

Stanford University

Department of Statistics

DEPARTMENTAL SEMINAR

4:30pm, Tuesday, November 5, 2019
McCullough Building (04-490) Room 115

Refreshments served at 4pm in Sequoia Lounge.

Speaker: Audrey Qiuyan Fu, *University of Idaho*

Title: **Imputation and causal inference in genomics**

Abstract:

Genomic data can be complex, large, noisy and sparse. Here I will discuss two problems we have worked on. The first problem deals with the highly sparse data from single-cell experiments of gene expression. These data contain a large number of zeros ($> 80\%$); many of these zeros are missing values rather than no expression. Underlying these data are complex regulatory relationships among genes, as well as potentially many cell types with different gene expression profiles. We took a deep learning approach and designed imputation methods based on autoencoders. We generated synthetic data using real single-cell data to evaluate the performance, although the theoretical properties of autoencoders for imputation are yet to be understood.

The second problem deals with causal inference: can we learn the biological mechanism directly from genomic data? For example, which genes regulate which other genes? And which genes are targeted by drugs? Genetic variation makes this inference possible (under certain assumptions), as it provides randomization among the individuals: this is known as the principle of Mendelian randomization in genetic epidemiology. We extended the interpretation of this principle to capture more causal relationships. We also developed an algorithm for learning causal graphs based on the PC algorithm, a classical algorithm in computer science for inferring directed acyclic graphs.