Learning with Tensors: From Theory to Application

Miaoyan Wang, Department of Statistics

I work at the intersection of **Machine Learning** and **Statistics**, with a particular focus on developing new methods for large-scale data fusion and signal processing.

Increased demand for high-dimensional data analysis throughout all areas of biomedical science has led to the explosive growth in machine learning in the past decade. My research is driven by the need for statistical insights into the success of machine learning in emerging biomedical problems. **The permeating theme in this proposal is to develop science-driven models to analyze high-dimensional data.** My group will release user-friendly software packages that facilitate individuals, academia, industry, and society for analyzing complex data in high dimension, namely, "big data".

Tensor data: new link between signal processing and data fusion

Tensors are high-dimensional arrays (Figure 1). Recent advances in high-throughput sequencing technology have transformed biomedical research into a data-intensive field where data are naturally generated in tensor form. One example of biomedical tensor data arising from my current collaboration is multi-tissue, multi-individual gene expression (Figure 1a). The completion of Genotype-Tissue Expression provides a huge compendium of tensor data consisting of millions of expression measurements from $\sim$ 20,000 genes across 544 individuals and 53 human tissues. Understanding the multifactorial patterns of whole-genome transcriptome variation is crucial to unravel gene networks and tissue functions, thereby broadly facilitating research efforts to unravel genetic basis for personalized disease.

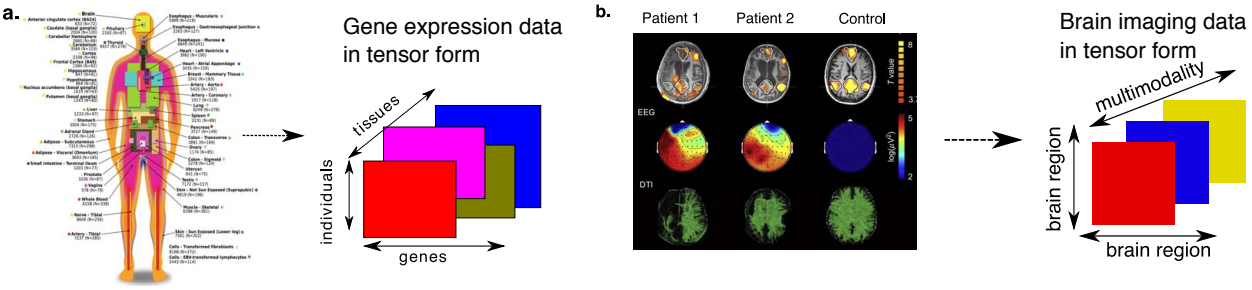

Figure 1: Examples of tensor data in biomedical research. (a) GTEx project collects gene expression profiles of over 20,000 genes from 544 individuals across 53 human tissues. (b) HCP collects multimodality imaging including EEG, DTI, fMRI from over 1,200 individuals.

Methods built on tensors provide generalized tools to capture complex structures in data that lower-order methods may fail to exploit. Existing analytical methods mostly focus on two-dimensional data, namely matrices. Early study on tensor methods seeks parsimonious representation within the high-dimensional tensor data. However, tensor-based methods are fraught with challenges due to increasing dimensionality and complexity. Ideally modern tensor methods will allow researchers to examine complex interactions among tensor entries and between multiple tensors, thereby providing solutions to questions that cannot be addressed by traditional analysis.

The proposal focuses on two main directions. The outlined project will build new links between mathematics and data science, and also produce new areas in which machine learning and biomedical applications can combine and complement each other.

Efficient algorithms for learning high-dimensional tensors. Tensor decomposition problems arise frequently in applications such as neuroimaging, recommendation system, topic modeling, and sensor network localization. In the simplest form, tensor decomposition can be formulated as finding the latent factors from a noisy tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ (Figure **??**c). Classical statistical theory only studies the asymptotical estimation accuracy in an ideal scenario when the computational resource is unlimited. This simplified assumption, however, is barely satisfied in modern science. Data is often incomplete with missing or costly

labels, the distribution may change overtime, and the practical considerations such as time and memory imposes additional constraints on the algorithm.

My general approach to the above problem is to carve out a broad range of specially-structured tensors that are useful in practice, and to understand, mathematically, the trade-off between statistical accuracy and computational efficiency. I have shown in a series of recent work that the empirical estimator often exhibits a two-component error for a broad range of tensor models. The key factor that governs the trade-off is the *intrinsic dimension*, which tightly captures, in a minimax sense, the complexity of the high-dimensional tensor problems. This notion reveals the fundamental interplay between the computational efficiency and statistical limit. My main aim for this direction is to advance the theory for designing accurate, robust and efficient tensor algorithms.

Impact and Significance:
- Generate new theoretical understanding on high-dimensional data problems that inform practice.
- Manage and tackle uncertainties related to statistical modeling with large datasets.
- Foster interdisciplinary research partnerships that integrate mathematics with biomedical challenges.

Integrative analysis of omic data. My goal in this direction is to integrate the most appropriate state-of-the-art statistical principals and machine learning tools into going omics research. My group has recently developed an efficient tensor-based clustering tool to successfully identify key expression signatures in the GTEx data. Our decomposition algorithm discovers three-way clusters with higher accuracy, while being 11× faster, than the competing methods. We are currently generalizing the method for integrative analysis of omics data, where multiple types of omics measurements (such as gene expression, DNA methylation, microRNA) are collected in the same set of individuals. Specifically, we will apply efficient statistical inference methods for identifying higher-order expression modules arising from the interactions among individuals, genes, tissues, and multiple modalities. We will develop a useful software package that allows researchers to utilize deep learning in biomedical research. The elucidation of expression signatures at the whole-genome scale have the potential to inform the personalized disease risk and variability in drug response.

Impact and Significance:
- Rigorous analysis of large-scale data from scientific studies on genomics and neuroimaging.
- Powerful predictive models for complex phenomena in the gene-gene interaction, brain network connectivity, and high-order interaction in the omics data.
- Efficient forecasting and decision-making tools that facilitate personalized medicine.

Deliverables and Milestones