Research Statement (Miaoyan Wang)

My research is in the intersection of statistics, machine learning, and optimization. In particular, I am working on developing higher-order tensor methods with statistical and computational guarantees. Much of my work is directly motivated by data analytic problems in genetics and genomics. These problems often involve large-scale, high-dimensional data, for which novel statistical methods are required. Examples include data from genome-wide association studies (GWAS) and multi-tissue gene expression experiments. Through the development of new statistical methods for "big data", I hope to advance our ability to make scientific discoveries in data-intensive fields.

Below, I briefly describe highlights of ongoing work. I will also summarize some of my collaborations and outline my vision for future work.

Current Research and Extensions

Tensors of order 3 or greater, known as higher-order tensors, have recently attracted increased attention in many fields across science and engineering. Among numerous examples, tensors have been used to model higher-order cumulants and to detect patterns in multi-way array data. However, tensor-based methods are fraught with challenges. Tensors are not simply matrices with more indices; rather, they are mathematical objects possessing multilinear algebraic properties. To this end, I have established several spectral properties for tensors which provide the mathematical foundations of novel tensor-based computations.

Spectral relationships between higher-order tensors and their unfoldings. A common paradigm in tensor-related algorithms advocates unfolding a tensor into a matrix and applying classical methods developed for matrices. Despite the popularity of such techniques, how the functional properties of a tensor change upon unfolding is currently not well understood. In this work, we investigate 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. Given an order-k tensor, we represent each possible unfolding operation using a partition π of $\{1, \ldots, k\}$, where a block in π corresponds to the set of modes that should be combined into a single mode. Figure 1 illustrates an example for $k = 3$.

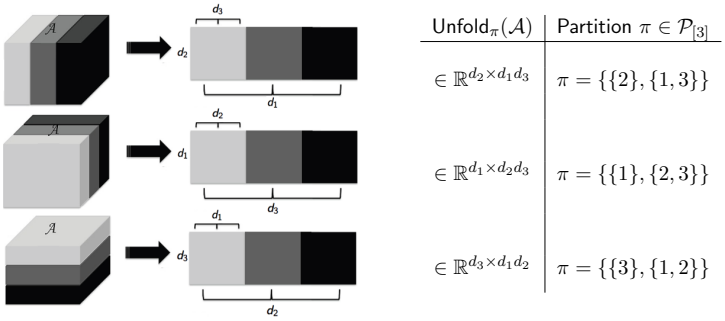


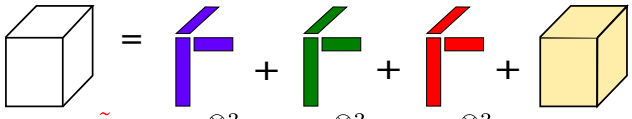
Figure 1: An order-3 tensor and its matricizations. The set of all possible unfoldings is in one-to-one correspondence with the set of partitions of $\{1, 2, 3\}$.

In contrast to the body of existing work which has focused almost exclusively on matricizations, I present a new framework representing all possible tensor unfoldings using the partition lattice and establish a set of general inequalities quantifying the impact of tensor unfoldings on the operator norms of the resulting tensors. I show that the comparison bounds scale polynomially in the dimensions $\{d_n\}$ of the tensor with powers depending on the corresponding partition and block sizes for the pair of tensor unfoldings being compared. The operator norms of all possible tensor

unfoldings together define what we coin a "norm landscape" on the partition lattice. To our knowledge, this result is the first attempt to provide a full picture of the norm landscape over all possible tensor unfoldings. This work is published in *Linear Algebra and Its Applications, 2017* [1].

I would like to emphasize that while this work focuses on theory rather than computational tractability, it possesses practical implications as well. The norm inequalities allow us to compare different unfolding schemes and evaluate the worst-case theoretical bounds of the corresponding algorithms. In addition to what is presented in [1], I have recently pushed the results further by characterizing 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. This is an on-going work with Anna Seigal (Math, UC Berkeley) and Yun S. Song (Stat & EECS, UC Berkeley).

Orthogonal tensor decomposition and its perturbation bound. Tensor decompositions can be used to estimate parameters in various latent variable models via the method of moments and also lead to the development of efficient denoising techniques in independent component analysis. In these scenarios, one is interested in decomposing a symmetric and (nearly) orthogonal decomposable (SOD [2]) tensors of the following form,



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

where $\{\mathbf{u}_i\}$ is a set of orthonormal vectors in $\mathbb{R}^d$, λ_is are positive scalars in $\mathbb{R}$, and the perturbation $\mathcal{E}$ is a symmetric but otherwise arbitrary tensor with $\|\mathcal{E}\|_2 \leq \varepsilon$. Our goal is to estimate the factor pairs $\{(\mathbf{u}_i, \lambda_i)\}$ from the noisy observation $\widetilde{\mathcal{A}}$.

In joint work with Yun S. Song, I present a new decomposition algorithm, TM-HOSVD (**T**wo-**m**ode **H**igher-**o**rder **S**ingular-**v**alue **D**ecomposition), and establish an analogue of Wedin's perturbation theorem for higher-order tensors:

Theorem 1 (Wang & Song, 2017)**.** *Let $\widetilde{\mathcal{A}} \in (\mathbb{R}^d)^{\otimes k}$ be a nearly-SOD tensor and $\{(\mathbf{u}_i, \lambda_i) \in \mathbb{R}^d \times \mathbb{R}\}$ be the output of the TM-HOSVD algorithm. Suppose $\varepsilon \leq \lambda_{\min}/[c_0 d^{(k-2)/2}]$, where c_0 is a sufficiently large constant that does not depend on d. Then, there exists a permutation π on $[r]$ such that for all $i \in [r]$,*

$$Loss(\widehat{\mathbf{u}}_i, \mathbf{u}_{\pi(i)}) \leq \frac{2\varepsilon}{\lambda_{\pi(i)}} + o(\varepsilon), \quad Loss(\widehat{\lambda}_i, \lambda_{\pi(i)}) \leq 2\varepsilon + o(\varepsilon). \tag{1}$$

Our contribution is to develop an effective decomposition algorithm, TM-HOSVD, and show that its error bound in (1) does not depend on the eigen-gaps (i.e., $|\lambda_i - \lambda_j|$). This feature fundamentally distinguishes tensor decompositions from matrix decompositions and reflects Kruskal's uniqueness characterization. The main idea in TM-HOSVD is to unfold the tensor along *two* modes and use rank-1 constraints to characterize the underlying components. Earlier work (Anandkumar et al, 2014) showed that the eigen-components of a 3rd cumulant tensor are closely related to parameter estimation in graphical models. In this case, we relax the perturbation tolerance from $\varepsilon \leq O(d^{-1})$ to $O(d^{-1/2})$, which is substantial. This work is published in *Proceedings of Machine Learning Research (Artificial Intelligence and Statistics, 2017)* [2].

There are many avenues worthy of pursuit based on this work. Recent works have shown that some non-orthogonal tensors can be converted to orthogonal tensors by an additional whitening

step. Therefore, the decomposition framework is applicable to a broad class of structured tensors. Motivated by applications, I have extended the TM-HOSVD algorithm to semi-nonnegative tensor decomposition and developed a novel tensor-based method for multi-mode clustering [3]. The proposed decomposition lends itself well to this context.

Machine learning and statistical applications to genetics. In addition to developing machine learning and statistical theory, I have a long-standing interest in developing statistical methods to address important problems in genetics and genomics. Here I briefly describe my current work on two-way mixed-effects methods for genetic association studies; other work in this vein can be found in [4–7].

GWAS are powerful tools for the identification of genetic variants underlying complex traits. However, existing methods typically perform GWAS within a single species; methods allowing the description of the genomic landscape of interactions between species have yet been developed. In joint work with Mary Sara McPeek (Statistics, UChicago), we develop the ATOMM method (for **A**nalysis with a **T**wo-**O**rganism **M**ixed **M**odel) to simultaneously detect genetic variants on a pair of genomes that are associated with a trait of interest.

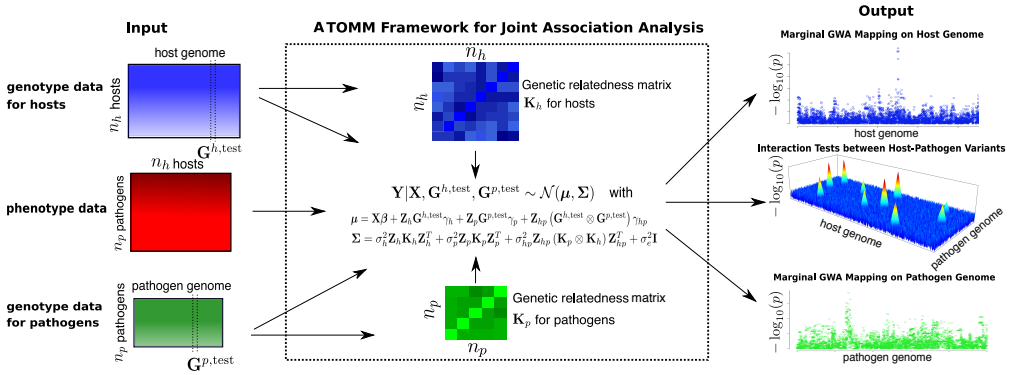


Figure 2: Illustration of the ATOMM for joint association analysis. In this schemaic, a conditional multivariate normal trait is assumed, but we have also developed a version of ATOMM for binomial-like traits.

Figure 2 illustrates the schematic diagram of ATOMM in a host-pathogen association study, ATOMM takes as input (i) phenotypic values obtained from host-pathogen pairs and (ii) whole-genome sequencing data for both host and pathogen samples, and outputs results from (i) marginal GWAS mapping on host and pathogen genomes and (ii) G $\times$ G interactions between host and pathogen variants.

The ATOMM method addresses the challenges of confounding due to host and pathogen population structure by utilizing a two-way, mixed effects model with three types of genetic relatedness matrices (GRMs): one each for the host and pathogen additive polygenic random effects individually, and the third for the additive-by-additive polygenic interaction random effects between the two genomes. The testing procedure is parallelizable in a very straightforward way, making our model computationally feasible for large-scale studies. Our method identifies candidate genes conferring host-pathogen specificity and uncovers genomic regions that are likely to carry evidence of co-evolution. As several consortium efforts have emerged to sequence the genomes of a variety of organisms, I believe that ATOMM provides a new opportunity to integrate available whole-genome sequence data and to decipher the genetic architecture of complex traits in finer detail than previously possible. This work is currently under the second-round review of *Proceedings of the National Academy of Sciences* (direct submission) [4].

Future directions

The prevailing theme in my research is to develop powerful statistical and machine learning theory and methods for advancing knowledges in data-intensive fields. In this regard, my work has spanned the spectrum from theoretical statistics to applied data analysis. I feel fortunate to have worked-closely with many fellow scientists on cutting-edge big data problems [4–6], which have shaped my scientific approach and spurred new ideas in my statistics research. Here I list two directions I would like to pursue: (1) nonparametric modeling of latent sample structure in GWAS, and (2) randomized algorithms for higher-order tensors.

Nonparametric modeling of latent sample structure in GWAS. In genetic association analyses, it is critical to increase power while avoiding false discoveries by properly controlling for latent sample structure. The linear mixed model (LMM) has emerged as a popular method to do so, however, the ability of LMMs to control for latent structure is based on relatively simple but restricted assumptions. To control for polygenic effects and population structure, LMMs assume that (1) the random effects from background variants are additive and i.i.d.; and (2) confounding effects for a pair of individuals are correlated according to the expected coalescence time between the two. Although both assumptions have been generally applicable in previous studies, the increasing sample size in modern GWAS is creating new challenges. With hundreds of thousands of sampled individuals available, many previously small genetic effects become detectable whereas previously small confounding effects become influential. I will develop nonparametric methods to allow fewer functional and distributional assumptions for the variant effect sizes. I will furthermore investigate into alternative to GRMs based on different probabilistic generative processes in order to better control for population confounding. The required high-dimensional integration will likely require stochastic variational approximations to scale up to the GWAS settings, but I expect this approach to outperform the current *ad hoc* greedy selection approach (Segura et al, 2012).

Randomized algorithms for higher-order tensors. Tensorial data arising from genetics and genomics 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. My goal here is to understand how randomized methods interact with classical techniques and to develop efficient randomized tensor algorithms with theoretical guarantees. Understanding the behavior of spectral norms of random tensors is one of my specific aims as it is a problem of considerable importance in many machine learning applications. As a preliminary result, I have obtained a non-asymptotic bound for the spectral norm of a Gaussian random tensor:

Theorem 2 (Wang, 2017)**.** *Let $\mathcal{T} \in (\mathbb{R}^d)^{\otimes k}$ be a random tensor with i.i.d. standard Gaussian entries, then we have*

$$d^{1/2} < \mathbb{E}(\|\mathcal{T}\|_2) < kd^{1/2}. \tag{2}$$

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

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

The proof uses stochastic monotonicity and Slepian-type inequalities. Compared to earlier result (Tomioka & Suzuki, 2014), the inequality (2) holds for *every* dimension d and every order k. The bound implies $\|\mathcal{T}\|_2 \asymp O(\sqrt{d})$ 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. For example,

Open Problem (Wang, 2017). Is there a tensor analogy of the Tracy-Widom law for the top eigenvalues of random symmetric tensors? Can we generalize the Bernstein inequality for the spectral norm of a sum of independent random tensors of the same dimensions?

References

[1] **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 (2017), 44–66.

[2] **Miaoyan Wang** and Yun S. Song, Tensor Decompositions via Two-Mode Higher-Order SVD (HOSVD), *Proceedings of Machine Learning Research* 54 (2017), 614–622.

[3] **Miaoyan Wang**, Jonathan Fischer, and Yun S Song, Three-way clustering of multi-tissue gene expression data using constrained tensor decompositions, *Preprint* (2017).

[4] **Miaoyan Wang**, Fabrice Rouxc, Claudia Bartolic, Carine Huard-Chauveauc, Christopher Meyere, Hana Lee, Dominique Roby, Mary Sara McPeek, and Joy Bergelson, Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes, Under second-round review by *Proceedings of the National Academy of Sciences (PNAS)* (2017).

[5] **Miaoyan Wang**, Johanna Jakobsdottir, Albert V Smith, and Mary Sara McPeek, G-strategy: Optimal selection of individuals for sequencing in genetic association studies, *Genetic Epidemiology* 40 (2016), no. 6, 446–460.

[6] Brett W Engelmann, Yohan Kim, **Miaoyan Wang**, Bjoern Peters, Ronald S Rock, and Piers D Nash, The development and application of a quantitative peptide microarray based approach to protein interaction domain specificity space, *Molecular & Cellular Proteomics* 13 (2014), no. 12, 3647–3662.

[7] Duo Jiang and **Miaoyan Wang**, Recent developments in statistical methods for GWAS and high-throughput sequencing studies of complex traits, Under review by *Biostatistics & Epidemiology* (2017).