A method for statistical learning in large databases of heterogeneous imaging, cognitive and behavioral data.

Luigi Antelmi, Marco Lorenzi, Valeria Manera, Philippe Robert, Nicholas Ayache

EPICLIN 2018

A method for statistical learning in large databases of heterogeneous imaging, cognitive and behavioural data.

Authors: Luigi Antelmi[1,*], Marco Lorenzi[1], Valeria Manera[2,3], Philippe Robert[3,4], Nicholas Ayache[1]

* corrisponding author: luigi.antelmi@inria.fr

Affiliations:
1 UCA, Inria Sophia Antipolis, Epione Research Project.
2 UCA, Inria Sophia Antipolis, Stars Research Project.
3 CoBTeK, University of Nice Sophia Antipolis.
4 Centre Memoire, CHU de Nice.

The authors declare no conflict of interest

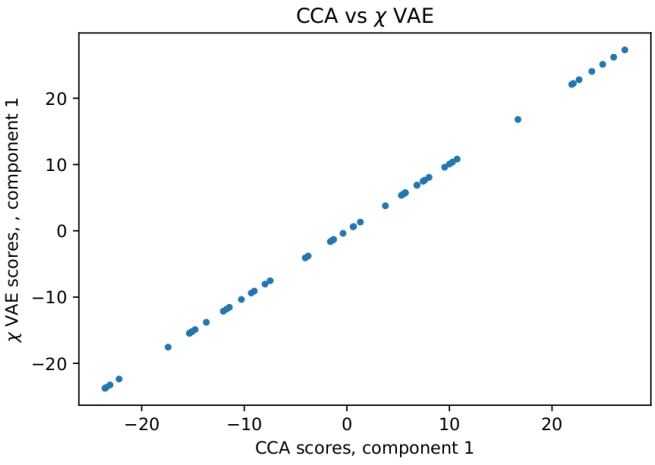
Abstract

Introduction The aim of this study is to develop a generative and probabilistic statistical learning model for the joint analysis of heterogeneous biomedical data. The model will be applied to the investigation of neurological disorders from collections of brain imaging, body sensors, biological and clinical data available in current large-scale health databases. The resulting methodological framework will be tested on the UK Biobank, as well as on pathology-specific clinical data, as provided by the ADNI, or INSIGHT initiatives.

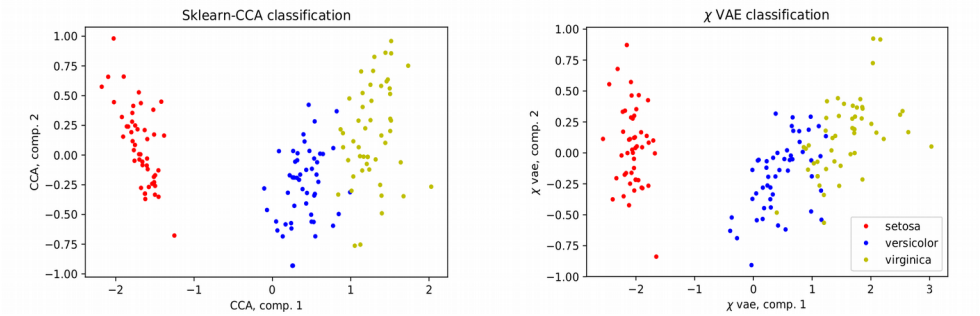
Methods We propose a variational approximation of Bayesian Canonical Correlation Analysis (CCA). The proposed formulation is inspired by current advanced in variational learning, and offers the potential to scale to high-dimensional observations, such as medical images and arrays of biological data.. We proved that the variational lower bound can be optimized through modern learning libraries such as Torch and TensorFlow.

Results We currently benchmarked the method with respect to classical CCA on both synthetic data and on the classical benchmarking datasets in machine learning (IRIS dataset). With respect to the synthetic dataset (Fig. 1A), we observed a strong agreement between the score components computed with classical CCA and our method. Moreover, the classification results on IRIS showed that the two methods essentially provide the same latent representation (Fig. 1B).

Conclusion Our method shows promising results for the future application to medical data. The method is computationally efficient and scalable, hence able to process complex multivariate multidimensional datasets. We expect to highlight meaningful relationship among biomarkers that could be used to develop optimal strategies for disease classification, quantification, and prediction. In the future the proposed approach will be tested in several experimental settings:
1) Classification/stratification;
2) Prediction and imputation from a set of observed data (e.g. predict biological and clinical output from medical imaging information).



(A)



(B)

Fig 1. (A) Comparison of score components computed with classical CCA and our method from the synthetic dataset. The perfect linear relationship between two methods applied to this datasets means that they both capture the same latent representation. (B) Classification of the IRIS dataset. Classical CCA on the left; our method on the right.