Working Group on Statistical Learning Seminar

Title: Compressing DNA data for prediction: the benefits of downsizing

Speaker: Silvia D'Angelo (University College Dublin)

Date: Mon 29th March 2021 at 3:00PM

Location: Online

Abstract: Prediction of phenotypes from DNA is a relevant yet complex task to perform, due to the usually prohibitive dimensions of DNA data. Indeed, while large data dimensionality potentially yields more detailed information on phenotypes, it also poses multiple computational issues. Recent works proposed to tackle such issues via the implementation of reduced ranked regression methods.

While reduced ranked regression approaches have substantially lowered computational times, their focus is on a restricted pool of models, that is regression models. Our intent is to provide a general framework to work with large DNA datasets in computationally feasible times, while allowing the investigation of more complex model specifications, as for example mixed effect models.

Exploiting specific characteristics of SNPs data, we propose a lossy compression strategy that allows to reduce data dimensionality, while still preserving the original information contained in the data. Compressed data can be then used to feasibly investigate a multitude of candidate models.

Results from simulated and real world examples will be presented, offering both a comparison with existing reduced ranked regression approaches and an illustration of mixed effect models.

Join Zoom Meeting: https://ucd-ie.zoom.us/j/68316324831