Sparse Proteomics Analysis

Toward a Mathematical Theory for Feature Selection from Forward Models

Martin Genzel

AFG Oberseminar, WiSe 2015/16
Technische Universität Berlin

email: genzel@math.tu-berlin.de

December 17, 2015

Abstract

Tumor diseases, such as cancer, rank among the most frequent causes of death in Western countries. The clinical research of the last decades has shown that the pathological mechanisms of many diseases are manifested on the level of *protein activities*. In order to improve the clinical *treatment options* and *early diagnostics*, it is therefore necessary to better understand protein structures and their interactions. The related research field of *proteomics* focuses on analyzing the so-called *proteome*, which denotes the entire set of proteins of a human individual at a certain point of time. Unfortunately, proteomics-data, e.g., produced by *mass spectrometry*, is usually extremely *high-dimensional*. Therefore, it is a very difficult task to extract a *disease fingerprint*, which is a small set of proteins allowing for an appropriate classification of a patient's health status.

In the first part of this talk, we will see that the assumption of *sparsity* can help us to cope with this challenge. In this context, the method of *Sparse Proteomics Analysis* (SPA) will be introduced, enabling us to build sparse and reliable classifiers. The second part of the talk is then devoted to a theoretical foundation of SPA. Relying on a simple linear *forward model* for the data, we will see that very recent results from *high-dimensional estimation theory* can be used to prove rigorous recovery guarantees.