

Spectral methods for high-dimensional tensor data

1 Research goals and significance

Rapid developments in modern technologies have made large-scale multidimensional data readily available in science and engineering. Tensors, or multi-way arrays, provide a generalized data structure and serve as a foundation for multivariate analysis. Analyzing tensor data with increasing dimensionality and ever-growing complexity requires the development of novel statistical methods. We consider two motivating examples from PI’s previous collaborations, one on a human whole-genome transcriptom study and the other on a neuroimaging study.

Multi-tissue, multi-individual gene expression. The advent of high-throughput sequencing technologies has led to an increasing availability of large multi-tissue data sets which contain gene expression measurements across different tissues and individuals. 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 transcriptom 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 transcriptom 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 2018, the human connectome project (HCP) has released massive datasets representing the anatomical and functional connectivities within human brains from over 1,200 individuals. Networks (or adjacency matrices) 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 common feature is that the acquired networks are huge in size and each possesses complex spatial temporal structure. Like most neuroimaging applications, the connectivity must be understood in the context of measurement modality and subject-level characteristics (e.g, age, gender, disease status). Integrative analysis is thus essential for investigating the commonality and variability between the networks.

In the first GTEx example and many other genomics studies, scientists are interested in identifying a subset of genes that behave similarly in subsets of tissues and individuals. In a broader sense, it is of great importance to identify meaningful low-dimensional structure within the high-dimensional tensor data. As part of the GTEx project, Consortium et al. [2015] performed differentially expression analyses in each of the 53 human tissues with the goal of identifying covariate-related genes. However, their analysis summarizes the gene-by-individual variability in isolation rather than simultaneously analyzing all tissues, the later of which is a goal of this proposal. In the second HCP example, a collection of side information places networks in context. The research question goes beyond the traditional network analysis: we are interested in the distribution over network-valued “objects” where the objects can be images, matrices, networks, or in general, tensors. **We propose to develop a framework of statistical models, scalable algorithms, and relevant theory**

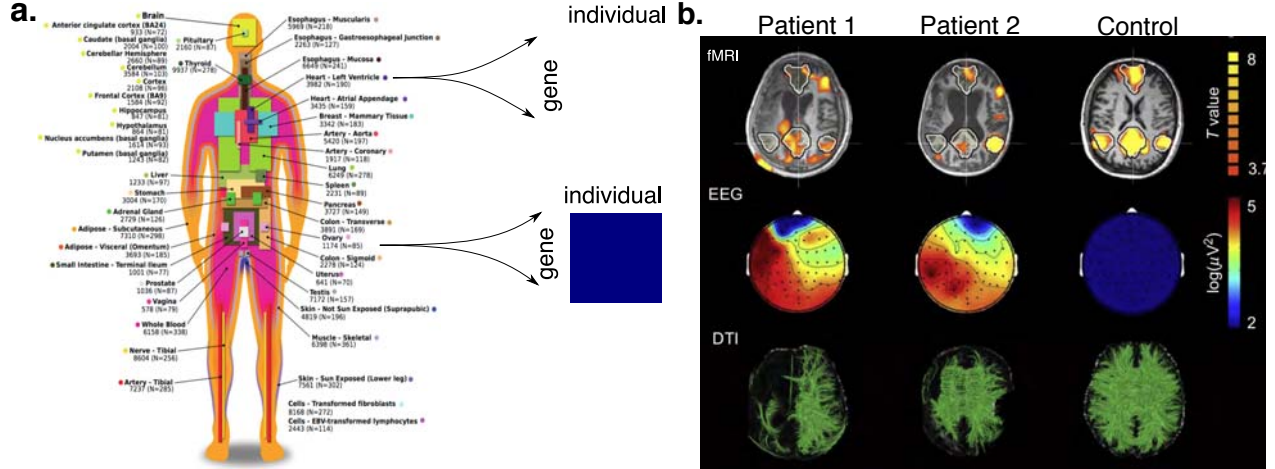


Figure 1: (a) GTEx project collects gene expression profiles of over 20,000 genes from 544 individuals across 53 human tissues. (b) HCP collects multimodality imaging including EEG, DTI, fMRI from over 1,200 individuals. Panel (a) is modified based on [Timpson et al. \[2018\]](#) and Panel (b) based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

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 matrix analysis. The proposed research will focus on spectral methods because of the existing rich set of numerical techniques that one may leverage. Spectral methods view a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ as a multilinear functional that maps K -tuples of vectors to numbers: $(\mathbf{x}_1, \dots, \mathbf{x}_K) \mapsto \mathcal{A}(\mathbf{x}_1, \dots, \mathbf{x}_K) \in \mathbb{R}$. In the matrix case, such techniques rely on singular-values and singular-vectors of the data matrix or the covariance matrix; their statistical properties are well understood [[Johnstone, 2001](#), [Stewart, 1990](#), [Yu et al., 2014](#), [Cai et al., 2013](#)]. However, spectral properties for tensors of order $k \geq 3$ (data tensor or higher-order statistics such as cumulants) are fraught with challenges, and in the worst case the computational problems have proven to be NP-hard [[Hillar and Lim, 2013](#)]. The difficulty lies in the limit from linear algebra to multi-linear algebra, an active topic that is currently studied in algebraic geometry, numerical analysis, and theoretical computer sciences. Fortunately, the tensors sought in applications often possess some special structures, such as being (nearly) low-rank, sparse, nonnegative, or orthogonal decomposable. **Our general strategy is to carve out a broad range of specially-structured tensors that are useful in practice, and to develop efficient statistical methods for analyzing these high-dimensional tensor data.**

Accordingly, this project will provide a statistical framework for spectral methodology in three aspects to facilitate high-dimensional tensor data analysis. The PI's previous research on bridging tensors to matrices [[Wang et al., 2017b](#)] opens up a new inquiry on spectral techniques for tensors. The proposed research will build upon this framework and develop a suite of theory, methodology, and practice for analyzing high-dimensional tensor data.

Aim 1: Spectral theory for specially-structured or random tensors. Tensors are not simply matrices with more indices; rather, they are mathematical objects possessing multilinear algebraic properties. The proposed research will focus on several spectral quantities for deterministic tensors and random tensors. The techniques developed will facilitate the development of novel tensor-based modeling and computations. See Section 2.

Aim 2: Estimation of low-rank tensors from discrete observations. Many real-world multiway datasets are instances of discrete-valued tensors, in which all tensor entries take on discrete values (e.g., ranking) or binary indicators 0/1 (e.g., presence vs. absence). Current tensor decomposition methods do not take the discrete nature into consideration. The proposed research will (1) show how current estimators are suboptimal, (2) develop new estimators that obtain faster rates of convergence, and (3) quantify the information loss due to discretization. See Section 3.

Aim 3: Joint estimation of mean and covariance for tensor-valued data. The dependence among tensor entries can be characterized by a probabilistic distribution on the tensor-valued random variable. Our research will show, under suitable constraints on the mean and covariance structure, it is possible to obtain consistent estimations for both, given only one instance of the tensor data realization. The proposed research will address the problem of multi-way clustering for array data with correlated entries. See Section 4.

Research team and prior NSF support. The PI is a young faculty in the Department of Statistics at UW-Madison. The PI has not been supported by NSF grant before. The proposed research builds on the PI’s established methodological work on functional properties of higher-order tensors [Wang and Song, 2017, Wang et al., 2017b] and applied work on tensor-based genomic research [Wang et al., 2018a,b]. The PI has a unique combination of training backgrounds in statistics (2010-2015), mathematics (2016-2017), computer science (2017-2018), and genomics (2013-2018). The PI’s previous work has a transformational nature involving connections across diverse disciplines, and the PI strives to push the boundary of interdisciplinary research further.

To accomplish the proposed research goals, the PI plans to devote one year to each project. Estimation methods, statistical properties, and computational performance will be carefully studied. Applications to data-intensive fields especially in genomics and neuroimaging will be emphasized. The funded PhD students and Postdoc will contribute to aims 1–3. Undergraduate students will also participate in various aspects of the research. These aims will be incorporated into research training and statistical education of students. The PI’s group will also develop publicly available software packages that are ready for use by fellow scientists.

2 AIM 1: Spectral theory for specially-structured or random tensors

2.1 Backgrounds: from matrices to tensors

Tensors of order 3 or greater, known as higher-order tensors, have recently drawn increased attention in statistical and machine learning applications. An important reason for such an increase is the effective representation of multiway data using a tensor structure. Methods built on tensors provide powerful tools to capture complex structures in data that lower-order methods may fail to exploit. See Anandkumar et al. [2014], Richard and Montanari [2014], Vempala and Xiao [2015], Zhou et al. [2015], McCullagh [2018], Liu et al. [2017], for example.

Before describing the proposed research in Section 2.2, this section introduces the multilinear properties of higher-order tensors. The PI’s established research on tensor spectral properties [Wang et al., 2017b, Wang and Song, 2017] opens up a new horizon of inquiry on spectral techniques for tensors.

An order- K tensor $\mathcal{A} = \llbracket a_{i_1 \dots i_K} \rrbracket \in \mathbb{F}^{d_1 \times \dots \times d_K}$ over a field \mathbb{F} is a hypermatrix with dimensions (d_1, \dots, d_K) and entries $a_{i_1 \dots i_K} \in \mathbb{F}$, for $1 \leq i_k \leq d_k$, $k = 1, \dots, K$. In this section, we focus on real tensors, $\mathbb{F} = \mathbb{R}$, and in Section 3 we will extend to binary-valued tensors, $\mathbb{F} = \{0, 1\}$, and

integer-valued tensors, $\mathbb{F} = \mathbb{N}_+$. We define the *operator norm* of a tensor \mathcal{A} using the associated K -multilinear functional.

Definition 2.1 (Lim [2005]). Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be an order- K tensor. For any $1 \leq p \leq \infty$, the l^p -norm of \mathcal{A} is defined as

$$\|\mathcal{A}\|_p = \sup \left\{ \langle \mathcal{A}, \mathbf{x}_1 \otimes \dots \otimes \mathbf{x}_K \rangle : \|\mathbf{x}_k\|_p = 1, \mathbf{x}_k \in \mathbb{R}^{d_k}, k \in [K] \right\}, \quad (1)$$

where $\|\mathbf{x}_k\|_p$ denotes the vector l^p -norm of \mathbf{x}_k and $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1, \dots, i_K} a_{i_1, \dots, i_K} b_{i_1, \dots, i_K}$ denotes the entrywise inner product between two tensors, \mathcal{A} and \mathcal{B} , of identical order and dimensions. We use the shorthand $[K]$ to denote the K -set $\{1, \dots, K\}$. The special case $p = 2$ is called the spectral norm.

We are currently investigating the operator norm of a tensor viewed as a multilinear functional, because this quantity plays an central role in tensor completion and low-rank approximation problems. A common paradigm in tensor-related algorithms advocates unfolding a tensor into a matrix and applying classical methods developed for matrices. Given an order- K tensor, each possible unfolding operation is represented using a partition π of $\{1, \dots, K\}$, 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 $K = 3$.

The **previous literature** has exclusively studied matricization of \mathcal{A} , i.e., unfolding the higher-order tensor \mathcal{A} into a matrix. Despite the popularity of such techniques, how the functional properties of a tensor change upon unfolding is currently not well understood. Furthermore, tensors are not simply matrices with more indices; rather, they are mathematical objects possessing multilinear algebraic properties. To this end, we propose to study all possible tensor unfoldings.

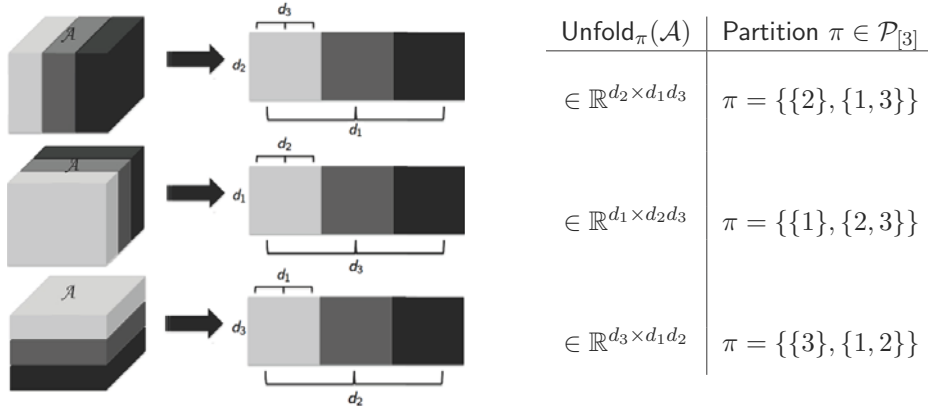


Figure 2: An order-3 tensor and its matricizations. The set of all possible unfoldings is in one-to-one correspondence with the set of partitions of $\{1, 2, 3\}$.

2.2 Spectral norm landscape on the partition lattice

The spectral norms of all possible tensor unfoldings together define a “norm landscape” on the partition lattice. The PI’s established work [Wang et al., 2017b] shows that the comparison bounds scale polynomially in the dimensions $\{d_n\}$ of the tensor with powers depending on the corresponding partitions and block sizes of the unfoldings being compared.

Let $\mathcal{P}_{[K]}$ denote the set of all partitions of $[K]$. Given $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$, we define the map $\dim_{\mathcal{A}}: \mathcal{P}_{[K]} \times \mathcal{P}_{[K]} \rightarrow \mathbb{N}_+$ as

$$\dim_{\mathcal{A}}(\pi_1, \pi_2) = \prod_{B \in \pi_1} \left[\max_{B' \in \pi_2} \left(\prod_{n \in B \cap B'} d_n \right) \right], \quad \text{where } \pi_1, \pi_2 \in \mathcal{P}_{[K]}.$$

The following results establish the general inequalities relating the l_p -norms of two tensor unfoldings corresponding to two arbitrary partitions π_1, π_2 in $\mathcal{P}_{[K]}$. Note that the two partitions π_1, π_2 are not necessarily comparable.

Theorem 2.2 (Wang et al. [2017b]). *Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be an arbitrary order- K tensor, and π_1, π_2 any two partitions in $\mathcal{P}_{[K]}$. Define $\dim(\mathcal{A}) = \prod_{k=1}^K d_k$. Then,*

1. For any $1 \leq p \leq 2$,

$$\frac{[\dim_{\mathcal{A}}]^{-1/p}}{[\dim_{\mathcal{A}}(\pi_1, \pi_2)]^{-1/2}} \|\text{Unfold}_{\pi_1}(\mathcal{A})\|_p \leq \|\text{Unfold}_{\pi_2}(\mathcal{A})\|_p \leq \frac{[\dim(\mathcal{A})]^{1/p}}{[\dim_{\mathcal{A}}(\pi_2, \pi_1)]^{1/2}} \|\text{Unfold}_{\pi_1}(\mathcal{A})\|_p.$$

2. For any $2 \leq p \leq \infty$,

$$\frac{[\dim(\mathcal{A})]^{\frac{1}{p}-1}}{[\dim_{\mathcal{A}}(\pi_1, \pi_2)]^{-1/2}} \|\text{Unfold}_{\pi_1}(\mathcal{A})\|_p \leq \|\text{Unfold}_{\pi_2}(\mathcal{A})\|_p \leq \frac{[\dim(\mathcal{A})]^{1-\frac{1}{p}}}{[\dim_{\mathcal{A}}(\pi_2, \pi_1)]^{1/2}} \|\text{Unfold}_{\pi_1}(\mathcal{A})\|_p.$$

To our knowledge, this is the first result to provide a full picture of the norm landscape over all possible tensor unfoldings. Using the partition lattice, we are able to quantify the impact of tensor unfoldings on the operator norms of the resulting tensors. The norm inequalities allow us to compare different unfolding schemes and evaluate the worst-case theoretical bounds of the corresponding algorithms. A open question remains to characterize the degree to which operator norm relations on the partition lattice restrict the original tensor. Essentially, this is a converse problem asking the conditions under which the spectral norm remains invariant within specific subsets of unfolding operations.

2.3 Two extremes

Tensors arisen in applications are often specially-structured. We will study the spectral properties of these tensors. The approach is novel because it explores the connection between high-dimensional statistics, multilinear algebra, and random tensor theory.

Corollary 2.3 (Tensor and its unfoldings, W. 2018+).

$$\frac{1}{d^{(K-\ell)/2}} \max_{\pi \in \mathcal{P}_{[K]}^{\ell}} \|\text{Unfold}_{\pi}(\mathcal{A})\|_2 \leq \|\mathcal{A}\|_2 \leq \min_{\pi \in \mathcal{P}_{[K]}^{\ell}} \|\text{Unfold}_{\pi}(\mathcal{A})\|_2. \quad (2)$$

Corollary 2 implies that the spectral norm of a tensor is lower and upper bounded by that of its unfoldings. Here $\pi \in \mathcal{P}_{[K]}^{\ell}$ denotes the partition with ℓ blocks; that is, $\text{Unfold}_{\pi}(\mathcal{A})$ is an unfolded tensor of order ℓ , where $1 \leq \ell \leq K$. In particular, $\ell = 2$ corresponds to matricization.

As we can see, an unfolding operation may inflate the operator norm by up to a $\text{poly}(d)$ factor in the worst case. However, the result should be interpreted with caution because the comparison bounds deal with arbitrary tensors rather than those often sought in applications. Tensors arisen in applications are often specially-structured. This motivate us to investigate the following two extreme cases.

Example 2.4. The class of orthogonally decomposable (ODE) tensors

$$\mathcal{A} = \lambda_1 \mathbf{a}_1^{(1)} \otimes \cdots \otimes \mathbf{a}_K^{(1)} + \cdots + \lambda_R \mathbf{a}_1^{(R)} \otimes \cdots \otimes \mathbf{a}_K^{(R)},$$

obtains the upper bound of (2), where $\{\mathbf{a}_1^{(r)}\}_{r \in [R]}, \dots, \{\mathbf{a}_K^{(r)}\}_{r \in [R]}$ are orthonormal vectors.

My ongoing work shows that ODE tensors belong to a more general class of structured tensors which I coin “partially-ODE” tensors. The spectral norm remains invariant for partially-ODE tensors under specific subsets of unfolding operations. Figure 3 gives an example of an order-6 partial-ODE tensor with respect to the partition $\pi = [12|34|56]$. The gray area is norm-preserving, meaning that all unfoldings in this region, including certain matricizations, leave the spectral norm unaffected.

Example 2.5. The class of Gaussian random tensors is defined as

$$\mathcal{A} = \llbracket a_{i_1 \dots i_K} \rrbracket \sim \mathcal{N}(\mathbf{0}, \Phi_1 \otimes \cdots \otimes \Phi_K),$$

if $\text{vec}(\mathcal{A})$ has zero mean and covariance $\Sigma = \Phi_1 \otimes \cdots \otimes \Phi_K$. In particular, when Φ_k are non-degenerate for all $k \in [K]$, we can define $\mathcal{C} = \mathcal{A}(\Phi_1^{-1/2}, \dots, \Phi_K^{-1/2}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I} \otimes \cdots \otimes \mathbf{I})$ which is a random tensor containing i.i.d. standard Gaussian entries. In such a case, \mathcal{C} is called to belong to the standard Gaussian tensor ensemble, and we are able to show that \mathcal{C} obtains the lower bound (up to a multiplicative factor) of (2) with high probability.

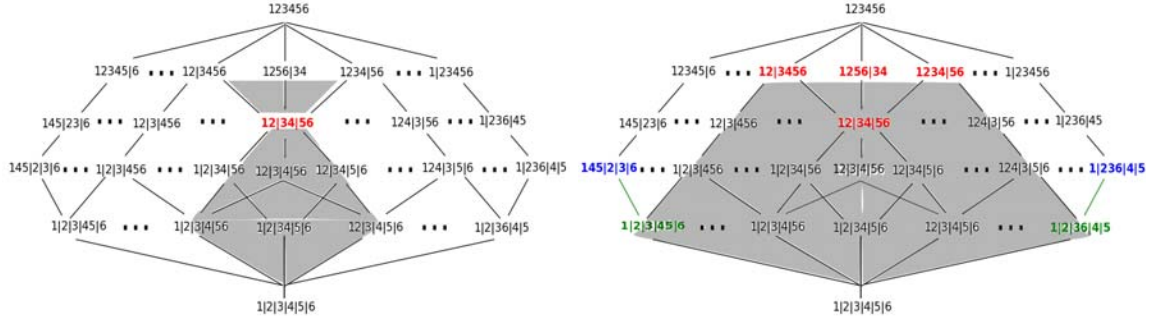


Figure 3: Norm-preserving region for a partially-ODE tensor. The tensor norm remains unchanged in the gray region. The norm comparison bounds between the two partitions (in blue) outside the region can be obtained by relating to the two partitions (in green) inside the region.

From the previous two examples, we see that the spectral norm of a normal tensor is inflated by up to a $\text{poly}(d)$ factor, whereas the structured tensors permit much tighter bounds. In statistical modeling, we are often in the intermediate scenario in which the signal is well approximated by a specially-structured tensor and the noise can be modeled using a Gaussian random tensor. The above results thus provide the theoretical foundations of novel tensor-based spectral methods.

2.4 Spectral properties for Gaussian tensor ensemble

Tensor data arising from genomics and neuroimaging applications are often extremely large and unwieldy. Fortunately, recent advances in randomized matrix techniques have brought a highly successful paradigm to scientific computation: finding structure with randomness [Halko et al., 2011]. Using our results on spectral relationship, we aim to understand how randomized methods interact with classical techniques and to develop efficient randomized tensor algorithms. Characterizing the behavior of spectral norms of random tensors is one of my specific aims as it is of great importance in many statistical and machine learning problems. As a preliminary result, I show that the spectral norm of a Gaussian random tensor satisfies the following non-asymptotic bound.

Theorem 2.6 (Concentration inequality, Unpublished). *Let $\mathcal{A} \in (\mathbb{R}^d)^{\otimes K}$ be a random tensor in the standard Gaussian tensor ensemble defined as in Example 2.5. Then we have*

$$d^{1/2} < \mathbb{E}(\|\mathcal{T}\|_2) < Kd^{1/2}. \quad (3)$$

Furthermore, $\|\mathcal{T}\|_2$ concentrates tightly around its expectation; namely, for any $s > 0$,

$$\mathbb{P}(|\|\mathcal{T}\|_2 - \mathbb{E}(\|\mathcal{T}\|_2)| \geq s) \leq 2e^{-s^2/2}.$$

Compared to earlier result [Tomioka and Suzuki, 2014, Nguyen et al., 2015], our inequality (3) holds for *every* dimension d and every order K . The bound implies $\|\mathcal{T}\| \asymp O(\sqrt{d})$ asymptotically for large d with fixed K . This can be viewed as a nice generalization of the classical result in random matrix theory. Since random tensor theory is far from reaching the level of maturity of random matrix theory, further forays in this direction are desirable.

Breaking previous limits: Spectral norm is one aspect of tensor spectral properties. We aim to investigate a range of (deterministic or probabilistic) spectral properties. For example, what is the tensor analogy of the Tracy-Widom law? Is there a Bernstein-type inequality for series of random tensors?

3 AIM 2: Estimation of low-rank tensors from discrete observations

This section proposes low-rank tensor estimation from contaminated sign observations. The estimation problem has never been studied in this context. Our preliminary results demonstrate a “dithering” effect in which stochastic noise is essential for recovering the signal from binary observations. This is in contrast to many statistical problems involving continuous observations.

We consider the problem of low-rank tensor estimation based on binary measurements. Such tensor data arise in several applications such as recommender system [Sun et al., 2017], social networks [Hoff, 2015, Rai et al., 2015], sensor network localization [Bhaskar and Javanmard, 2015], and neuroimaging [Zhou et al., 2013, Wang et al., 2017a]. One example is the recommender system [Bi et al., 2018], which can be naturally described as a three-way tensor of user \times item \times context and each entry indicates the user-item interaction. Another example is the DBLP database [Zhe et al., 2016] (see Figure 4), which is organized into a three-way tensor of author \times word \times venue and each entry indicates the co-occurrence of the triplets.

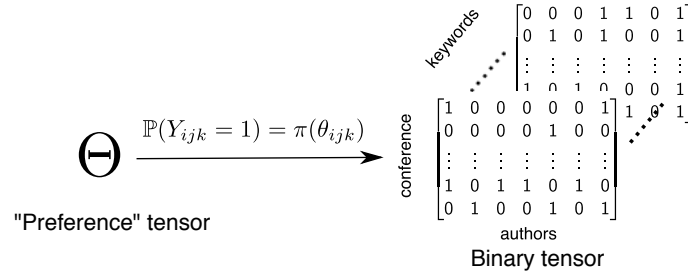


Figure 4: Latent variable interpretation of a binary tensor based on DBLP database.

Binary tensors encountered in practice are often noisy and high-dimensional, and it is thus crucial to develop dimension reduction methods for visualizing, summarizing and analyzing these datasets. A number of successful tensor decomposition methods have been proposed [Kolda and Bader, 2009,

Anandkumar et al., 2014, Wang and Song, 2017, Zhang and Xia, 2018], revitalizing the classical methods such as CANDECOMP/PARAFAC (CP) decomposition [Hitchcock, 1927] and Tucker decomposition [Tucker, 1966], as well as developing new ones such as tensor train decomposition [Osledeets, 2011]. However, all these methods treat tensor entries as continuous-valued, and are therefore not suitable to handle binary tensors. There is a relative paucity of binary tensor decomposition methods.

In this project, the PI aims to fill this gap by developing methodology and theory for binary tensor decomposition. Instead of assuming the data tensor is approximately low-rank, we propose to assume the parameter tensor, under a suitable link function, resides in a low-rank space. This generalization is analogous to the relationship between Gaussian linear model and generalized linear model (GLM). We draw inspiration from 1-bit compression and show that the generative model for a binary tensor can be regarded as the entrywise quantization from a noisy real-valued tensor. Such connection allows to fully characterize the impact of signal-to-noise ratio (SNR) on the recovery accuracy, as our preliminary results identify three different phases for tensor recovery according to SNR; see Table 1 in Section 3.3. When SNR is bounded by a constant, the loss in binary tensor decomposition is comparable to the case of continuous-valued tensor, suggesting very little information is lost by quantization. On the other hand, when SNR is sufficiently large, stochastic noise turns out to be helpful, and is in fact essential, for estimating the low-rank signal tensor.

3.1 Low-rank binary tensor model

This section exploits the connection between continuous-valued tensors and binary tensor decomposition. The proposed binary tensor model can be regarded as an entrywise quantization of a noisy continuous-valued low-rank tensor.

Consider an order- K tensor $\Theta = \llbracket \theta_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$ with rank bounded by R . Suppose that we cannot not directly observe Θ ; instead, we observe the quantized version $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \{0, 1\}^{d_1 \times \dots \times d_K}$ in the following way:

$$y_{i_1, \dots, i_K} = \begin{cases} 1 & \text{if } \theta_{i_1, \dots, i_K} + \varepsilon_{i_1, \dots, i_K} \geq 0, \\ 0 & \text{if } \theta_{i_1, \dots, i_K} + \varepsilon_{i_1, \dots, i_K} < 0, \end{cases} \quad (4)$$

where $\mathcal{E} = \llbracket \varepsilon_{i_1, \dots, i_K} \rrbracket$ is a noise tensor with i.i.d. entries following a distribution specified later.

In this setting, the tensor Θ represents an underlying, real-valued “preference tensor” whose noisy discretization gives \mathcal{Y} (Figure 4). The model (4) in fact is equivalent to the a generalized multilinear model if the cumulative distribution function of $\varepsilon_{i_1, \dots, i_K}$ is specified by a strictly increasing link function $f: \mathbb{R} \mapsto [0, 1]$:

$$\mathbb{P}(y_{i_1, \dots, i_K} = 1) = f(\theta_{i_1, \dots, i_K}). \quad (5)$$

We consider three choices of f , or equivalently, the distribution of $\mathcal{E} = \llbracket \varepsilon_{i_1, \dots, i_K} \rrbracket$.

Example 3.1. (Logistic link/Logistic noise). The logistic model is represented by (6) with $f(\theta) = (1 + e^{-\theta/\sigma})^{-1}$ and the scale parameter $\sigma > 0$. Equivalently, the noise $\varepsilon_{i_1, \dots, i_K}$ in (4) follows i.i.d. logistic distribution with the scale parameter σ .

Example 3.2. (Probit link/Gaussian noise). The probit model is represented by (6) with $f(\theta) = \Phi(\theta/\sigma)$, where Φ is the cumulative distribution function of a standard Gaussian. Equivalently, the noise $\varepsilon_{i_1, \dots, i_K}$ in (4) follows i.i.d. $N(0, \sigma)$.

Example 3.3. (Laplacian link/Laplacian noise). The Laplacian model is represented by (6) with

$$f(\theta) = \begin{cases} \frac{1}{2} \exp\left(\frac{\theta}{\sigma}\right), & \text{if } \theta < 0, \\ 1 - \frac{1}{2} \exp\left(-\frac{\theta}{\sigma}\right), & \text{if } \theta \geq 0, \end{cases}$$

and the scale parameter $\sigma > 0$. Equivalently, the noise $\varepsilon_{i_1, \dots, i_K}$ in (4) follows i.i.d. Laplace distribution with the scale parameter σ .

3.2 Rank-constrained likelihood-based estimation

This section proposes a rank-constrained maximum likelihood estimator (MLE) for the latent tensor Θ . The optimization is not jointly convex, but it is convex in each factor matrix individually with all other factor matrices fixed. This feature enables a block relaxation type minimization.

The new method is built in the framework of multilinear generalized model. For the binary tensor $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \{0, 1\}^{d_1 \times \dots \times d_K}$, we assume its entries are realizations of independent Bernoulli random variables, in that

$$\mathcal{Y} | \Theta \sim \text{Bernoulli}\{f(\Theta)\}, \quad \text{with } \mathbb{P}(y_{i_1, \dots, i_K} = 1) = f(\theta_{i_1, \dots, i_K}), \quad (6)$$

for all $(i_1, \dots, i_K) \in [d_1] \times \dots \times [d_K]$. Furthermore, we assume the parameter tensor Θ admits a rank- R CP decomposition,

$$\Theta = \sum_{r=1}^R \lambda_r \mathbf{a}_r^{(1)} \otimes \dots \otimes \mathbf{a}_r^{(K)}, \quad (7)$$

where $\lambda_r \in \mathbb{R}_+$, with $\lambda_1 \geq \dots \geq \lambda_K$ without loss of generality, $\mathbf{a}_r^{(k)} \in \mathbf{S}^{d_k-1}$, for $r \in [R], k \in [K]$, and Θ cannot be written as a sum of fewer than R outer products. In this model, $f: \mathbb{R} \rightarrow [0, 1]$ is a strictly increasing function.

We propose to estimate the unknown parameter tensor Θ in model (6) using a constrained likelihood approach. By incorporating the CP structure (7), the constrained MLE can be expressed as:

$$\hat{\Theta}_{\text{MLE}} = \min_{\Theta \in \mathcal{D}} \mathcal{L}_{\mathcal{Y}}(\Theta) = - \sum_{i_1, \dots, i_K} \log f(q_{i_1, \dots, i_K} \theta_{i_1, \dots, i_K})$$

where $\mathcal{D} = \{\Theta \text{ satisfies (7) with rank } R, \text{ and } \|\Theta\|_{\infty} \leq \alpha\},$

for a given rank $R \in \mathbb{N}_+$ and a bound $\alpha \in \mathbb{R}_+$. Here the feasible set \mathcal{D} consists of tensors defined by two constraints. The first is that Θ admits the CP structure (7) with rank R . As discussed in Section 1, the low-rank structure (7) is a commonly used dimension reduction tool in tensor data analysis. Moreover, all the subsequent estimation and theoretical results hold as long as the working rank is no smaller than the true rank. The second constraint is that all the entries of Θ are bounded in absolute value by a constant $\alpha \in \mathbb{R}_+$. We refer to α as the “signal” bound of Θ . This infinity-norm condition is a technical assumption to aid the recovery of Θ in the noiseless case. Similar conditions have been employed for the matrix case [Davenport et al., 2014, Bhaskar and Javanmard, 2015, Cai and Zhou, 2013].

3.3 Preliminary results

This section studies the convergence property of $\hat{\Theta}_{MLE}$. The classical asymptotic analysis does not apply straightforwardly because the number of parameters increases with the sample size. The proposed approach shows that the proposed estimate enjoys the minimax convergence rate as tensor dimensions increase.

The proposed approach will focus on the estimation accuracy using the deviation in the Frobenius norm. For the true coefficient tensor $\Theta^{true} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ and its estimator $\hat{\Theta}$, define

$$\text{Loss}(\hat{\Theta}, \Theta^{true}) = \frac{1}{\sqrt{\prod_k d_k}} \|\hat{\Theta} - \Theta^{true}\|_F.$$

Theorem 3.4 (Spectral norm, Unpublished) *Suppose $\mathcal{Y} \in \{0, 1\}^{d_1 \times \dots \times d_K}$ is an order- K binary tensor following model (6), with the link function f , and the true coefficient tensor $\Theta^{true} \in \mathcal{D}$. Then there exist an absolute constant $C_1 > 0$, and a constant $C_2 > 0$ that depends only on K , such that, with probability at least $1 - \exp(-C_1 \log K \sum_k d_k)$,*

$$\text{Loss}(\hat{\Theta}_{MLE}, \Theta^{true}) \leq \min \left\{ 2\alpha, \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}} \right\}. \quad (10)$$

Here L_α, γ_α are two quantities controlling the “steepness” and “convexity” of the link function f . Specifically,

$$L_\alpha \stackrel{\text{def}}{=} \sup_{|\theta| \leq \alpha} \left\{ \frac{\dot{f}(\theta)}{f(\theta)(1-f(\theta))} \right\}, \quad \text{and} \quad \gamma_\alpha \stackrel{\text{def}}{=} \inf_{|\theta| \leq \alpha} \left(\frac{\dot{f}^2(\theta)}{f^2(\theta)} - \frac{\ddot{f}(\theta)}{f(\theta)} \right),$$

where $\dot{f}(\theta) = df(\theta)/d\theta$, and α is the bound on the entry-wise infinity-norm of Θ . When α is a fixed constant and f is a fixed function, all these quantities are bounded by some fixed constants independent of the tensor dimension.

The next theorem establishes the information-theoretical lower bound for any estimator $\hat{\Theta}$ in $\mathcal{D}(R, \alpha)$ under model (6).

Theorem 3.5 (Minimax bound, Unpublished) *Consider a binary tensor $\mathcal{Y} \in \{0, 1\}^{d_1 \times \dots \times d_K}$ generated according to the probit model (3). Suppose $R \leq \min_k d_k$ and $\max_k d_k \geq 8$. Let $\hat{\Theta}$ denote an estimator of Θ^{true} . There exist absolute constants $\beta_0 \in (0, 1)$ and $c_0 > 0$ such that*

$$\inf_{\hat{\Theta}} \sup_{\Theta^{true} \in \mathcal{D}} \mathbb{P} \left(\text{Loss}(\hat{\Theta}, \Theta^{true}) \geq c_0 \min \left\{ \alpha, \sigma \sqrt{\frac{R d_{\max}}{\prod_k d_k}} \right\} \right) \geq \beta_0. \quad (11)$$

We now compare the lower bound (11) to the upper bound (10), as the tensor dimension $d_k \rightarrow \infty$ while the signal bound α and the noise level σ are fixed. Since $d_{\max} \leq \sum_k d_k \leq K d_{\max}$, both the bounds are of the form $C \sqrt{d_{\max}} (\prod_k d_k)^{-1/2}$, where C is a factor that does not depend on the tensor dimension. Henceforth, the estimator $\hat{\Theta}_{MLE}$ is rate-optimal in dimension.

Open Problem. An important question is whether the estimator is also rate-optimal with respect to the rank R . It remains an open problem to find a minimax convex solver. A relevant open problem also arises in the context of tensor completion [Raskutti et al., 2015], even for real-valued tensors.

Table 1 summarizes the error bound for binary tensor decomposition when $d_1 = \dots = d_K = d$ under probit link. When the signal and noise are not treated as fixed, our Theorems 3.4 and 3.5 reveal three different regimes of $\text{SNR} = \|\Theta\|_\infty/\sigma$:

		$\text{SNR} \gg \mathcal{O}(1)$	$\mathcal{O}(1) \gtrsim \text{SNR} \gg \mathcal{O}(d^{-(K-1)/2})$	$\mathcal{O}(d^{-(K-1)/2}) \gtrsim \text{SNR}$
Binary tensor	UB	$\sigma e^{\alpha^2/\alpha^2} d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α
	LB	$\sigma d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α
Continuous-valued tensor	UB	$\sigma d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α
	LB	$\sigma d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α

Table 1: Error bound for low-rank tensor estimation with probit link. For ease of presentation, we omit the constants that depend on K and R . UB: upper bound; LB: lower bound.

- 1). In the high SNR regime $\text{SNR} \gg \mathcal{O}(1)$, the error decreases with noise, suggesting increasing the noise level would lead to an improved tensor estimation accuracy (“dithering” effect);
- 2). In the moderate SNR regime $\mathcal{O}(1) \gtrsim \text{SNR} \gg \mathcal{O}(d^{-(K-1)/2})$, the error increases linearly with noise. In this case, the loss in binary tensor decomposition is comparable to the case of continuous-valued tensor, suggesting very little information is lost by discretization;
- 3). In the weak SNR regime $\text{SNR} \lesssim \mathcal{O}(d^{-(K-1)/2})$, a consistent estimation of Θ is impossible.

We conduct a preliminary study by simulating order-3 binary tensors from various models. Figure 5a plots the estimation error as a function of the tensor dimension d while holding the noise level fixed for three different ranks. Consistent with our theoretical results, the decay in the error appears to behave on the order of $d^{-1/2}$. Figure 5b plots the estimation error as a function of the noise level σ while holding the dimension fixed for three different ranks $R \in \{1, 3, 5\}$. A larger estimation error is observed when the noise is either too small or too large. The non-monotonic behavior confirms the changes in the three phases with respect to the SNR. Particularly, in the high SNR regime, the random noise is seen to improve the recovery accuracy. Finally, we comment that, although our approach is based on low-rank CP model, our method is robust to potential model misspecification. In particular, Table 2 reports the performance of our method for three-way block models, which is different from the multiplicative model in (6). It is seen that our method is able to recover the signal tensors well in all three scenarios.

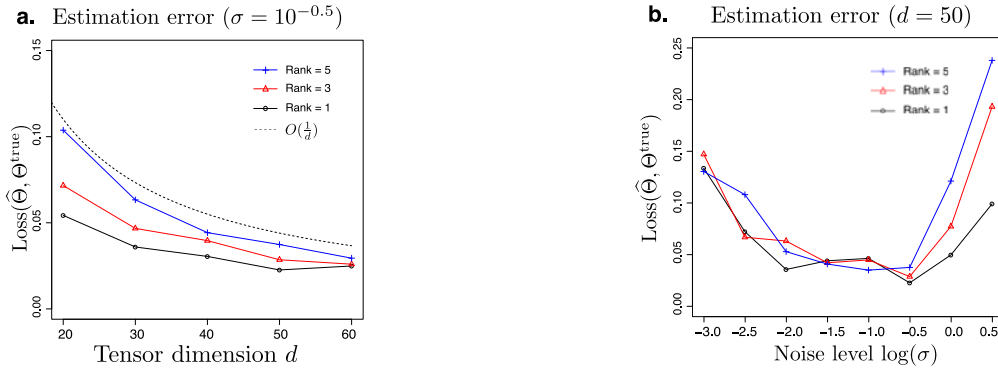
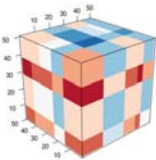
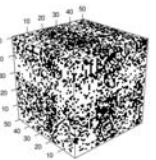
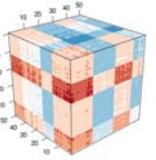
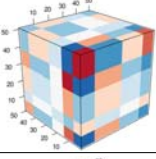
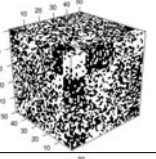
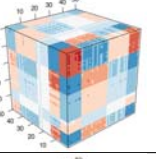
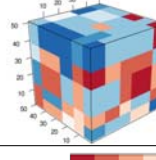
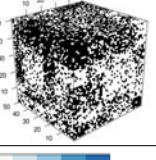
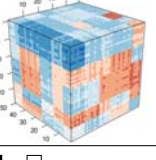
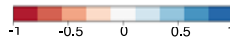



Figure 5: Estimation error of binary tensor decomposition. (a) Estimation error as a function of the tensor dimension $d = d_1 = d_2 = d_3$. (b) Estimation error as a function of the noise level.

Block model	Experiment			Relative Loss	Rank Estimate	Time (sec)
	True signal	Input tensor	Output tensor			
Additive				0.23(0.05)	1.9(0.3)	4.23(1.62)
Multiplicative				0.22(0.07)	1.0(0.0)	1.70(0.09)
Combinatorial				0.48(0.04)	6.0(0.9)	10.4(3.4)

Continuous-valued tensor
Binary-valued tensor

Table 2: Performance under the three-way block-mean models. Columns 2–4 are color images of the simulated tensor under different block mean models. Reported are the relative loss $\|\hat{\Theta}_{MLE} - \Theta^{true}\|_F / \|\Theta^{true}\|_F$, the estimated rank via BIC, and the running time, averaged over 30 data replications, with the standard error shown in the parenthesis.

4 Aim 3: Joint estimation of mean and covariance for tensor-valued data

This section proposes joint mean and covariance modeling for tensor data. Many genomics and neuroimaging datasets exhibit dependencies along multiple modes. This gives rise to the challenging problem of analyzing unreplicated high-dimensional data with unknown mean and dependence structures.

4.1 Background and outline

The research is motivated by the biological data described in Section 1. In the GTEx data, variation in expression levels results from complex interactions among genes, individuals, and tissues. This is illustrated in Table 3, where a group of genes may 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	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 3: Three-way clustering analysis of gene expression in the brain [Wang et al., 2018a]. Bold numbers indicate p -value < 0.001 .

Clustering has proven useful to reveal latent structure in high-dimensional expression data [Tibshirani et al., 1999, Lazzeroni and Owen, 2002, Liu et al., 2008, Tan and Witten, 2014]. Traditional

clustering methods (such as K-means, PCA, and t-SNE [Maaten and Hinton, 2008]) assume that gene expression patterns persist across one of the different contexts (either tissues or individuals), or assume that samples are i.i.d. or homogeneous. Direct application of these algorithms to multi-tissue expression data requires concatenating all available samples from different tissues into a single matrix, precluding potential insights into tissue \times individual specificity [Bahcall, 2015]. Alternatively, inferring gene modules separately for each tissue ignores commonalities among tissues and may hinder the discovery of differentially-expressed genes that characterize tissues or tissue groups.

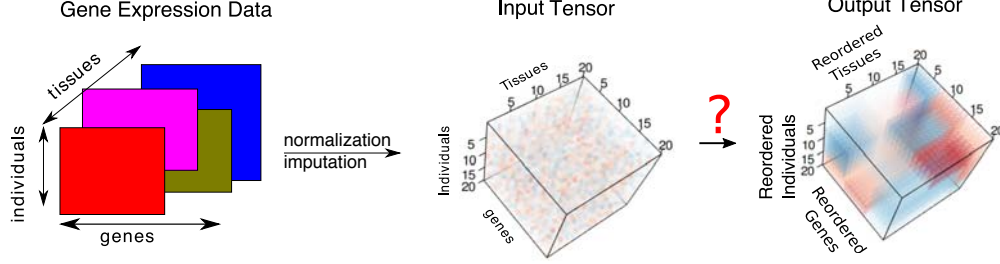


Figure 6: Schematic figure for robust multi-way clustering.

Robust Multi-way Clustering. To identify subsets of genes that are similarly expressed within subsets of individuals and tissues, we seek local blocks in the expression tensor while accounting for background correlation.

As illustrated in Figure 6, multi-tissue multi-individual gene expression measurements can be organized into a three-way array, or order-3 tensor, with gene, tissue, and individual modes. Our recent work [Wang et al., 2018a] developed a CP tensor method to simultaneously cluster genes, tissues, and individuals. The model assumes the data entries are independent conditionally on the group memberships. However, violations of the assumption is possible in practice. Data with background correlations are common in a broad range of settings, including neuroscience (spatial and temporal correlations) and genomics (latent correlation due to batch effects). The presence of background correlations may affect the accuracy on cluster inferences. Our goal is to robustly infer subsets of genes that are similarly expressed in subsets of tissues and individuals; mathematically, this reduces to detecting three-way blocks in the expression tensor while accounting for background correlation (Figure 6).

4.2 Multi-way block model with kronecker covariance

Let $\mathcal{Y} = \llbracket Y_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be an order- K data tensor. The tensor \mathcal{Y} is said to admit a multi-way block structure if there exists a checkbox structure module some unknown reordering along each of its mode. Specifically, the tensor mean admits the following Tucker structure:

$$\mathbb{E}(\mathcal{Y}) \equiv \mathcal{S} = \mathcal{C} \times_1 \mathbf{M}_1 \cdots \times_K \mathbf{M}_K, \quad (12)$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \dots \times r_K}$ represents the block mean tensor and $\mathbf{M}_k \in \{0, 1\}^{r_k \times d_k}$ is the group membership matrix at k -th mode, for all $k \in [K]$. By the definition of multilinear rank, $\mathbb{E}(\mathcal{Y})$ has multilinear-rank bounded by (r_1, \dots, r_K) , which is typically much smaller than (d_1, \dots, d_K) .

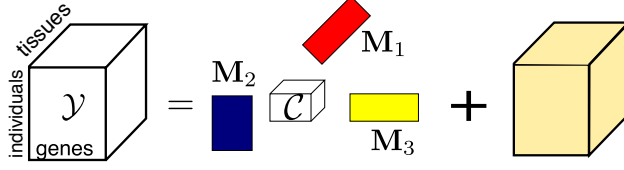


Figure 7: Schematic figure of a multi-way block model for an order-3 tensor.

What distinguishes our approach from those commonly used in the literature is that we do not impose the independence assumptions on the tensor entries. Instead, we introduce a more general tensor-variate distribution (c.f. Example 2.5) on the \mathcal{Y} :

$$\mathcal{Y} \sim \mathcal{N}(\mathcal{S}, \Phi_1 \otimes \cdots \otimes \Phi_K), \quad (13)$$

where $\Phi_k \in \mathbb{R}^{d_k \times d_k}$ represents the covariance along the k -th mode for all $k \in [K]$.

A further simplification to the model (13) is natural. Though we might expect correlation present among all entries, correlations between entries in two different clusters are less easily interpreted. This leads to a block-diagonal model for each of $\{\Phi_k\}$:

$$\Phi_k = \begin{bmatrix} \Phi_{k,1} & & \\ & \ddots & \\ & & \Phi_{k,r_k} \end{bmatrix}, \quad \text{where } k = 1, \dots, K \quad (14)$$

where $\Phi_{k,\ell}$ is the covariance matrix for entries within ℓ -th cluster in the mode k , where $\ell = 1, \dots, r_k$. Combining equations (12), (13) and (14) yields our final model. The parameters of interest are \mathcal{C} and $\{\mathbf{M}_k\}_{k \in [K]}$, whereas the nuisance parameters are $\{\Phi_{k,\ell}\}_{\ell \in [r_k], k \in [K]}$. Note that the covariance matrices are identifiably only up to scaling.

4.3 Further considerations

The joint modeling approach above will enable accurate estimation of cluster structure. To enhance the interpretability, we will consider various forms of penalties including (inverse) covariance sparsity. One possibility is to impose a graphical lasso penalty on the $\{\Phi_k\}$ and an l_1 -penalty on the \mathcal{S} . This leads to the following optimization:

$$f(\mathcal{C}, \{\mathbf{M}_k\}, \{\Phi_{k,\ell}\}, \lambda, \beta) = \text{Log-like}(\mathcal{C}, \{\mathbf{M}_k\}, \{\Phi_{k,\ell}\}) - \lambda \sum_{k=1}^K \sum_{\ell=1}^{r_k} |\Phi_{k,\ell}^{-1}|_1 - \beta |\mathcal{C}|_1, \quad (15)$$

where $|\cdot|$ denotes the entrywise l_1 norm and λ, β are nonnegative turning parameters that determine the extent of penalization. To maximize (15), we will take an iterative approach in which we update each of the parameters sequentially, holding all other parameters fixed as we update the current set of parameters. We will explore using the state-of-the-art graphical lasso and k -mean algorithm and modify them to incorporate more realistic features.

Our proposed joint modeling has the following advantages: (1) inferences on the block structure become correctly calibrated, and (2) the graphical networks along each modes are able to be inferred. An important question in this setting is to select the tuning parameters λ and β . We will take a cross-validation approach by casting the clustering problem into a supervised prediction problem. Specially, one can leave out a random subset of elements from the data tensor \mathcal{Y} , impute those left-out elements using the overall mean, and cluster the resulting tensor data. We will use the predicted mean squared loss as the utility.

Statistical inference based on tensor factors can be further extended. One possible approach would be performing parametric bootstrap [Efron and Tibshirani, 1994] to assess the uncertainty in the estimation. For example, one can simulate tensors from the fitted low-rank model (13) based on the estimates, and then assess the empirical distribution of the block estimates. This approach has been applied in matrix factorization [Milan and Whittaker, 1995] and can be extended to tensor factorization. While being simple, the parametric bootstrap assesses uncertainty only in the estimation procedure but not the modeling. The inclusion of lasso penalty also adds further complication to the post-selection inference. We leave it for future study.

5 Borader impacts

The educational components of this proposal are targeted at several different populations of students: PhD students, undergraduate students, visiting international students, AP statistics students underrepresented students at West High School in Madison. The project will support the development of the PI, a young faculty, to create an interdisciplinary working group of researchers.

5.1 Undergraduate education

The PI is one of the few female faculty members in her home department. As a young faculty, PI has been striving to encourage more under-represented students into her research field. The personnel constitution in our education environment is becoming increasingly diverse, reflecting the variety witnessed in our broader society. Diversity can be measured across many dimensions – gender, ethnicity, origin, age, etc. However, diversity is not limited to these visible physical differences; it also consists of difference in each student’s personality (introvert vs. extravert), learning skills (listening, speaking, reading, or writing skills), emotional development, and every aspect that distinguishes one individual from another. PI believes that every student is unique in his/her own way, and as professors we should create an open, inclusive education environment that prepares students to become better professionals in the larger, diverse society.

To achieve this goal, the PI has been actively involved in Madison Teaching and Learning Excellence program. The PI plans to open up an undergraduate course on *introduction to data sciences*. 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. Statistics, as a discipline, is concerned with transforming “data” into “information”. However *data are not just numbers, they are numbers with context*. Carefully-chosen examples are crucial for all levels of students, even from pure fields: they need to understand that statistics use the language of mathematics, but is *not* only about mathematics. Training our students to read critically, translate statistical results into plain English (or vice versa), and be aware of possibly spurious situations plays an important role in teaching statistics, both at the undergraduate and the graduate level.

5.2 Outreach activities

The PI has a unique combination of training backgrounds in statistics, computer science, and genomics. The PI’s established collaborations on cutting-edge scientific problems has motivated the methodological ideas presented in this proposal. Ongoing collaborations include Genomics lab at the University of Pennsylvania, Evolution and Ecology lab at the University of Chicago, Laboratoire des Interactions Plantes Micro-organismes (LIPM) at CNRS, France and data science group at Chan Zuckerberg Biohub. The resulting solutions are expected to have a direct impact on a broad range of scientific fields. The PI also plans to release software packages to include the new methods resulting from this proposal, as well as other popular methods in the area of tensor data analysis.

References

- Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15: 2773–2832, 2014.
- Orli G Bahcall. Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals. *Nature Reviews Genetics*, 16(7):375, 2015.
- Sonia A Bhaskar and Adel Javanmard. 1-bit matrix completion under exact low-rank constraint. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*, pages 1–6. IEEE, 2015.
- Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.*, 46(6B):3308–3333, 12 2018. doi: 10.1214/17-AOS1659.
- Marie-Aur lie Bruno, D Fern andez-Espejo, Remy Lehenbre, Luaba Tshibanda, Audrey Vanhau-denhuysse, Olivia Gosseries, Emilie Lommers, M Napolitani, Quentin Noirhomme, M lanie Boly, et al. Multimodal neuroimaging in patients with disorders of consciousness showing ?functional hemispherectomy? In *Progress in brain research*, volume 193, pages 323–333. Elsevier, 2011.
- T Tony Cai, Zongming Ma, Yihong Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.
- Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.
- GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.
- Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.
- Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.
- Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.
- Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169, 2015.
- Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.
- Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

- Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica sinica*, pages 61–86, 2002.
- Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. In *Computational Advances in Multi-Sensor Adaptive Processing, 2005 1st IEEE International Workshop on*, pages 129–132. IEEE, 2005.
- Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. *arXiv preprint arXiv:1702.07449*, 2017.
- Yufeng Liu, David Neil Hayes, Andrew Nobel, and JS Marron. Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
- Peter McCullagh. *Tensor methods in statistics*. Courier Dover Publications, 2018.
- Luis Milan and Joe Whittaker. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, pages 31–49, 1995.
- Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229, 2015.
- Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- Piyush Rai, Changwei Hu, Matthew Harding, and Lawrence Carin. Scalable probabilistic tensor factorization for binary and count data. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 3770–3776, 2015.
- Garvesh Raskutti, Ming Yuan, and Han Chen. Convex regularization for high-dimensional multi-response tensor regression. *arXiv preprint arXiv:1512.01215*, 2015.
- Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.
- Gilbert W Stewart. Matrix perturbation theory. 1990.
- Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017. doi: 10.1111/rssb.12190.
- Kean Ming Tan and Daniela M Witten. Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics*, 23(4):985–1008, 2014.
- Robert Tibshirani, Trevor Hastie, Mike Eisen, Doug Ross, David Botstein, Pat Brown, et al. Clustering methods for the analysis of DNA microarray data. *Dept. Statist., Stanford Univ., Stanford, CA, Tech. Rep*, 1999.
- Nicholas J Timpson, Celia MT Greenwood, Nicole Soranzo, Daniel J Lawson, and J Brent Richards. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, 19(2):110, 2018.

- Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.
- Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966.
- Santosh S Vempala and Ying Xiao. Max vs min: Tensor decomposition and ica with nearly linear sample complexity. In *Conference on Learning Theory*, pages 1710–1723, 2015.
- Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of multiple networks. *arXiv preprint arXiv:1707.06360*, 2017a.
- Miaoyan Wang and Yun Song. Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622, 2017.
- Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and its Applications*, 520:44–66, 2017b.
- Miaoyan Wang, Jonathan Fischer, and Yun S Song. Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition. *Accepted by Annals of Applied Statistics*, 2018a.
- Miaoyan Wang, Fabrice Roux, Claudia Bartoli, Carine Huard-Chauveau, Christopher Meyer, Hana Lee, Dominique Roby, Mary Sara McPeck, and Joy Bergelson. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences*, page 201710980, 2018b.
- Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.
- Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018.
- Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang-chih Lee, Zenglin Xu, Yuan Qi, and Zoubin Ghahramani. Distributed flexible nonlinear tensor factorization. In *Advances in Neural Information Processing Systems*, pages 928–936, 2016.
- Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.
- Jing Zhou, Anirban Bhattacharya, Amy H Herring, and David B Dunson. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512):1562–1576, 2015.