Invitation à la soutenance publique de thèse de

Emilie RENARD
Master ingénieur civil en mathématiques appliquées

Pour l’obtention du grade de Docteur en sciences de l’ingénieur et technologie

« Extracting information from multiple datasets by matrix factorization and common subspace computation »

qui se déroulera
le mardi 03 septembre 2019 à 16h
Auditoire A03
Place des Sciences
1348 Louvain-la-Neuve

Jury members :

Prof. Pierre-Antoine Absil (UCLouvain), supervisor
Prof. Vincent Blondel (UCLouvain), supervisor
Prof. Raphaël Jungers (UCLouvain), chairperson
Prof. Michel Verleysen (UCLouvain), secretary
Prof. Andrew Teschendorff (University College London, UK)
Prof. Kyle Gallivan (Florida State University, Florida, USA)
Dr. Thibault Helleputte (DNAlytics, Belgique)

With the increasingly easiness to measure and store data, dealing with different datasets representing similar phenomena is becoming more and more common. An example in bioinformatics is gene expression data, where the same disease can be studied in different hospitals at different times by measuring the gene expression levels of patients. It is then desirable to merge the different measures obtained, in order to improve the robustness of the subsequent analysis.

In this thesis, we investigate different ways of extracting common information from multiple datasets with common features. We propose two different types of approach: removing the differences across datasets, and, keeping the common components.

To remove the differences across datasets, we use a matrix factorization approach: we develop a new spatiotemporal version of independent component analysis that we apply to all concatenated datasets, and use the obtained components to model the differences to be removed. The approach is validated on gene expression datasets, and compared to other existing methods in the wider context of a classification task.

In the second part of the thesis, we propose a minimax formulation to model the problem of finding a common subspace across a set of subspaces associated with the considered datasets. We then provide different algorithms to solve this minimax problem, with some convergence guarantees, and compare their performances on synthetic and real data.