Dirichlet-tree multinomial mixtures for clustering microbiome compositions

Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

- [ ] This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).
- [x] I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The case studies in this paper is implemented with the July 29, 2016 version of the fecal data from the American Gut Project. The American Gut Project collects fecal, oral, skin, and other body site samples from thousands of participants. Lifestyle and dietary data about each participant is also collected, making it possible to study the associations between the human microbiome and various factors. The data is freely available to public through the link ftp://ftp.microbio.me/AmericanGut. The original American Gut sequences and metadata are available at The European Bioinformatics Institute under the accession ERP012803. Cleaned data including the OTU table, the phylogenetic tree and the covariate information about the participants are available at the data portal ftp://ftp.microbio.me/AmericanGut. Licensing information of the data can be found at https://github.com/biocore/American-Gut/blob/master/LICENSE.

In the simulation studies, we reanalyze the dataset in *Dethlefsen, Les, and David A. Relman. "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation." Proceedings of the National Academy of Sciences 108.Supplement 1 (2011): 4554-4561.* The dataset is available at https://www.pnas.org/content/108/Supplement_1/4554.

Availability

- [x] Data **are** publicly available.
- [ ] Data **cannot be made** publicly available.

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

- [ ] Data are available online at:
- [x] Data are available as part of the paper's supplementary material.
- [ ] Data are publicly available by request, following the process described here:
- [ ] Data are or will be made available through some other mechanism, described here:

Description

File format(s)

- [ ] CSV or other plain text.
- [x] Software-specific binary format (.Rda, Python pickle, etc.): pkcle
- [ ] Standardized binary format (e.g., netCDF, HDF5, etc.):
- [ ] Other (please specify):

Data dictionary

- [x] Provided by authors in the following file(s): please refer to the README files in the data folder
- [ ] Data file(s) is(are) self-describing (e.g., netCDF files)
- [ ] Available at the following URL:

Part 2: Code

Abstract

The code includes all functions to generate the simulated datasets in Section 3 as well as all functions to perform the analyses in Section 3 and Section 4. Specifically, the folder “numerical_examples” contains all functions to reproduce the numerical studies in Section 3. The folder “case_studies” contains all functions to reproduce the case studies in Section 4.

Description

Code format(s)

- [x] Script files
  - [x] R
  - [ ] Python
  - [ ] Matlab
  - [ ] Other:
- [x] Package
  - [x] R
  - [ ] Python
  - [ ] MATLAB toolbox
  - [ ] Other:
- [ ] Reproducible report
  - [ ] R Markdown
  - [ ] Jupyter notebook
  - [ ] Other:
- [ ] Shell script
- [ ] Other (please specify):

Supporting software requirements

Version of primary software used

R version 3.6.0

Libraries and dependencies used by the code

R packages to run the numerical and case study analyses:

- ape (suggested version 5.4)
- phyloseq (suggested version 1.32.0)
- DirichletMultinomial (suggested version 1.30.0)
- clusteval (suggested version 0.1)
- cluster (suggested version 2.1.0)
- kernlab (suggested version 0.9-29)
- vegan (suggested version 2.5-6)

Additional R packages to reproduce the figures:

- ggplot2 (suggested version 3.3.2)
- reshape2 (suggested version 1.4.4)
- viridis (suggested version 0.5.1)

Parallelization used

- [x] No parallel code used
- [ ] Multi-core parallelization on a single machine/node
  - Number of cores used:
- [ ] Multi-machine/multi-node parallelization
  - Number of nodes and cores used: 3 nodes, 243 cores

License

- [x] MIT License (default)
- [ ] BSD
- [ ] GPL v3.0
- [ ] Creative Commons
- [ ] Other: (please specify below)

Scope

The provided workflow reproduces:

- [ ] Any numbers provided in text in the paper
- [x] All tables and figures in the paper
- [ ] Selected tables and figures in the paper, as explained and justified below:

Workflow

Format(s)

- [ ] Single master code file
- [ ] Wrapper (shell) script(s)
- [ ] Self-contained R Markdown file, Jupyter notebook, or other literate programming approach
- [ ] Text file (e.g., a readme-style file) that documents workflow
- [ ] Makefile
- [x] Other (more detail in *Instructions* below)

Instructions

1. Simulation studies

The simulation studies in Section 3.1 can be reproduced with files in the folder "numerical_examples/simulations". Functions to generate the simulated datasets are provided in the folder "dgp". Each run of the file in the folder generate a dataset for a single round of the simulation study. For example, the folder "results" contains the simulation results under scenario 2 with 90 samples. The functions in folder "analyses" read the simulation results and produce the numebers to generate the tables and figures in Section 3.1. Here we only include results for a specific simulation scenario in the folder "results" for illustration. Results for all other scenarios can be reproduced and saved in the same manner.

The workflow to reproduce the simulation studies is described as follows:

- Generate data and run the competing methods. Run each R file in the folder "study" 100 times with different random seed. Save the results in the folder "results".
- Run the R files in the folder "analyses" to reproduce the simulation results (figures and numbers for the tables).

2. Validation

The folder "validation" contains code to reproduce the validation study in Section 3.2. Files "D.R", "E.R", and "F.R" contain code to analyze the datasets of subject D, E, and F with DTMM, respectively. Results of directly running these files are provided as datasets "D_5.RData", "E_5.RData", and "F_5.RData". An entire run of the file "validation_analysis.R" provides all figures in the validation study.

3. Case studies

The folder "case_studies" contains code to reproduce the case studies in Section 4. The subfolder "IBD" contains the case study on the IBD dataset in Section 4.1. Specifically, the file "AG_ibd_job.R" contains the R code to fit the IBD data with DTMM, the results are saved in the dataset "res_ag_ibd.RData" (this dataset is obtained by an entire run of the file "AG_ibd_job.R"). "AG_ibd_analysis.R" contains all functions to generate the figures in Section 4.1. Similarly, all code and functions to generate results in Section 4.2 are provided in folder "Diabetes".

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

- [ ] < 1 minute
- [ ] 1-10 minutes
- [ ] 10-60 minutes
- [ ] 1-8 hours
- [ ] > 8 hours
- [x] Not feasible to run on a desktop machine, as described here: Each simulation round in Section 3.1 takes 1-8 hours to run. However, it is not feasible to run the entire simulation study on a desktop machine. The validation study in Section 3.2 on each patient takes 5-8 hours to run on a desktop with a 2.5 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory. The case studies take more than 24 hours to run on the same desktop.