

STATISTICS SEMINAR

Huizi Wang, Ph.D. Candidate

Efficient Classification Methods for Sparse High-Dimensional Count Data with Model Selection as Motivated by Microbe Finder

**Friday, October 10, 2025
2:30 – 3:30 pm
310 MSCS**

Abstract: Existing classification methods for high-dimensional (HD) count data, such as penalized multiple logistic regression (MLR) and a penalized multivariate discriminant analysis, aim to model zero-inflation and/or over-dispersion, avoid overfitting, and be computationally efficient. However, no one method addresses all of these issues simultaneously. This dissertation provides new methods for HD classification analysis with count data that are computationally efficient, can model zero-inflation and over-dispersion, and allow for model/variable selection while avoiding over-fitting. First, we show analytically that multivariate model classifiers can be written as MLR classifiers, where the regression coefficients include parameters for zero-inflation and over-dispersion. Then, we develop penalized method of moments (PMOM) estimators for the regression coefficients, where a soft-thresholding function applies an ℓ_1 penalty. Thus, methods allow for variable selection and avoid overfitting. Further, they are computationally efficient because PMOM estimators use soft-thresholding functions and avoid iterative procedures. The methodology is applied to two HD Microbe Finder data sets, where the goal is to classify a sample as pathogenic or healthy. Analysis highlights that our methods allow for “e-probes” to be selected, and indeed model zero-inflation and over-dispersion well. Cross-validation estimates of the area under the receiver operating characteristic curves also illustrate that the proposed methodology typically has higher classification power than the competing method. Simulation study shows that our methods typically perform better than the competing method with various p .

Refreshments immediately following seminar in Room 309 MSCS.