ORIE 5741 - Project Proposal

Jiacheng Cao (jc2992), Fuhan Cong (fc334)

1. Background

Parkinson's disease is a serious neurological disorder that affects various bodily functions and currently has no cure. By 2037, it's estimated that 1.6 million people in the US will be affected, resulting in significant economic costs. Through the use of data science, gaining a better understanding of these abnormalities could lead to important discoveries for developing new pharmacotherapies that can slow down or cure Parkinson's disease.

2. Question

We try to predict the course of Parkinson's disease (PD) using protein abundance data and hope to improve our understanding of the molecular mechanisms underlying PD. Ultimately, this data analytics could give some suggestions for the development of new treatments or therapies for Parkinson's disease.

3. Dataset

1. **Peptides Dataset.** Peptides are the component subunits of proteins. This dataset includes associated proteins, amino acids included in the peptide and the frequency of the amino acid in the sample for each patient diagnosed with Parkinson’s Disease.
2. **Proteins Dataset.** This dataset includes protein expression frequencies aggregated from the peptide level data, as some proteins contain repeated copies of a given peptide.
3. **Clinical Dataset.** This dataset includes the patients’ scores on the Unified Parkinson's Disease Rating Scale, which indicates the severity of symptoms, and indicators for whether or not the patient is taking any medication during the UPDRS assessment.

4. Proposed Strategies

The first step is data preprocessing, involving missing data, outliers, and imbalanced situations. Then, we will conduct EDA to determine if there are any relationships among the features. For example, the levels of certain peptides may have a significant impact on the progression of PD in individual patients. After that, we are about to build predictive models to predict the progression of PD. We will try baseline and higher-level models and improve accuracy. Also since it is time series data, we may also try relative methods.