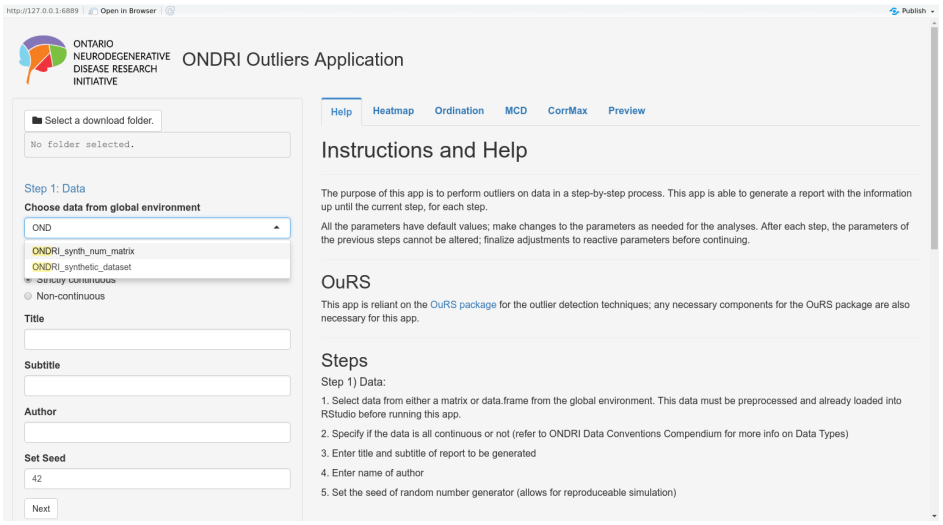
To start, load your data into your RStudio environment as a numeric matrix making sure there are no non-numeric values or NA within your data.

```
# Read our data into a data.frame
ONDRI_synthetic_dataset <- read.csv("baycrest/datapacks/synth_data/JUL26/
                                    OND01_ALL_01_NIBS_SYNTHDATA_2020JUL26_DATA.csv")

# Make sure there are no NAs in your data. This should be integer(0).
which(sapply(ONDRI_synthetic_dataset, is.na))

# Make sure all these values are TRUE for variables you are including
lapply(ONDRI_synthetic_dataset, is.numeric)

# ...

# Convert data frame into matrix
ONDRI_synth_num_matrix <- as.matrix(ONDRI_synthetic_dataset_prepped)
```

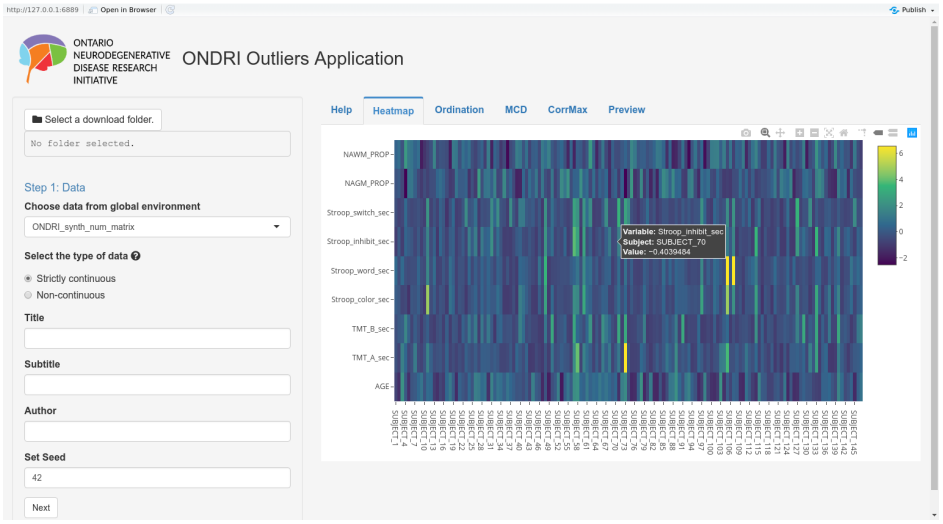
Once again, your data must be a numeric *matrix* and not a data frame after you make sure there are no NA or non-numeric values.

4.1 Step 1: Data and Document Parameters

Select your data in the Step 1 sidebar, then select *Strictly continuous* if all of your columns are continuous, otherwise select *Non-continuous*. If you are generating a report, set the title, subtitle, and author of your report and select the folder which the report files will be saved to. You may also choose a random seed for the bootstrapping procedure. Then, click the next button.

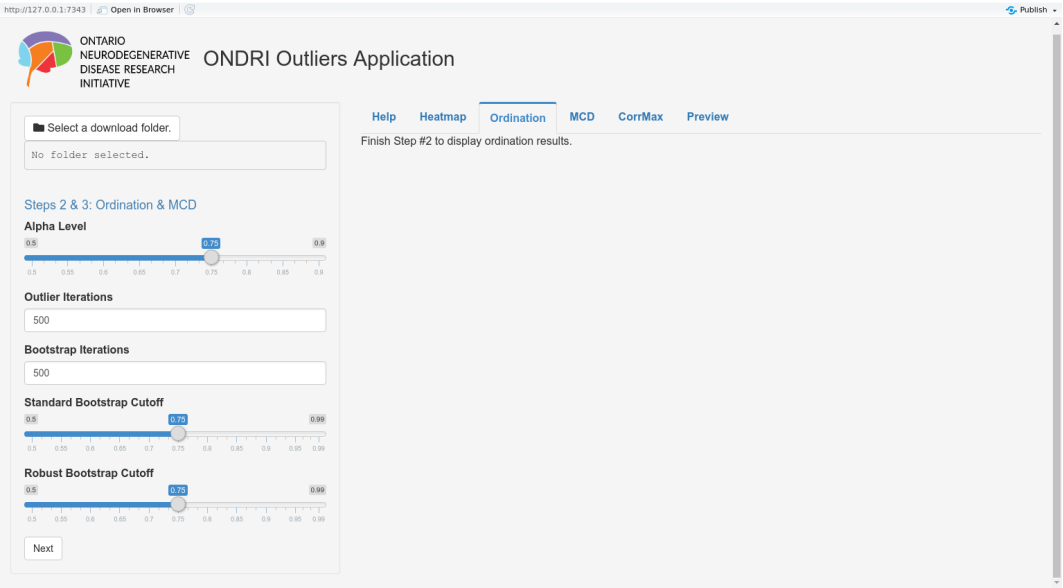


After choosing your data, you can view a scaled visualization of your data matrix in the *Heatmap* tab. You can zoom in by clicking and drawing a rectangle with the mouse to see subjects that may be omitted at a large scale. You can also double-click to zoom back out.



4.2 Step 2 & 3: Ordination & MCD

In this stage, we choose the parameters for the bootstrapping and MCD.



- *Alpha Level*: Alpha value used to compute sample size for MCD.
- *Outlier Iterations*: Number of subsets for MCD.
- *Bootstrap Iterations*: Number of samples in bootstrapping.
- *Standard Bootstrap Cutoff*: Cutoff for Mahalanobis distances in MCD plot which determines which subjects are flagged as outliers.
- *Robust Bootstrap Cutoff*: Cutoff for Robust Mahalanobis distances in MCD plot which determines which subjects are flagged as outliers.

After accepting these fields, two plots should render in the *Ordination* and *MCD* tab respectively. The left side of the Ordination plot contains the component scores of the variables and the right side contains the scores of the the subjects. The top two plots are plotted on the top two components (i.e. the ones explaining the most amount of the variance) and the bottom two plots are plotted with the last two components (i.e. the ones explaining the least amount of the variance). The MCD plot shows the standard and robust Mahalanobis distances of each subject and the bootstrap cutoffs on a log scale. After the MCD plot is rendered, the subjects on the Ordination plot will be coloured with the results of the MCD plot.

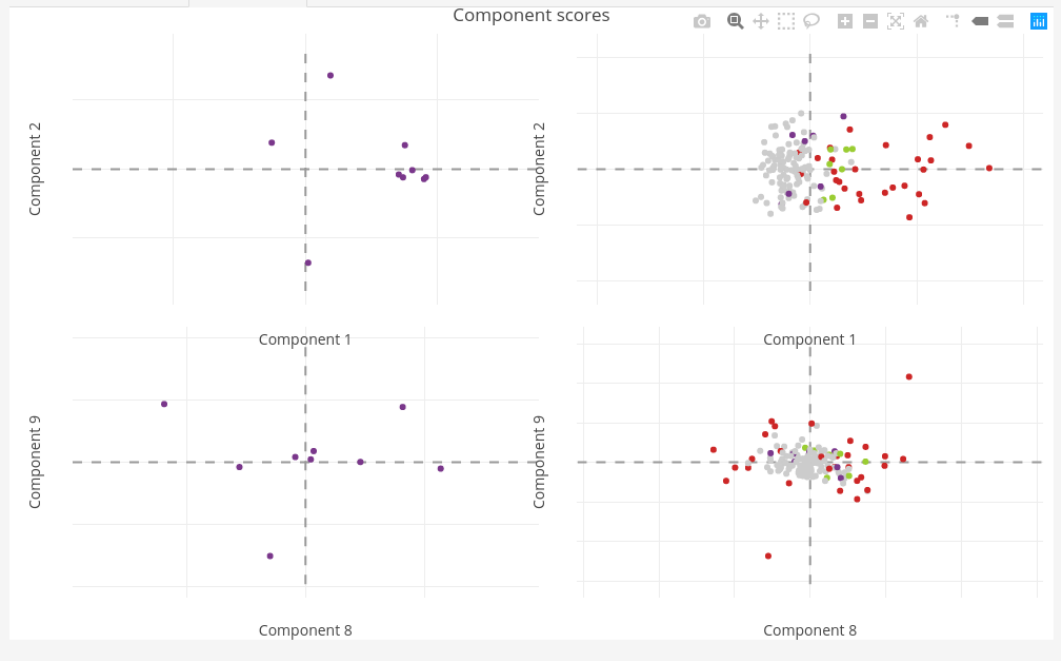


Figure 2: Plot of the columns and subjects by the first two and last two components respectively.

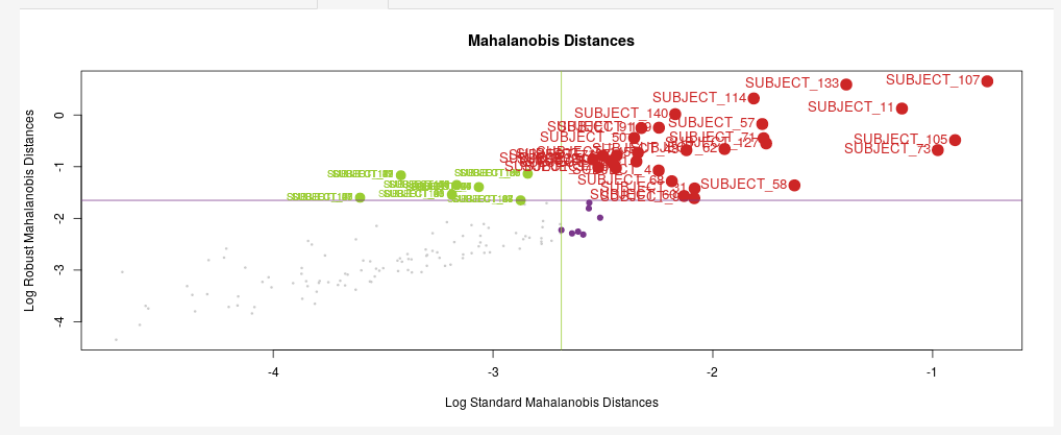
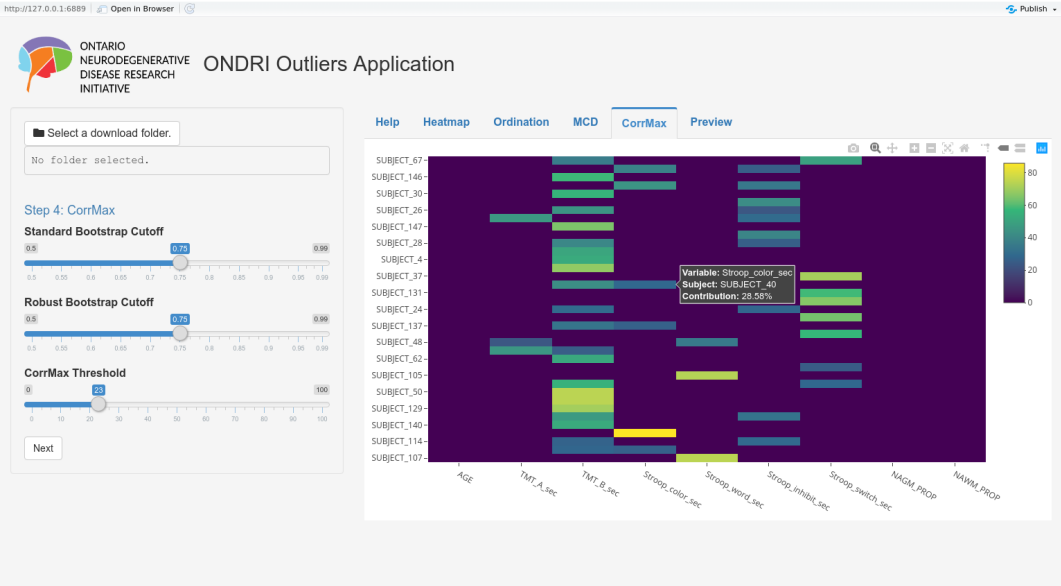


Figure 3: Plot of the Mahalanobis distances and cutoffs.

4.3 Step 4: CorrMax

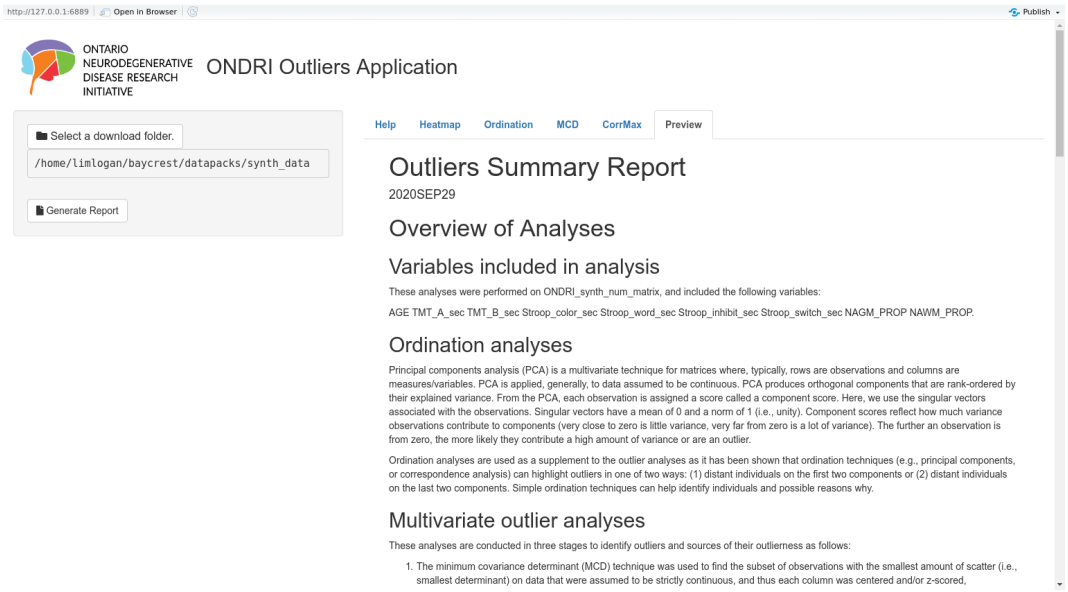
After Step 4 input is accepted, the *CorrMax* tab will show a heatmap of percentage contributions of the points flagged corresponding to the MCD plot. Like in the Heatmap tab, you can hover the cursor over the heatmap to see these values. You can also adjust the CorrMax threshold which will adjust the percentage at which a subject-variable pair will be displayed as non-zero in the CorrMax plot.



4.4 Generate Report

After the four steps are completed you can click the Generate Report button to create a PDF of the results, including the plots shown in Ordination, MCD, and CorrMax. You can preview the report at any time by visiting the *Preview* tab, but note that if a step has not yet been completed, the plots will not show in the preview.

Note if you have not yet viewed the MCD tab, the program will first take you to it when you hit the 'Generate Report' button. Then you will have to press the button again to generate the report. This is because in order to fully run the lazy-evaluated shiny processes in the MCD stage, it is required that the MCD plot be rendered by the user.



5 Common Errors

5.1 Including non-numeric values in your data or NAs

You might receive the " 'x' must be numeric" error or the 'non-numeric argument to a mathematical function' error.

Often, values that look like numbers in a data frame can be character vectors. To check that your entries are numeric, you can use

```
which(!sapply(mydata, is.numeric))
```

to see all non-numeric entries in your matrix. To find NA's, you can use

```
which(sapply(mydata, is.na))
```

5.2 Using a data.frame object instead of a matrix as data

This can cause a few problems with the internal mathematical functions that take matrices. Using a data.frame object as your data is not recommended or supported at this time.

To check if your object is a matrix or data.frame, use:

```
class(mydata)
```

To convert a data.frame to a matrix, you can use:

```
my_matrix <- as.matrix(mydata)
```

You can also view the object in RStudio to make sure your new object has the correct rows, columns and values:

```
View(my_matrix)
```

5.3 Using data that is too wide

If your data does not satisfy

```
nrow(data()) < ceiling(ncol(data())*.9)
```

then you will not be able to continue with the outlier detection process and may instead be given the 'Data is too wide' message.

6 References

Beaton, D., Sunderland, K. M.Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., Troyer, A. K.Binns, M. A., Abdi, H., & Strother, S. C. (2019). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. *bioRxiv*. https://doi.org/10.1101/333005

Garthwaite, P. H., & Koch, I. (2016). Evaluating the contributions of individual variables to a quadratic form. *Australian & New Zealand Journal of Statistics*, *58*(1), 99–119.

Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43.

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., Peltsch, A. J., Pieruccini-Faria, F., Sahlas, D. J., Swartz, R. H., & others. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: An application within the ondri project. *BMC Medical Research Methodology*, *19*(1), 1–16.