DSS Statistics Seminar

April 23, 2021, 12:00

https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09

Passcode: 432940

Probabilistic partial least squares methods for data integration

Jeanine Houwing-Duistermaat

Department of Statistics, University of Leeds, UK
Department of Statistics, University of Bologna, Italy
Department of Biostatistics, University Medical Center Utrecht, The Netherlands.

Many studies collect multiple omics datasets to gather novel insights about different stages of biological processes. For joint modelling of these datasets, several data integration methods have been developed. These methods address high dimensionality, within and across datasets correlation, and the presence of heterogeneity among datasets due to representing different biological processes and using different measurement technologies. Most methods, neither provide statistical evidence for a relationship between the datasets nor identify relevant variables that contribute to this relationship.

We propose a probabilistic latent variable modelling framework for inferring the relationship between two omics datasets. These methods reduce dimensionality and capture correlations by forming components that are linear combinations of the variables. The correlation structure is modelled by joint and data specific components. We propose maximum likelihood estimation of the parameters and formulate a test statistic for the null hypothesis of no relationship.

We evaluate our methods via simulations. Under the null hypothesis, the test statistic appears to approximately follow the normal distribution. Our method outperforms existing methods for small and heterogeneous datasets in terms of selecting relevant variables and prediction accuracy. We illustrate the methods by analysing omics datasets from a population cohort and a case control study.

SAPIENZA
UNIVERSITÀ DI ROMA