EnsMOD Monte Carlo (EnsMOD_MC) User Guide

Nathan P. Manes, Jian Song, Aleksandra Nita-Lazar

Laboratory of Immune System Biology, National Institute of Allergy
and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA

Contact us at EnsMOD-team@list.nih.gov

Introduction

If one or more outliers are detected using EnsMOD, EnsMOD_MC can be used to reanalyze the user's dataset to estimate the outlier detection false positive rate (FPR, which equals the probability of falsely detecting an outlier). EnsMOD_MC uses the Monte Carlo method, and a good introduction to the Monte Carlo method is William L. Dunn and J. Kenneth Shultis, Exploring Monte Carlo Methods 2nd Ed., Elsevier Publishing, 2022.

EnsMOD_MC does not make any assumptions about the user's dataset (e.g., approximately Gaussian variance is not presumed). EnsMOD_MC is available as a stand-alone Rmarkdown script. EnsMOD and EnsMOD_MC are open-source and freely available (https://github.com/niaid/EnsMOD).

EnsMOD_MC Overview

First, delete the detected outlier(s) from the input dataset and save it as "Gene_Expression_Table.xlsx". The resulting number of columns needs to be ≥9. Run EnsMOD_MC. Use the same four cutoff parameter values (CCC, SC, robpca, PcaGrid) that were used to detect the outlier(s). One row (e.g., gene) at a time, some or all of the input data are shuffled. The EnsMOD HCA and rPCA tests are performed, and false positive outliers are detected, if any. This is repeated a preset number of times. The estimated FPR is the number of detected outliers (i.e., false positives) divided by the number of simulations.

By default, all columns are included during shuffling. To prevent a column from being shuffled, append "_fixed" to the end of the column header (i.e., the sample unique identifier) of the input dataset. The experimental conditions that did not correspond to an outlier do not need to be shuffled. If multiple experimental conditions correspond to detected outliers, they can be analyzed separately using EnsMOD_MC.

Installing and Running EnsMOD_MC

1. Install R (https://www.r-project.org/).
2. Install RStudio (https://www.rstudio.com/).
3. Acquire "EnsMOD_Monte_Carlo_stand-alone_v1_0.Rmd" (https://github.com/niaid/EnsMOD).
4. The script automatically installs and updates all of the required R packages.
5. **Provide the table of input data.**
    a. The input table needs to be in "Gene_Expression_Table.xlsx", and it needs to be located in the same directory as the EnsMOD_MC Rmarkdown file.
    b. The XLSX file should contain only one worksheet and only one table.
        i. The first row (the header) should contain unique identifiers of the samples.
        ii. The detected outlier columns should be deleted from the table.
        iii. To prevent a column from being shuffled, append "_fixed" to the end of the column header (i.e., the sample unique identifier).
    c. Rows may contain missing values (such as "NaN"), but these rows will be excluded from the analyses (note that "0" is not treated as a missing value).
    d. EnsMOD example datasets are provided at (https://github.com/niaid/EnsMOD/tree/main/app/EnsMOD_Examples)
    e. During our testing of EnsMOD, a minimum of nine samples (i.e., columns) were required. Fewer samples resulted in the robpca step failing ("Error in robpca: Something went wrong with the outlyingness computations.").

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	

6. **Ensure that the EnsMOD_MC settings are correct.**
    a. Use RStudio to edit the parameter values and click Save.
    b. EnsMOD and EnsMOD_MC should use the same four outlier detection cutoff parameter values. They are:
        i. The minimum CCC threshold
        ii. The maximum SC threshold
        iii. The robpca probabilistic threshold
        iv. The PcaGrid probabilistic threshold
    c. The number of Monte Carlo simulations (num_MC_Tests) should be at least ~500, but note that the runtime could be many hours or even days (try 10 to estimate the runtime for 500 simulations).
7. Run EnsMOD_MC.
    a. To run EnsMOD_MC, click the "Knit" button. EnsMOD_MC will produce an HTML output file that can be reviewed using a web browser.