Internship project

Subject

Large-scale data depth for brain imaging

Possibility to continue as a PhD candidate

YES (Funding to be confirmed, topic will be enlarged)

About the grant "Large-scale data depth: computation and applications" (LS-Depth-CaP)

The internship will be funded by the grant of the French National Agency for Research in the category Artificial Intelligence (ANR JCJC 2021) : Today, data ordering techniques, together with most of machine learning methodology, are faced with *challenges of large-scale data* often *impaired by diversity of sources and curse of high dimensionality* on scales unimaginable until recently, with hyperspectral imagery, financial high-frequency series and social networks being only a few examples. Distinguishable from existing methods, *data depth* defines *centrality-based order*, and therefore extends median, quantiles, ranks, and outliers to *higher dimensions*, and eventually to *more complex data*. Although data depth has become increasingly important in applications, practitioners are limited by *computational burden of algorithms* and *shortage of implementations*, which impedes its usage in machine learning. Large-scale real-data applications of data depth are underdeveloped. To bring data depth to the computational level that allows its wide usage by the machine learning community, the *LS-Depth-CaP* project addresses this issue in a systematic way, pursuing four following objectives:

- Establishing a connection between statistical and computational properties of the data depth.
- Development of large-scale optimization techniques and obtaining statistical guarantees.
- Implementation as a freely-accessible software.
- Illustration on exemplifying applications involving large-scale data depth computation.

Description of the internship

Supervision

Pavlo Mozharovskyi (https://perso.telecom-paristech.fr/mozharovskyi/)

Location and dates of the internship

Address : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date of the beginning of the internship : beginning 2024

Team where the thesis will be written

Department IDS, Team Signal, Statistique et Apprentissage (S2A)

Keywords

Data depth, computational statistics, brain imaging, robustness, large-scale setting, non-parametric statistics, computational complexity, approximate computation, DT-MRI fibers.

Detailed subject

Current internship has as its main objective application of the data depth methodology to statistical treatment of fMRI brain images. Large-scale solution for disease cause investigation in the area of *brain imaging* based on DTI scans of the Old Australian Twins Study (OATS) which includes the tasks

of curve registration, outlier detection, and statistical comparison of monozygotic and dizygotic twins, for high-dimensional imaging on (potentially) thousands of individuals is envisaged.

Axonal fibers of the brain are retrieved using fMRI-processing software tracing them based on preliminary information, *e.g.*, about their origin and destination. Thus, abnormal paths can be included, and shall be identified as outlying not to perturb subsequent analysis. Further, since patients can position differently in the scanner, brain fibers should be aligned, *e.g.*, by using *the deepest observation*; *invariance properties of depth* are thus crucial here. These preliminary transformations constitute necessary pre-treatment in the study of heredity, which shall be conducted on *hundreds to thousands of aged twins* (for each measured thousands of brain fibers) originating from the OATS data set. With both monozygotic (100% of shared genes) and dizygotic (50% of shared genes) twins included in the OATS, the difference in the development of brain fibers can give insight about origination and heredity of brain ageing diseases (*e.g.*, Alzheimer or Parkinson). These questions shall be thoroughly analyzed by means of statistical testing, which will *require large-scale computation of data depth*.

Collaboration with University of New South Wales, Sydney is envisaged for (already collected) data acquisition, pre-processing, area-specific and statistical expertise.

Candidate profile

Student having or approaching master 2 level with following competences:

- Statistical learning, bases of probability

- Good level of programming (R, C/C++, Python)

- Good command of English

Application

To send on pavlo.mozharovskyi@telecom-paris.fr:

- Curriculum Vitae

- Personalized motivation letter that explains interest of the candidate in the subject (can be directly in the body of the email)

- Grade reports for recent years

- Contact of a person willing to give recommendation

Incomplete applications will not be considered.

References

[1] Dyckerhoff, R., Mozharovskyi, P., and Nagy, S. (2021): Approximate computation of projection depths. *Computational Statistics and Data Analysis*, 157, 107166.

[2] Lafaye De Micheaux, P., Mozharovskyi, P., and Vimond, M. (2021): Depth for curve data and applications. *Journal of the American Statistical Association*, 116(536), 1881–1897.

[3] Mosler K. and Mozharovskyi P. (2022): Choosing among notions of multivariate depth statistics. *Statistical Science*, 37(3), 348–368.

[4] Wen, W., Thalamuthu, A., Mather, K.A., Zhu, W., Jiang, J., Lafaye de Micheaux, P., Wright, M.J., Ames, D., Sachdev, P.S. (2016): Distinct genetic influences on cortical and subcortical brain structures. *Scientific Reports*, 6, 32760.

[5] Zuo, Y. and Serfling, R. (2000): General notions of statistical depth function. *The Annals of Statistics*, 28(2), 461–482.