ONTARIO
NEURODEGENERATIVE
DISEASE RESEARCH
INITIATIVE

Outliers App 0.10.2.9001 - Reference Guide

On behalf of the Neuroinformatics and Biostatistics (NIBS) Platform

Updated as of March 22, 2021

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 and Deployment

1. Install R first and then RStudio. Please choose the correct installer carefully as it will depend on your computer's operating system.

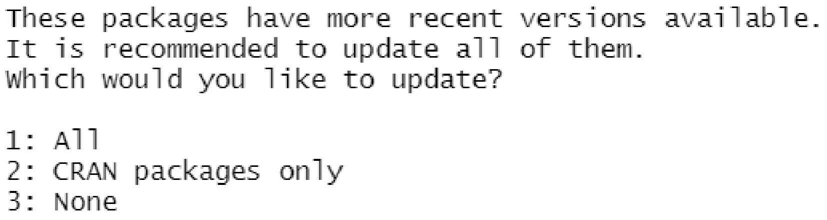
2. Install the `GSVD` and `ours` packages (which are not available through CRAN) with the following lines of code:

```
if (!require("devtools")){
  install.packages("devtools")
}

if (!require("GSVD")){
  Sys.setenv(R_REMOTES_NO_ERRORS_FROM_WARNINGS = TRUE)
  devtools::install_github("derekbeaton/GSVD")
}
if (!require("ours")){
  Sys.setenv(R_REMOTES_NO_ERRORS_FROM_WARNINGS = TRUE)
  devtools::install_github("derekbeaton/OuRS", subdir = "/OuRS")
}
```

3. Download and install the shiny app directly with the following lines of code:

```
if (!require("devtools")){
  install.packages("devtools")
}
devtools::install_github(repo = "ondri-nibs/outliers_app")
```

4. Type `ONDRIOutliersApp::installPackages()` to install any missing packages and/or dependencies. If you get the following message in your RStudio console, please type 3.



```
These packages have more recent versions available.
It is recommended to update all of them.
Which would you like to update?

1: All
2: CRAN packages only
3: None
```

Figure 1: This message may appear multiple times during the installation process.

5. When installation is complete, type `ONDRIOutliersApp::runApp()` to open the app.

3 How to Use

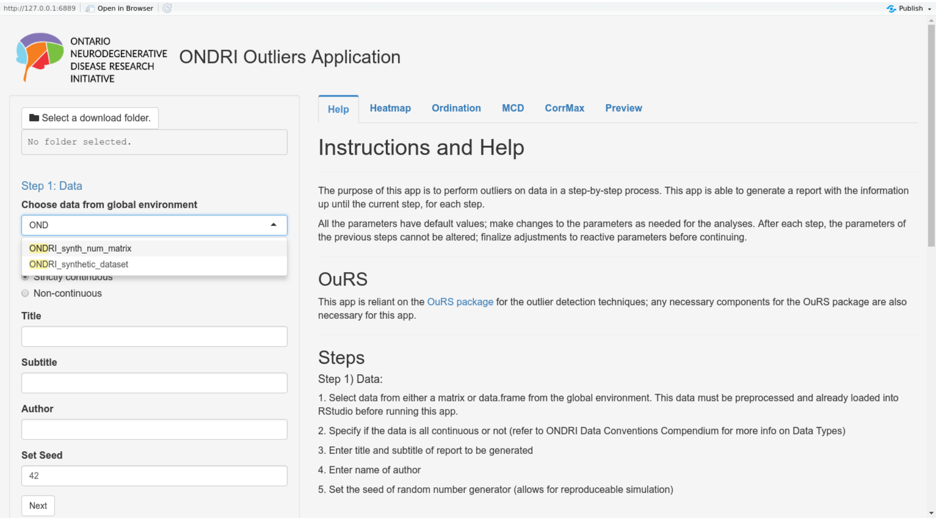
To start, load your data into your RStudio environment as a numeric matrix making sure there are no non-numeric values or NA within your data.

```r
# 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.

3.1 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.



ONDRI Outliers Application

No folder selected.

Step 1: Data

Choose data from global environment

OND

ONDRI_synth_num_matrix

ONDRI_synthetic_dataset

● Strictly continuous

○ Non-continuous

Title

Subtitle

Author

Set Seed

42

Instructions and Help

The purpose of this app is to perform outliers on data in a step-by-step process. This app is able to generate a report with the information up until the current step, for each step.

All the parameters have default values; make changes to the parameters as needed for the analyses. After each step, the parameters of the previous steps cannot be altered; finalize adjustments to reactive parameters before continuing.

OuRS

This app is reliant on the OuRS package for the outlier detection techniques; any necessary components for the OuRS package are also necessary for this app.

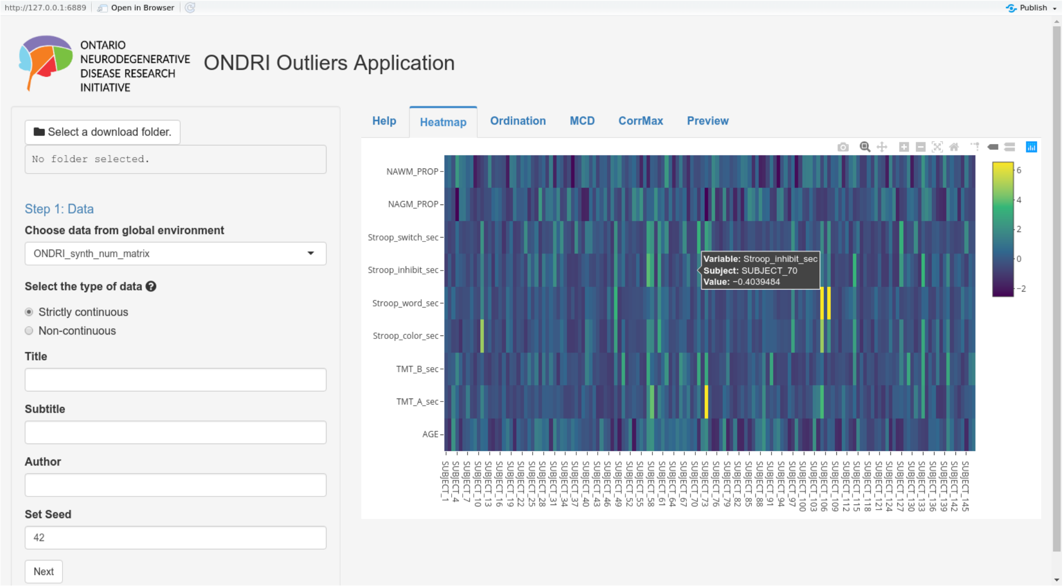
Steps

Step 1) Data:

1. Select data from either a matrix or data.frame from the global environment. This data must be preprocessed and already loaded into RStudio before running this app.
2. Specify if the data is all continuous or not (refer to ONDRI Data Conventions Compendium for more info on Data Types)
3. Enter title and subtitle of report to be generated
4. Enter name of author
5. Set the seed of random number generator (allows for reproduceable simulation)

Figure 2: All output from the app will be added to the download folder selected.

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.



ONDRI Outliers Application

No folder selected.

Step 1: Data

Choose data from global environment: ONDRI_synth_num_matrix

Select the type of data

● Strictly continuous

○ Non-continuous

Title

Subtitle

Author

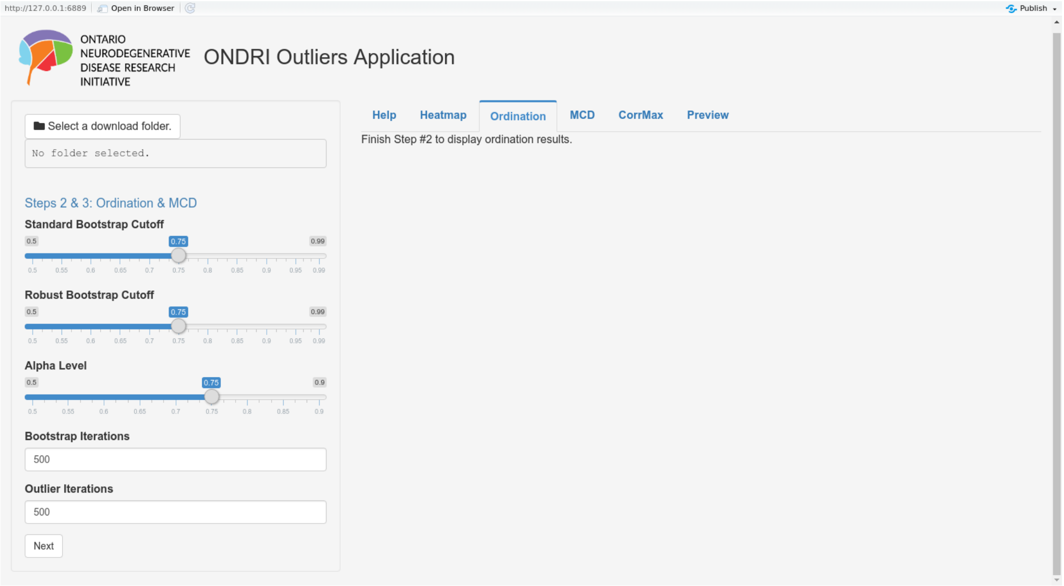
Set Seed: 42

Variable: Stroop_inhibit_sec
Subject: SUBJECT_70
Value: −0.4039484

Figure 3: Heatmap tab.

3.2 Step 2 & 3: Ordination & MCD

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



ONDRI Outliers Application

No folder selected.

Finish Step #2 to display ordination results.

Steps 2 & 3: Ordination & MCD

Standard Bootstrap Cutoff: 0.75 (range 0.5 – 0.99)

Robust Bootstrap Cutoff: 0.75 (range 0.5 – 0.99)

Alpha Level: 0.75 (range 0.5 – 0.9)

Bootstrap Iterations: 500

Outlier Iterations: 500

Figure 4: Selection of parameters prior to ordination.

- *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.

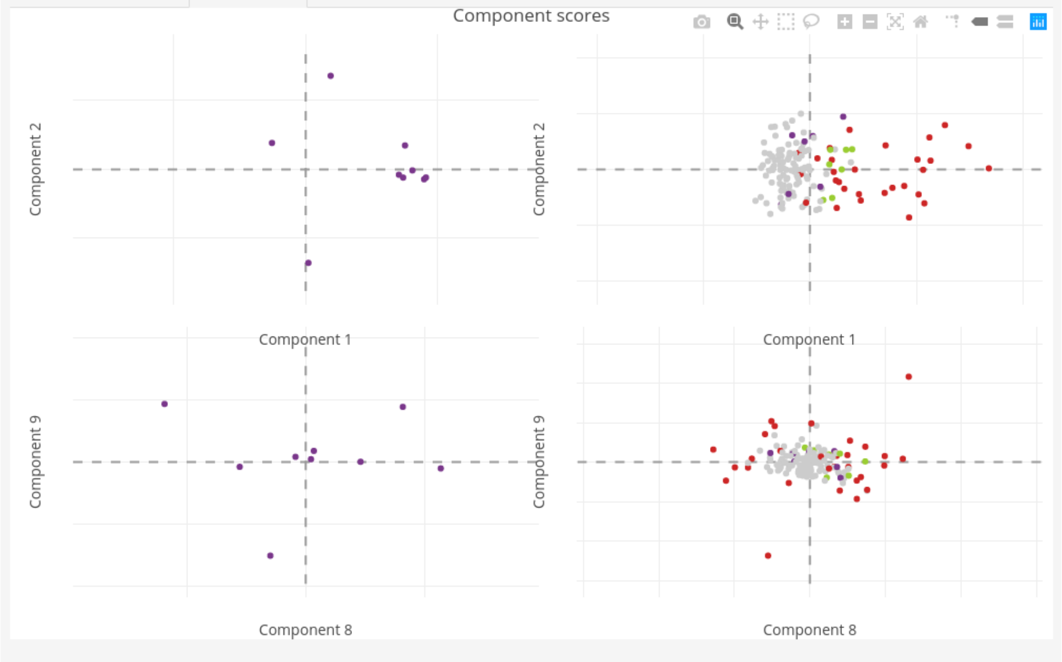


Figure 5: Plot of the columns and subjects by the first two and last two components respectively.

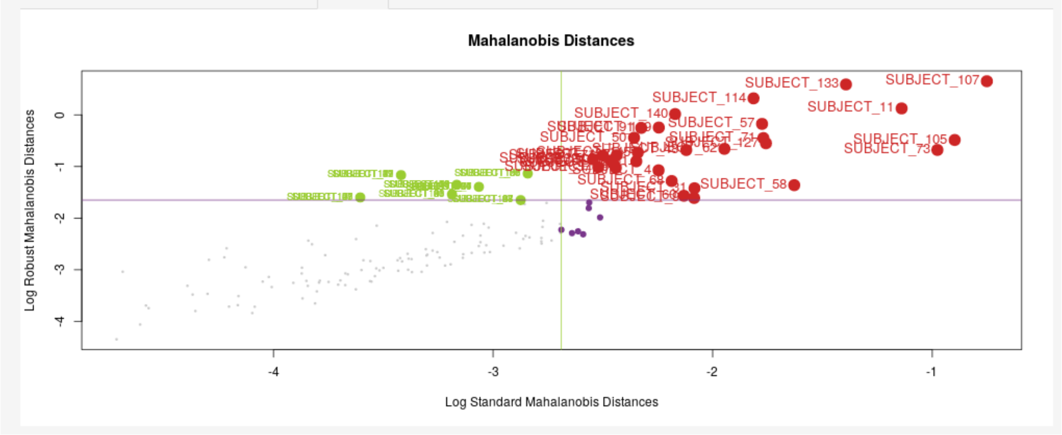
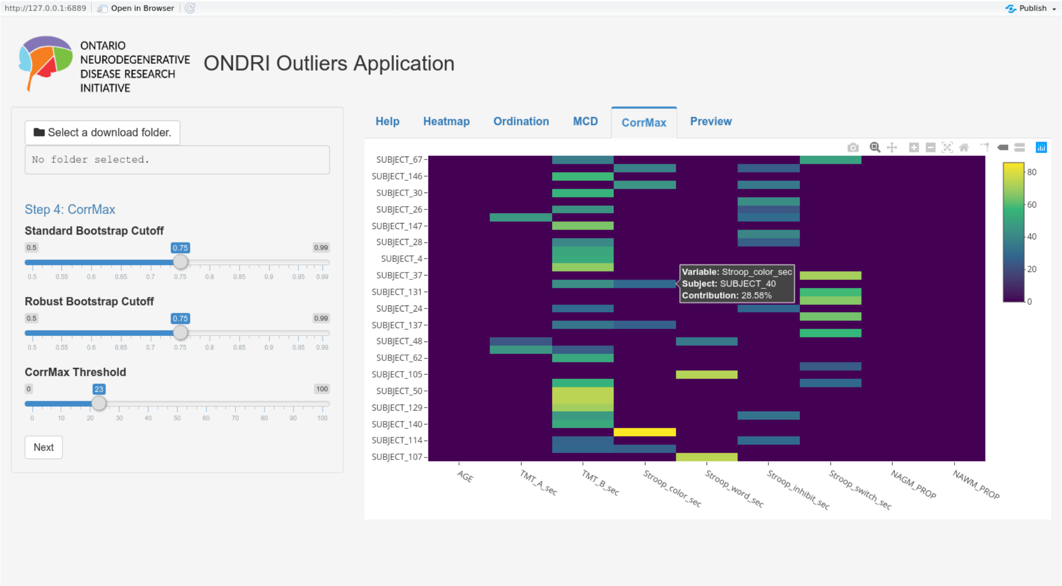


Figure 6: Plot of the Mahalanobis distances and cutoffs.

3.3 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.



```
No folder selected.
```

Step 4: CorrMax

Standard Bootstrap Cutoff: 0.5 – 0.75 – 0.99

Robust Bootstrap Cutoff: 0.5 – 0.75 – 0.99

CorrMax Threshold: 0 – 23 – 100

Variable: Stroop_color_sec
Subject: SUBJECT_40
Contribution: 28.58%

Figure 7: CorrMax tab.

3.4 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.

4 Common Errors

4.1 Including non-numeric values in your data or NAs

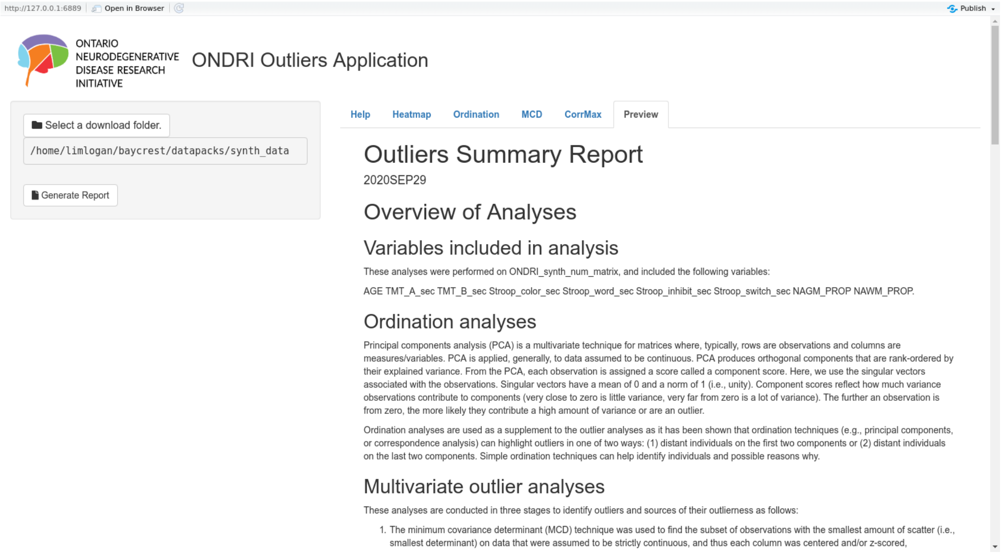
You might receive the " 'x' must be numeric" error or the 'non-numeric argument to a mathematical function' error.

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))
```



http://127.0.0.1:6889
Open in Browser
Publish
ONTARIO NEURODEGENERATIVE DISEASE RESEARCH INITIATIVE
ONDRI Outliers Application
Select a download folder.
/home/limlogan/baycrest/datapacks/synth_data
Generate Report
Help
Heatmap
Ordination
MCD
CorrMax
Preview

Outliers Summary Report

2020SEP29

Overview of Analyses

Variables included in analysis

These analyses were performed on ONDRI_synth_num_matrix, and included the following variables:

AGE TMT_A_sec TMT_B_sec Stroop_color_sec Stroop_word_sec Stroop_inhibit_sec Stroop_switch_sec NAGM_PROP NAWM_PROP.

Ordination analyses

Principal components analysis (PCA) is a multivariate technique for matrices where, typically, rows are observations and columns are measures/variables. PCA is applied, generally, to data assumed to be continuous. PCA produces orthogonal components that are rank-ordered by their explained variance. From the PCA, each observation is assigned a score called a component score. Here, we use the singular vectors associated with the observations. Singular vectors have a mean of 0 and a norm of 1 (i.e., unity). Component scores reflect how much variance observations contribute to components (very close to zero is little variance, very far from zero is a lot of variance). The further an observation is from zero, the more likely they contribute a high amount of variance or are an outlier.

Ordination analyses are used as a supplement to the outlier analyses as it has been shown that ordination techniques (e.g., principal components, or correspondence analysis) can highlight outliers in one of two ways: (1) distant individuals on the first two components or (2) distant individuals on the last two components. Simple ordination techniques can help identify individuals and possible reasons why.

Multivariate outlier analyses

These analyses are conducted in three stages to identify outliers and sources of their outlierness as follows:

1. The minimum covariance determinant (MCD) technique was used to find the subset of observations with the smallest amount of scatter (i.e., smallest determinant) on data that were assumed to be strictly continuous, and thus each column was centered and/or z-scored,

Figure 8: Report preview.

4.2 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)
```

4.3 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.

5 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.