Siva Rajesh Kasa

Research Statement

✉ kasa@u.nus.edu
kasakh.github.io

High Dimensionality and Misspecification

My research aims to address the question, 'How we can we model high dimensional real-world data in a statistically robust, flexible, and scalable manner and draw business insights from it?' High dimensional data is becoming increasingly relevant in a today's world - the number of subjects is usually fixed; however, the number of features/attributes we have about each subject is growing everyday, thanks to IOT devices, wearables, sensors, etc. The present deep learning research addresses the scalability issue, however it lacks in interpretability. On the other hand, classic statistical techinques whilst being interpretable, do not scale well to large datasets. My research attempts to bridge this gap.

High dimensionality is a relevant and common problem across multiple domains - health care, natural language processing, marketing (customer segmentation), etc. In the past, researchers have well studied and documented the problems with high-dimensionality - two main problems are non-full rank matrices and overparametrization. While most of the earlier modeling techniques have been geared towards low-dimensional paradigm, there is a pressing need for analogous techniques in the high-dimensional spectrum. Another aspect while modeling real world data using statistical inference is the assumption that data indeed follows our model. However, this assumption is often simplistic and rarely true. This leads to the question how sensitive are our insights/conclusions to our modeling assumptions?

The problems of high-dimensionality and misspecification are further exacerbated in the unsupervised learning. In our work on 'Clustering with Misspecified Models', we demonstrate how a slight misspecification in our model can lead to a drastic reduction in clustering accuracy. We propose a solution to alleviate this problem which uses automatic differentiation tools. We demonstrate the usefulness of our proposed solution by clustering the electronic medical records in MIMIC-III dataset (100+ GB) and drawing useful business implications from these clusters.

Background

My formal exposure to advanced statistical methods provides me with a solid foundation to develop statistical modeling techniques for high-dimensional data. Having TAed a course in Information Theory, I have a decent exposure to information-theoretic understanding of modeling techniques and what can be the upper bound on the amount of information we can extract from the data. I also led a research-project towards developing semi-parametric modelling solutions for high-dimensional data, which is published in Bioinformatics, Oxford Press. Details of my previous course and research work are available in my CV, but I would like to highlight my interests in developing scalable and flexible models for high-dimensional data. I believe my knack for problem-solving coupled with a flair for rigorous theory is what puts me in good stead.