Tensor Methods and Applications to the Data Science

Miaoyan Wang

My research is in the intersection of mathematics, statistics, and computer science, with a focus on biomedical data analysis. The proposed project is to develop a framework of mathematical theory, statistical models, and efficient algorithms to analyze high-dimensional multiway array data. Recent advances in high-throughput sequencing technology has fundamentally transformed biomedical research into a data-intensive filed. As data continues to grow at an ever increasing rate, so does the need for efficient and powerful methods to extract information from data. Through developing efficient methods for analyzing "big data", I strive to push the boundary of biomedical research further, in a form that is useful to individuals, academia, industry, and society.

I have a unique combination of training backgrounds in mathematics (2006-2010), statistics (2010-2015), genetics (2015-2017), and computer science (2015-2017). During my career, I am fortunately to have worked closely with many fellow scientists on cutting-edge big data problems, which have spurred new goals in my math research and shaped my scientific approach. **The prevailing theme in my research is to develop powerful machine learning theory and methods for advancing knowledges in biomedical science.** In this regard, my work will link math and life science, the two areas I have been working on extensively, and also develop new mathematical areas based on the questions raised in the applied endeavors.

Tensor Data: New Links between Math and Life Science

Big data, Deep learning are terms one encounters everywhere, and we are told that they will define a new science. Rapid developments in modern technologies have made large-scale data readily available across science and engineering. Tensors, or multi-way arrays, provide a generalized data structure and serve as a foundation in many learning procedures. Methods built on tensors provide powerful tools to capture complex structures in data that lower-order methods may fail to exploit. However, tensor-based methods are fraught with challenges. Extending familiar matrix concepts to tensors is hard, and most computational problems regarding tensors are NP-hard. Analyzing tensor data with increasing dimensionality and growing complexity presents serious challenges – the classical off-the-shelf methods cannot keep up with them, and new approaches have to be developed.

Below I give two examples of biomedical tensor data arising from my current collaborations, and for the further development of which I will seek new linkages with mathematics.

Multi-tissue, multi-individual gene expression. A typical multi-tissue experiment collects gene expression profiles from different individuals in a number of tissues. The recent completion of Genotype-Tissue Expression (GTEx, Figure 1a) project has provided unprecedented opportunities to investigate transcriptome diversity and complexity. The study results in a huge compendium of tensor data consisting of millions of expression measurements from $\sim$ 20,000 genes across 544 individuals and 53 human tissues, including 13 brain regions, adipose, heart, artery, skin, and more. In this setting, variation in the expression levels arises due to contributions specific to genes, tissues, individuals, and interactions thereof. 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.

Multimodal neuroimaging data analysis. Neuroimaging studies aim to characterize the human brain connectivity in response to stimulus or physiological changes. As of the fall of 2019, the human connectome project (HCP) has released massive datasets representing the anatomical and functional connectivities within human brains from over 1,200 individuals. Adjacency matrices (or networks) are common tools to describe the brain connectivity, where edges (or connections) join a set of nodes (or brain voxels). Different imaging measurements are utilized to construct the brain networks (Figure 1b), including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and diffusion tensor imaging (DTI). A key feature is that the acquired networks are huge in size and each possesses complex spatial temporal structure. Integrative analysis is thus essential for investigating the commonality and variability between the networks.

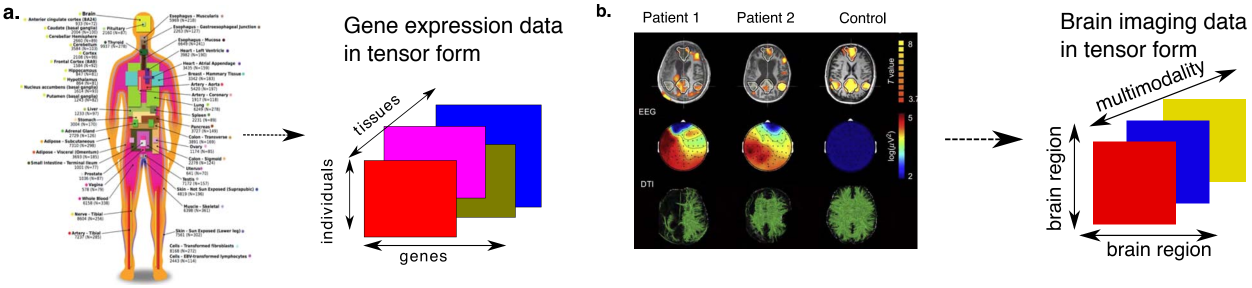


Figure 1: (a) GTEx project collects gene expression profiles of over 20,000 genes from 544 individuals across 53 human tissues. Figure modified based on Timpson et al. [2018]. (b) HCP collects multimodality imaging including EEG, DTI, fMRI from over 1,200 individuals. Figure modified based on Bruno et al. [2011].

In the above two examples and many others applications, scientists are interested in identifying interpretable low-dimensional structure within the high-dimensional tensor data. The research question goes beyond the traditional multivariate analysis: we are interested in the distribution over network-valued "objects" where the objects can be images, networks, manifolds, and arrays. **The main goal of this proposal is to develop a framework of statistical models, scalable algorithms, and mathematical theory to analyze tensor-valued data.** This 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.

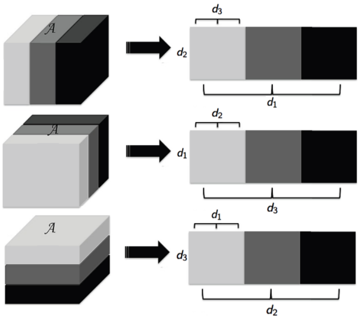
Research Goals and Strategies

The proposed project focuses on tensors of order 3 or greater, known as higher-order tensors. A tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a higher-order generalization of matrix and can be viewed as a multilinear functional that maps K-tuples of vectors to numbers: $(x_1, \ldots, x_K) \mapsto A(x_1, \ldots, x_K) \in \mathbb{R}$. However, higher-order tensors are not simply matrices with more indices; rather, they are mathematical objects possessing special algebraic properties. The challenge stems from the gap from linear algebra to multilinear algebra. The multi-linearity is an active topic that is currently studied in algebraic geometry and has been shown to closely connect to the long-standing problems in computer science: P vs. NP and matrix multiplication.

Below I outline three main directions I plan to pursue with the WiSTEM^{2}D grant. This would build new links between math and natural science, and also produce new areas in which mathematical theory and machine learning applications can combine and complement each other.

Spectral properties for specially-structured tensors. This aim focuses on the spectral properties of higher-order tensors. Here I take the operator norm for illustration but other spectral properties can similarly be discussed. The operator norm of a tensor is characterized by the best rank-one approximation in the least square sense. Given an order-K tensor, each possible unfolding operation is represented using a partition π of $\{1, ..., K\}$, where a block in π corresponds to the set of modes that should be combined into a single mode. Figure 2 illustrates an example for $K = 3$. Earlier I established a new framework representing all possible tensor unfoldings using the partition lattice, and showed that the spectral norm comparison bounds scale polynomially in the dimensions $\{d_n\}$ of the tensor with powers depending on the corresponding partition and block sizes. The operator norms of all possible tensor unfoldings together define what we coin a "norm landscape" on the partition lattice. To our knowledge, this is the first result to provide a full picture of the norm landscape over all possible tensor unfoldings.

My earlier results are already useful in several machine learning applications but should be interpreted with caution: the comparison bounds deal with the worst-case scenarios with arbitrary tensors. Tensors arisen in applications are often specially-structured. The structures of interest include, but are not limited to, low-rankness, sparsity, non-negativity, orthogonality, or having blocks. We plan to study a range of spectral properties of these tensors. I expect that our newest attempt that focuses on structured tensors will provide the right framework to take this one step further. We will not only show the spectral properties of general tensors, but also finely disentangle the structured signals with the stochastic noise. These will provide mathematical foundations to mostly novel tensor-based learning algorithms.

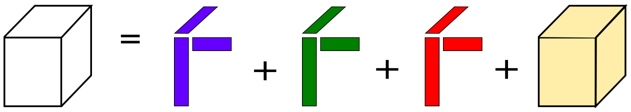


$\text{Unfold}_\pi(\mathcal{A})$	Partition $\pi \in \mathcal{P}_{[3]}$
$\in \mathbb{R}^{d_2 \times d_1 d_3}$	$\pi = \{\{2\},\{1,3\}\}$
$\in \mathbb{R}^{d_1 \times d_2 d_3}$	$\pi = \{\{1\},\{2,3\}\}$
$\in \mathbb{R}^{d_3 \times d_1 d_2}$	$\pi = \{\{3\},\{1,2\}\}$

Figure 2: An order-3 tensor and its matricizations. The set of all possible unfoldings is in one-to-one correspondence with the set of partitions of $\{1,2,3\}$.

Breaking previous limits: We plan to investigate the spectral properties for both deterministic and random tensors. For example, what is tensor analogy of the Tracy-Widom law? Solving these questions would nicely connects multilinear algebra with probability theory and machine learning applications.

Statistical and computational limits in the tensor decomposition. Tensor decomposition problems arise frequently in applications such as neuroimaging, recommendation system, topic modeling, and sensor network localization. Developing efficient and accurate algorithms for tensor decomposition receives much attention recently. My goal is to develop efficient tensor methods with theoretical guarantees and to make use of the powerful tensor tools for elucidating complex structure in multi-mode data. In the simplest form, tensor decomposition can be formulated as finding the latent factors $\{\mathbf{u}_r\}$ from a noisy tensor $\hat{\mathcal{A}} \in (\mathbb{R}^d)^{\otimes K}$:



$$\tilde{\mathcal{A}} = \lambda_1 \mathbf{u}_1^{\otimes 3} + \lambda_2 \mathbf{u}_2^{\otimes 3} + \lambda_3 \mathbf{u}_3^{\otimes 3} + \mathcal{E}$$

where the $\Theta = \sum_{r=1}^{R} \lambda_r \mathbf{u}_r^{\otimes K}$ is called the signal tensor, $\{\mathbf{u}_r\}$ is a set of unit vectors in $\mathbb{R}^d$, λ_is are positive scalars in $\mathbb{R}$, and the perturbation $\mathcal{E}$ represents a small (stochastic) noise tensor.

Classical statistical theory only studies the asymptotical estimation properties when the computational resource is unlimited. However, modern data science is more concerned on the trade-off between statistical limit and computational efficiency. In this regards, we aim to study some fundamental issues of structured tensor decomposition, including the signal-to-noise ratio under which a reliable estimation is possible, the intrinsic hardness, in terms of minimax rate, the estimation properties that are specific to a particular algorithm, and those general to all algorithms. My preliminary results reveal that, for a broad range of low-rank tensors, the empirical estimate often exhibits a two-component error bound:

$$\text{Loss}(\hat{\Theta}^{(t)}, \Theta_{\text{true}}) \leq \underbrace{C_1 \rho^t \text{Loss}(\hat{\Theta}^{(0)},\ \Theta_{\text{true}})}_{\text{algorithmic error}} + \underbrace{C_2 d^{-(K-1)/2}}_{\text{statistical error}},$$

where $\hat{\Theta}^{(0)}$ is the initialization, $\hat{\Theta}^{(t)}$ is the estimation at the t-th iteration, $\rho \in (0,1)$ is a contradiction parameter, and $C_1, C_2 > 0$ are two constants that do not depend on dimensions. This bound reveals the interesting interplay between the computational and statistical errors, and it immediate suggests a practical tradeoff between the two aspects. Understanding the gap between the algorithm property and the theoretical optimality is one of our main goals, which in turn offers a useful guide to the algorithm design.

Our general strategy is to carve out a broad range of specially-structured tensors that are useful in practice, and to develop theoretically sound and empirically efficient methods for analyzing these high-dimensional tensor data.

Tensor methods for integrative analysis of omic data. The earlier two directions start from applications but require mathematical development in order to establish the right framework. In this direction, I would like to take the results back to the applications. The aforementioned framework of tensor decomposition will be applied to array-based genomic datasets. My group has recently developed a tensor-based multiway clustering tool to investigate the variation in gene expression levels in the GTEx data. The results are illustrated in Table 1, where we successfully identified groups of genes that perform coordinated biological functions in certain contexts (e.g., specific tissues or individuals) but behave differently in other settings through tissue- and/or individual-dependent gene regulation mechanisms.

Tissue	Gene	Individual		
enriched region	enriched ontology	variance explained by		
		age	gender	ethnicity
cerebellum	dorsal spinal cord development	0.0%	**8.0**%	0.2%
cortex	behavior defense response	**16.7**%	0.6%	1.4%
basal ganglia	forebrain generation of neurons	1.3%	0.8%	1.7%
others	embryonic skeletal system morphogenesis	**10.5**%	0.7%	5.2%

Table 1: Tensor analysis of gene expression in the brain [Wang et al., AOAS'18]. Bold numbers indicate p-value < 0.001.

Along this direction, I will investigate into integrative analysis of omics data, in which multiple types of omics measurements (such as gene expression, DNA methylation, microRNA) are collected in the same set of individuals. In such cases, tensor decomposition may be applied to a stack of data matrices or correlation matrices, depending on the specific goals of the project. Other applications include multi-tissue gene expression studies under different experimental conditions in which one may be interested in identifying 4D expression modules arising from the interactions among individuals, genes, tissues, and conditions. The tensor framework can also be applied to time-course multi-tissue gene expression. We will treat time as the 4th mode and extend the tensor projection approach to identify the time trajectories of 3D expression modules. The elucidation of time-course patten will provide us deeper understanding of expression dynamics.

On the computational side, tensor data arising from genetics and genomics applications are often extremely large and unwieldy, and tensor-based computation is far from reaching the level of maturity of matrix computation. Fortunately, recent advances in randomized matrix techniques have brought a highly successful paradigm to scientific computation: *finding structure with randomness* [Halko et al 2009]. My goal here is to understand how randomized methods interact with classical techniques and to develop efficient randomized tensor algorithms with theoretical guarantees. The preliminary results have shown the promising applications. Because random tensor theory is far from reaching the level of maturity of random matrix theory, further forays in this direction are desirable.

New Possibilities and Outreaches

Undergraduate education. The PI's home school UW-Madison has recently launched a new undergraduate major in data science. As one of the few young faculty members, the PI has been striving to encourage more under-represented students into the STEM field. The PI plans to open up an undergraduate course on *introduction to data science*. This course will expose undergraduate students to a wide range of modern exploratory analytic tools for massive high-dimensional data, and will be closely tied with data-intensive scientific domains. Data science uses language of mathematics but is not only mathematics – it is concerned with transforming "data" into "information". Through project- and inquiry-based learning, I hope to inspire students in learning about data with scientific context.

Planned postdoctoral hires. If this proposal is accepted, the PI would like to hire one postdoctoral fellow on mathematics/statistical machine learning. The PI is a young faculty and currently has one active NSF grant on mathematical aspects of tensors. The PI has no external grant support for postdoc or for the applications proposed in this research though. With the support of STEM^{2}D program, the PI will take initiative to establish joint seminar that focuses on the scientific applications of tensor data analysis. The seminar will invite both mathematician and scientists in data-intensive fields, with the aim to spark new research goals at the interface and to encourage multidisciplinary collaborations.