BRIGHAM AND WOMEN'S HOSPITAL
HARVARD MEDICAL SCHOOL

181 Longwood Avenue
Boston, Massachusetts 02115-5804

Department of Medicine
Channing Division of Network Medicine

Channing Methods Seminar

December 19 (Tuesday), 2023, 11AM (ET)

MCP 5th-floor large conference room
https://us02web.zoom.us/j/579497999?pwd=cHNlWHMzWUlFUUVJTG1EeVJmY05aQT09
Meeting ID: 579 497 999
Passcode: 844168

Anna Neufeld, PhD
Fred Hutch Cancer Center

Data thinning to overcome double dipping

Abstract: We refer to the practice of using the same data to fit and validate a model as double dipping. Problems arise when standard statistical procedures are applied in settings that involve double dipping. To circumvent the challenges associated with double dipping, one approach is to fit a model on one dataset, and then validate the model on another independent dataset. When we only have access to one dataset, we typically accomplish this via sample splitting. Unfortunately, in many unsupervised problems, sample splitting does not allow us to avoid double dipping. In this talk, we are motivated by unsupervised problems that arise in the analysis of single cell RNA sequencing data. We first propose Poisson count splitting, which splits a single observation drawn from a Poisson distribution into two independent components. We show that Poisson count splitting allows us to avoid double dipping in the context of our motivating problems. We next generalize the count splitting framework to a variety of distributions, and refer to the generalized framework as data thinning. Data thinning is a very general alternative to sample splitting that is useful far beyond the context of single-cell RNA sequencing data, and, unlike sample splitting, can be applied in both supervised and unsupervised settings.

Bio: Anna Neufeld completed her PhD in statistics from the University of Washington in 2023, under the guidance of Professor Daniela Witten. She worked on problems related to "double dipping" and testing data-driven hypotheses, including problems related to the analysis of single cell RNA sequencing data. She is now a postdoctoral research fellow at the Fred Hutch Cancer Center in Seattle, WA, working with Professor Jeff Leek on problems related to multiple testing in the analysis of genomic data.

Hosted by Yang-Yu Liu