

EnsMOD Monte Carlo (EnsMOD-MC)

EnsMOD is used for Omics Sample Outlier Detection

The user can set four threshold parameters:

Cophenetic Correlation Coefficient (CCC):

0.8

Descriptive Statistic

Silhouette Coefficient (SC):

0.25

Descriptive Statistic

ROBust PCA algorithm (rob pca) cutoff:

0.975

Statistical Test (assumes normality)

Robust PCA algorithm (PcaGrid) cutoff:

0.975

Statistical Test (assumes normality)

An overall robust probability value was not calculated for the detected outlier(s)

Monte Carlo Methods

Monte Carlo methods are a broad class of algorithms that rely on repeated random sampling to obtain a numerical result(s).

Monte Carlo methods can be used to robustly solve any problem having a probabilistic interpretation.

Monte Carlo methods tend to follow a particular pattern:

1. Define the domain of possible input values
2. Generate inputs randomly
3. Perform a deterministic computation using each input to produce a result
4. Aggregate the results

False Positive Rate (FPR)

		Predicted condition	
		Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

If one or more outliers are detected using EnsMOD, then EnsMOD-MC can be used to reanalyze the dataset to estimate the FPR of the outlier detection

EnsMOD-MC does not make any assumptions about the user's dataset (e.g., normal variance is not assumed)

EnsMOD-MC is 100% separate from EnsMOD, and it is available as an open-source, freely available, stand-alone script at <https://github.com/niaid/EnsMOD>

Use of EnsMOD-MC:

1. Delete the outliers detected by EnsMOD from the input dataset
2. Optional: Append “_fixed” to the end of column headers to prevent sampling
3. Optional: Adjust the number of MC simulations and/or the random seed
4. Use exactly the same threshold parameters used for EnsMOD

EnsMOD-MC

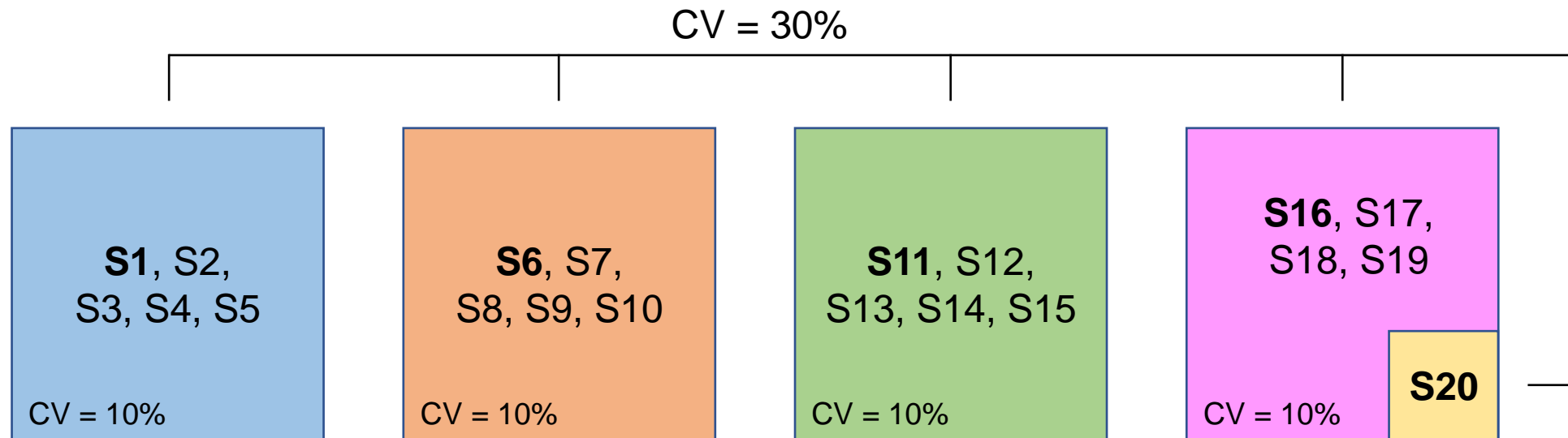
1. Randomly samples (without replacement) the non-outlier data
2. Performs EnsMOD outlier detection (detected outliers are false positives)
3. Loops to Step 1 to perform N simulations

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{\text{FP}}{\text{Loops}}$$

Note that $P = \text{TP} + \text{FN} = 0$

Simulated Proteomics Dataset

S1, S6, S11, S16, S20 = Gaussian random sampling
 $\mu = 1,000,000$, CV = 30%, N=100 proteins
Remaining Samples: CV = 10%



EnsMOD-MC Example: Simulated Proteomics Dataset

Original simulated proteomics dataset used for EnsMOD (s_20 is the simulated outlier).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	s_01	s_02	s_03	s_04	s_05	s_06	s_07	s_08	s_09	s_10	s_11	s_12	s_13	s_14	s_15	s_16	s_17	s_18	s_19	s_20	
2	1.74E+06	1.61E+06	1.76E+06	2.10E+06	1.75E+06	1.28E+06	1.48E+06	1.44E+06	1.21E+06	1.31E+06	1.17E+06	9.34E+05	1.08E+06	1.17E+06	1.04E+06	1.20E+06	1.36E+06	1.16E+06	1.27E+06	8.69E+05	
3	1.14E+06	1.26E+06	1.21E+06	1.38E+06	9.95E+05	7.07E+05	7.26E+05	6.21E+05	8.27E+05	7.02E+05	1.45E+06	1.43E+06	1.40E+06	1.29E+06	1.37E+06	1.28E+06	1.22E+06	1.35E+06	1.18E+06	1.31E+06	
4	9.53E+05	1.06E+06	9.32E+05	8.94E+05	8.20E+05	1.23E+06	1.30E+06	1.34E+06	1.19E+06	1.13E+06	8.15E+05	8.49E+05	8.83E+05	9.52E+05	8.95E+05	1.13E+06	1.08E+06	1.03E+06	1.19E+06	7.49E+05	
5	8.94E+05	8.74E+05	8.89E+05	7.87E+05	9.52E+05	1.34E+06	1.45E+06	1.32E+06	1.61E+06	1.46E+06	9.41E+05	9.50E+05	1.00E+06	1.09E+06	8.97E+05	1.04E+06	9.87E+05	9.40E+05	1.03E+06	1.37E+06	
6	1.32E+06	1.18E+06	1.54E+06	1.30E+06	1.18E+06	1.17E+06	1.24E+06	1.20E+06	9.94E+05	1.24E+06	1.23E+06	1.46E+06	1.42E+06	1.20E+06	1.02E+06	1.22E+06	1.34E+06	9.84E+05	1.10E+06	1.11E+06	

Simulated proteomics dataset used for EnsMOD_MC. Sample s_20 was deleted, and only the fourth experimental condition (it was s_16, s_17, s_18, s_19, s_20) was shuffled.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	s_01_fixe	s_02_fixe	s_03_fixe	s_04_fixe	s_05_fixe	s_06_fixe	s_07_fixe	s_08_fixe	s_09_fixe	s_10_fixe	s_11_fixe	s_12_fixe	s_13_fixe	s_14_fixe	s_15_fixe	s_16	s_17	s_18	s_19	
2	1.74E+06	1.61E+06	1.76E+06	2.10E+06	1.75E+06	1.28E+06	1.48E+06	1.44E+06	1.21E+06	1.31E+06	1.17E+06	9.34E+05	1.08E+06	1.17E+06	1.04E+06	1.20E+06	1.36E+06	1.16E+06	1.27E+06	
3	1.14E+06	1.26E+06	1.21E+06	1.38E+06	9.95E+05	7.07E+05	7.26E+05	6.21E+05	8.27E+05	7.02E+05	1.45E+06	1.43E+06	1.40E+06	1.29E+06	1.37E+06	1.28E+06	1.22E+06	1.35E+06	1.18E+06	
4	9.53E+05	1.06E+06	9.32E+05	8.94E+05	8.20E+05	1.23E+06	1.30E+06	1.34E+06	1.19E+06	1.13E+06	8.15E+05	8.49E+05	8.83E+05	9.52E+05	8.95E+05	1.13E+06	1.08E+06	1.03E+06	1.19E+06	
5	8.94E+05	8.74E+05	8.89E+05	7.87E+05	9.52E+05	1.34E+06	1.45E+06	1.32E+06	1.61E+06	1.46E+06	9.41E+05	9.50E+05	1.00E+06	1.09E+06	8.97E+05	1.04E+06	9.87E+05	9.40E+05	1.03E+06	
6	1.32E+06	1.18E+06	1.54E+06	1.30E+06	1.18E+06	1.17E+06	1.24E+06	1.20E+06	9.94E+05	1.24E+06	1.23E+06	1.46E+06	1.42E+06	1.20E+06	1.02E+06	1.22E+06	1.34E+06	9.84E+05	1.10E+06	

EnsMOD-MC Example: Simulated Proteomics Dataset --> Estimated FPR = 0%

```
# Set the number of Monte Carlo tests
# Note that ~1000 or more tests are recommended
num_MC_Tests <- 1000

# Set the initial seed for the Monte Carlo tests
MC_seed <- 90421

# Set the Cophenetic Correlation Coefficient
# The CCC is a measure of agreement between the original
# and the resampled dendrograms
CCC_min <- 0.8

# Set the Silhouette Coefficient
# The SC is a measure of how well the objects in each
# cluster are separated from the other clusters
SC_max <- 0.25

# Set the Robust PCA algorithm
# For normally distributed data
robtpca_prob <- 0.999

# Set the Robust PCA algorithm
# For non-normally distributed data
PcaGrid_prob <- 0.999
```



```
## [1] 100 19

# Remove rows (genes/proteins) having all zeros
input_data_0 <- input_data[!apply(input_data, 1, function(x) all(x == 0)), ]
# Remove any row (genes/proteins) with one or more invalid values ("NaN")
input_data_nna <- input_data_0[complete.cases(input_data_0), ]
# Display how many rows remain
input_data_nna_rows <- dim(input_data_nna)[1]
input_data_col <- dim(input_data_nna)[2]
dim(input_data_nna)

## [1] 100 19
```

2. Monte Carlo method applied to EnsMOD

Use the Monte Carlo method to estimate the outlier detection false positive rate (FPR) (i.e., the probability of falsely rejecting the null hypothesis). For each analyte (e.g., gene), some or all of the input data are shuffled. By default, all columns are shuffled. To prevent a column from being shuffled, append "_fixed" to the end of the sample unique identifier. The HCA and rPCA tests are performed, and false positive outliers are detected, if any. This is repeated, and the estimated FPR is the number of detected outliers divided by the number of tests.

Monte Carlo method reference: William L. Dunn and J. Kenneth Shultis, Exploring Monte Carlo Methods 2nd Ed., Elsevier Publishing, 2022

3. Monte Carlo test results

The number of Monte Carlo tests, and the total number of outliers

```
# The number of Monte Carlo tests:
num_MC_Tests

## [1] 1000

# The total number of outliers:
num_MC_Outliers

## [1] 0

# The estimated outlier detection false positive rate (i.e., the probability of falsely rejecting the null hypothesis):
MC_est_FPR <- num_MC_Outliers / num_MC_Tests
MC_est_FPR

## [1] 0
```

Spleen Phosphoproteomics of Mice with Anthrax

Prepared: 5 mice 15 mice 15 mice 25 mice

Toxin-, capsule-, asymptomatic, abortive infection Toxin+, capsule-, lethal in 2-4d

Time	No Injection	Vehicle	Δ Sterne	Sterne
0 h	5 mice			
24 h		5 mice	5 mice	5 mice
48 h		5 mice	5 mice	5 mice
72 h		5 mice	5 mice	1 mouse

The Sterne 72h sample functions as a pseudo-outlier.

Spleens, TiO₂ phospho-enrichment, Label-Free, LTQ-Orbitrap Classic

EnsMOD-MC Example: Anthrax Phosphoproteomics --> Estimated FPR = 1.4%

```
# Set the number of Monte Carlo tests
# Note that ~1000 or more might take more time
num_MC_Tests <- 500

# Set the initial seed for reproducibility
MC_seed <- 90421

# Set the Cophenetic Correlation Coefficient (CCC)
# The CCC is a measure of how faithful the model is
# between the original unmodeled data and the model
CCC_min <- 0.8

# Set the Silhouette Coefficient (SC)
# The SC is a measure of how similar an object is to its own cluster
SC_max <- 0.25

# Set the Robust PCA algorithm (robust PCA)
# For normally distributed data, this is the best choice
robust_pca_prob <- 0.999

# Set the Robust PCA algorithm (PcaGrid)
# For normally distributed data, this is the best choice
PcaGrid_prob <- 0.999
```

EnsMOD_Monte_Carlo_v1_0

File | C:/Users/nmane/OneDrive/Desktop/Omics%20outlier%20detection%202021/EnsMOD_MC/Test02%20-%20v1_0%20rPCA0.999%20-%20Anthrax%20Phosphoproteomics%20...

```
## [1] 5205 44

# Remove rows (genes/proteins) having all zeros
input_data_0 <- input_data[!apply(input_data, 1, function(x) all(x == 0)), ]
# Remove any row (genes/proteins) with one or more invalid values ("NaN")
input_data_nna <- input_data_0[complete.cases(input_data_0), ]
# Display how many rows remain
input_data_nna_rows <- dim(input_data_nna)[1]
input_data_col <- dim(input_data_nna)[2]
dim(input_data_nna)

## [1] 1714 44
```

2. Monte Carlo method applied to EnsMOD

Use the Monte Carlo method to estimate the outlier detection false positive rate (FPR) (i.e., the probability of falsely rejecting the null hypothesis). For each analyte (e.g., gene), some or all of the input data are shuffled. By default, all columns are shuffled. To prevent a column from being shuffled, append "_fixed" to the end of the sample unique identifier. The HCA and rPCA tests are performed, and false positive outliers are detected, if any. This is repeated, and the estimated FPR is the number of detected outliers divided by the number of tests.

Monte Carlo method reference: William L. Dunn and J. Kenneth Shults, Exploring Monte Carlo Methods 2nd Ed., Elsevier Publishing, 2022

3. Monte Carlo test results

The number of Monte Carlo tests, and the total number of outliers

```
# The number of Monte Carlo tests:
num_MC_Tests

## [1] 500

# The total number of outliers:
num_MC_Outliers

## [1] 7

# The estimated outlier detection false positive rate (i.e., the probability of falsely rejecting the null hypothesis):
MC_est_FPR <- num_MC_Outliers / num_MC_Tests
MC_est_FPR

## [1] 0.014
```

