May 18th, 4:00 PM - 4:30 PM

Multiclass classification and prevalence estimation with applications to SARS-CoV-2 antibody assays

Rayanne Luke
Johns Hopkins University, National Institute of Standards and Technology, rluke3@jhu.edu

Paul N. Patrone
National Institute of Standards and Technology, paul.patrone@nist.gov

Anthony J. Kearsley
National Institute of Standards and Technology, anthony.kearsley@nist.gov

Title: **Multiclass classification and prevalence estimation with applications to SARS-CoV-2 antibody assays**

Authors: Rayanne A. Luke, Paul N. Patrone, Anthony J. Kearsley

Abstract:

Accurate classification strategies are needed to interpret diagnostic test results. These include problems with more than two classes, such as to distinguish uninfected, previously infected, and vaccinated individuals. Classification is further complicated when the relative fraction of population in each class, or multiclass prevalence, is unknown. A common misconception is that prevalence estimation requires classification. We develop a multiclass prevalence estimation method independent of classification and an associated classification scheme that minimizes the prevalence-weighted average of false classifications for an arbitrary number of dimensions and classes. Our work hinges on constructing probability models for data that are inputs to an optimal-decision theory framework. This method is applicable to a variety of settings.

We study, as an example, the tri-class problem for SARS-CoV-2 negative, previously infected, and vaccinated individuals. We first use synthetic data to illustrate main ideas and show that practical uncertainties (< 2 %) for multiclass prevalence estimation and classification can be achieved in real-world settings. We then separate antibody data from Ainsworth et al. (2020) and Wei et al. (2021) into training and test populations, build conditional probability models based on the training data, estimate prevalences for the test population, and perform classification. Bootstrapping is performed to find an average multiclass prevalence estimate error of 4.24 % leading to classification errors that are marginally worse for the test data with estimated prevalences compared to the training data with known prevalences. This work is the first of its kind to estimate multiclass prevalence for SARS-CoV-2.