Overview of the dissertation *Statistical Methods for Mass Spectrometry Proteomics Data with Multiple Membership Structure*

This dissertation addresses practical challenges in the statistical analysis of mass spectrometry–based proteomics data, requiring both the development and benchmarking of new statistical models, as well as the design and implementation of supporting software. A central focus is on problems that fall into the class of *multiple membership* models. Unlike standard analysis of variance, where each observation belongs to exactly one group, multiple membership arises when observations are simultaneously associated with several groups, contributing information to the estimation of multiple effects at once. Two motivating examples of this structure are considered in detail.

In **protein quantification**, a related problem arises in the inclusion of *shared peptides*, that is, peptides that map to more than one protein. Standard approaches often discard these peptides, thereby reducing statistical power and introducing biases. Modeling shared peptides as belonging to multiple proteins creates a multiple membership problem: a single peptide's observed abundance informs the abundance estimates of each protein to which it maps. We proposed a statistical model that extends existing `MSstats` protein quantification approach by introducing weights that describe the consistency between observed peptide-level quantitative data and unknown protein-level abundances of interest. We fit the model using biconvex optimization methods. We evaluated the method on both simulated and real data, focusing on two key aspects: the precision of estimating relative protein abundances across biological conditions (e.g., healthy vs. diseased patients), and the statistical properties of the resulting significance tests.

In bottom-up **hydrogen–deuterium exchange mass spectrometry (HDX-MS)**, the experimental output is peptide-level spectra, whereas the scientific goal is to infer exchange probabilities at the level of protein segments, often individual residues or short contiguous stretches. Each observed peptide spans multiple residues, and many residues are covered by several overlapping peptides. As a result, a single residue contributes to the uptake of multiple peptides, and each peptide provides information about several residues. This overlap creates a multiple membership structure: peptide-level data must be modeled as a composite of the exchange behavior of all residues it contains, while each residue's parameters are informed by multiple peptides. We proposed a statistical model that represents overlapped segments of peptide sequences by multinomial random variables and models observed spectral data based on convolutions of their distributions. We discuss fitting this model under various assumptions and evaluate it in terms of its ability to recover the segment-level exchange probabilities based on isotopic distributions of peptides in observed spectra based on simulations and biological case studies.

Finally, this dissertation reports on the redesign and refactoring of `MSstats`, a widely used statistical software suite for the analysis of differential protein abundance from mass spectrometry data, developed by Prof. Olga Vitek's group at Northeastern University (US). We evaluate the proposed changes in terms of complexity of the structures of packages, extensibility, and performance.

We describe practical implementations of both methods. The statistical method for incorporating shared peptides introduced in this thesis has been integrated into the `MSstats` framework.