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).

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

This paper demonstrates the merits of the proposed methods by leveraging the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and the ADHD-200 Preprocessed dataset. Both of these datasets are publicly available and can be used to enhance the understanding of the human brain itself and its relationship with common factors such as gender, age, etc.

Availability

☒ 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:

☐ 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:

- **The ADNI dataset**. Access of the ADNI dataset is contingent on adherence to the ADNI Data Use Agreement and the publications' policies outlined in the ADNI website (for example, Data Sharing and Publication Policy). As a result, we do not have the ability to publicly deploy this dataset for other individuals or research groups. You may first need to obtain access from the ADNI team, and then kindly request the preprocessed dataset from Prof. Hongtu Zhu.
- **The ADHD-200 processed dataset**. Data usage is unrestricted for non-commercial research purposes. Therefore, we cannot arbitrarily publish this dataset, as it is intended solely for non-commercial research. In turn, the ADHD-200 consortium provides instruction that detail the procedure for obtaining the dataset. Simply follow these instructions to access the dataset mentioned in the paper.

Non-publicly available data

Description

File format(s)

☒ CSV or other plain text.
☐ Software-specific binary format (.Rda, Python pickle, etc.): pkcle
☐ Standardized binary format (e.g., netCDF, HDF5, etc.):
☒ Other (please specify): `.dat` and `.tsv` files

Data dictionary

- ☐ Provided by authors in the following file(s):
- ☒ Data file(s) is(are) self-describing (e.g., netCDF files)
- ☐ Available at the following URL:

Additional Information (optional)

Part 2: Code

Abstract

Our code includes the R package that implements the Cramér-von Mises test and the metric association test. The p-values for these tests can be derived either through permutation or by estimating the asymptotic distribution under null hypotheses.

Description

Code format(s)

- ☒ Script files
  - ☒ R
  - ☐ Python
  - ☐ Matlab
  - ☐ Other:
- ☒ Package
  - ☒ 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 3.5.3
- Python: version 3.9.12

Libraries and dependencies used by the code R packages:

- `Ball`: 1.3.13 (Please download this package from the github repository. Some advanced features haven't been integrated into the Ball package (version 1.3.13) in R CRAN. We expect to make the Ball package on R CRAN support these new features in version 1.3.14.)

The following R packages are available on R CRAN:

- `movMF`: 0.2.4
- `CovTools`: 0.5.3
- `snowfall`: 1.84.6.1
- `mvtnorm`: 1.0.10
- `energy`: 1.7.5

- `multivariance`: 2.4.1
- `abind`: 1.4.5
- `fda.usc`: 2.1.0
- `fda`: 6.1.4
- `lokern`: 1.1.10
- `fdapace`: 0.5.9
- `CVglasso`: 1.0
- `reshape`: 1.4.4
- `ggplot2`: 3.4.2

The following Python packages can be download from PyPI:

- `geomstats`: 2.4.2
- `numpy`: 1.22.3
- `pyreadr`: 0.4.4

Supporting system/hardware requirements (optional)

Parallelization used

☐ No parallel code used
☒ Multi-core parallelization on a single machine/node
  – Number of cores used: 50
☐ Multi-machine/multi-node parallelization
  – Number of nodes and cores used:

License

☐ MIT License (default)
☐ BSD
☒ GPL v3.0
☐ Creative Commons
☐ Other: (please specify)

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

☐ Any numbers provided in text in the paper
☐ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))
☐ All tables and figures in the paper
☒ Selected tables and figures in the paper, as explained and justified below:
  – Figure 3
  – Figure 4
  – Table 1
  – Table 2

Workflow

Location

The workflow is available:

☐ As part of the paper's supplementary material.
☒ In this Git repository: a public repository Nonparametric-Statistical-Inference-via-Metric-Distribution-Function-in-Metric-Spaces
☐ Other (please specify):

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
☐ Other (more detail in *Instructions* below)

Instructions

We have organized our scripts to improve the reproducibility of experiments. Specifically, each R script file corresponds to a certain part of the results in the paper, which are listed below:

- `large_homogeneity_n.R` <-> Results in Figure 3A
- `large_joint_independence_n.R` <-> Results in Figure 3B
- `compare_homo_methods.R` <-> Results in Figure 4A
- `compare_indep_methods.R` <-> Results in Figure 4B
- `real_data_adni_analysis.R` <-> Results in Table 1
- `real_data_adhd200_analysis.R` <-> Results in Table 2

Notice that, before conducting these scripts, please modify the your_path in each script accordingly.

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
☐ Not feasible to run on a desktop machine, as described here:

Additional information (optional)

Notes (optional)