WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Please Reply To:

Miaoyan Wang
Department of Statistics
University of Wisconsin–Madison
1300 University Ave
Madison, Wisconsin 53706

Email: miaoyan.wang@wisc.edu
http://pages.stat.wisc.edu/~miaoyan

July 3, 2020

Dear Committee Member,

I am writing to apply for the research funding from American Family Insurance. I am an assistant professor in the Department of Statistics at the University of Wisconsin-Madison. My research is in the intersection of mathematics, statistics, and computer science, with a focus on **signal processing and data fusion**. Specific interests include higher-order tensor methods, high dimensional statistics, and applications to network analyses and neuroimaging.

The prevailing theme in my proposal is to develop powerful machine learning methods and application for advancing knowledges in data fusion. Specifically, I will focus on developing statistical methods for high-dimensional high-order object data (a.k.a. tensors) with applications to social networks, brain imaging, and integrative analysis of multimodal data. Analyzing tensor data with increasing dimensionality and ever-growing complexity requires the development of novel machine learning tools. In this regard, my work will link theory and practice, and also expand core computational areas based on the questions raised in the applied endeavors.

My interdisciplinary research efforts have been reflected in my training. Prior to UW-Madison, I was a postdoc in Computer Science at UC Berkeley, where I was also affiliated with Chan-Zuckerberg Biohub at San Francisco. In 2015-2017, I was a Simons postdoc in Biology and Mathematics at University of Pennsylvania. I obtained my Ph.D. in Statistics from the University of Chicago in 2015 and B.S. in Mathematics from Fudan University

in 2010. I plan to leverage my interdisciplinary background to find new solutions to data problems arising from science and engineering.

This proposal lays out an ambitious plan aiming to aid decision-making in health industry via the development of efficient data fusion methods for high-dimensional tensors. The software packages resulting from this proposal, will be released freely, as well as related visualization tools for tensor data analyses. With the support of American Family Insurance program, I will organize a series of workshops that promote the applications of tensor data analysis in industry. As one of the few female faculty members in my home department, I have been striving to encourage more under-represented students into the emerging data science. I plan to create an undergraduate course on **introduction to data sciences** that inspire more students in learning data analytic tools to address important questions in daily life.

Enclosed please find my CV and research proposal. Thank you in advance for your consideration.

Sincerely,
Miaoyan Wang

Tensor Methods and Applications to Biomedical Challenges

Miaoyan Wang, Department of Statistics

I work at the intersection of **statistics**, **machine learning**, and **nuroimaging**, with a particular focus on higher-order tensor methods. My research is driven by data analytic problems in **signal processing and data fusion**. These problems often involve large-scale, high-dimensional data, for which novel statistical methods are required. Examples include data from multi-relational social networks and context-specific recommendation systems. **The proposed project is to develop a framework of statistical models, efficient algorithms, and mathematical theory to analyze large-scale tensor data.** Through the development of analytic tools for big tensor data, my research aims at increasing the efficiency in biomedical discoveries.

Tensor methods transform data to knowledge

Tensors are high-dimensional arrays (Figure 1a). Recent advances in high-throughput technology have transformed scientific research into data-intensive fields where data are naturally generated in tensor form. Figure 1b illustrates an example of tensor data arising from my current collaboration on multi-modal imaging analysis of brain connectivity. The Human Connectome Project provides a huge compendium of tensor data consisting of anatomical and functional connectivities within 1,200 human brains. Multiple imaging measurements are utilized to construct the brain networks, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and diffusion tensor imaging (DTI). Understanding key patterns among multimodal networks is crucial to unravel brain functions, thereby broadly facilitating research efforts towards personalized treatment for human diseases.

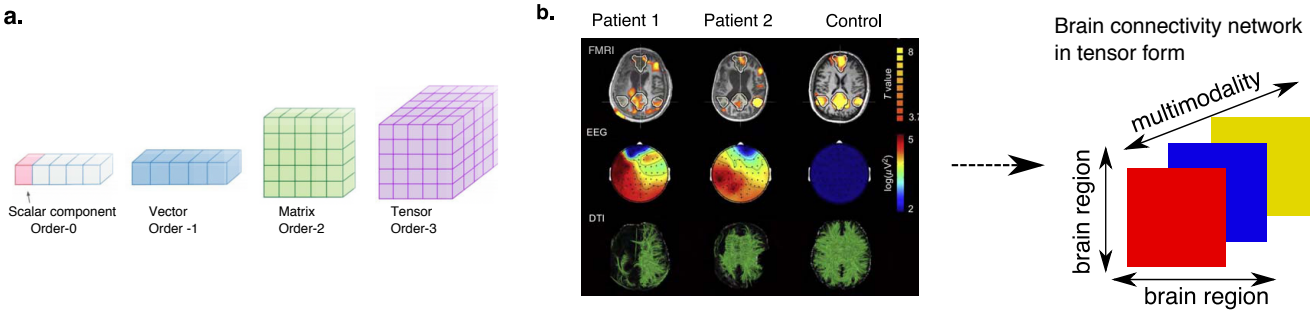


Figure 1: (a) Higher-order tensors are generalizations of matrices. (b) Tensor data from neuroimaging studies. Multiple measurements are used to construct brain connectivities within 1,200 individuals (Bruno et al., 2011).

Methods built on tensors provide generalized tools to capture complex data structure that the off-the-shelf methods may fail to exploit. In many data fusion applications, researcher are interested in interpretable low-dimensional structure within the high-dimensional tensor data. The question goes beyond the traditional multivariate analysis; we aim to characterize probabilistic distributions over higher-order "objects", where the objects can be images, networks, manifolds, or in general, tensors. Our project targets at developing theory, methodology, and practice for analyzing high-dimensional data. **The resulting prototypes will facilitate automatic detection of hidden salient structure in tensor data, thereby providing solutions to questions that cannot be addressed by classical methods.**

The proposal focuses on two main directions outlined below. The ultimate goal is to not only bridge data analytics to domain science, but also spark new areas where machine learning theory and scientific discovery can complement each other.

Project 1: Supervised learning with high-dimensional tensors. Prediction machine learning problems common arise in daily applications, i.e., identifying segmentations from images, classifying documents into topics, or deciding target customers in recommendation systems. These problems are often formulated as predicting a variable Y from explanatory variables X, where the data available is in the form of pairs $\{(X_i, Y_i) : i = 1, \ldots, n\}$. In contrast to traditional work that focuses on only univariate response, we will develop a general learning framework with tensor observation serving as a response, and features along multiple modes forming the predictor. The proposed model enables the prediction of a high-dimensional tensor $Y \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ from possibly multiple predictors $X^{(k)} \in \mathbb{R}^{d_k \times p_k}$, where $k = 1, \ldots, K$ indexes the modes of the tensor. The tensor approach boosts the prediction performance by incorporation of multi-way information across the different modes.

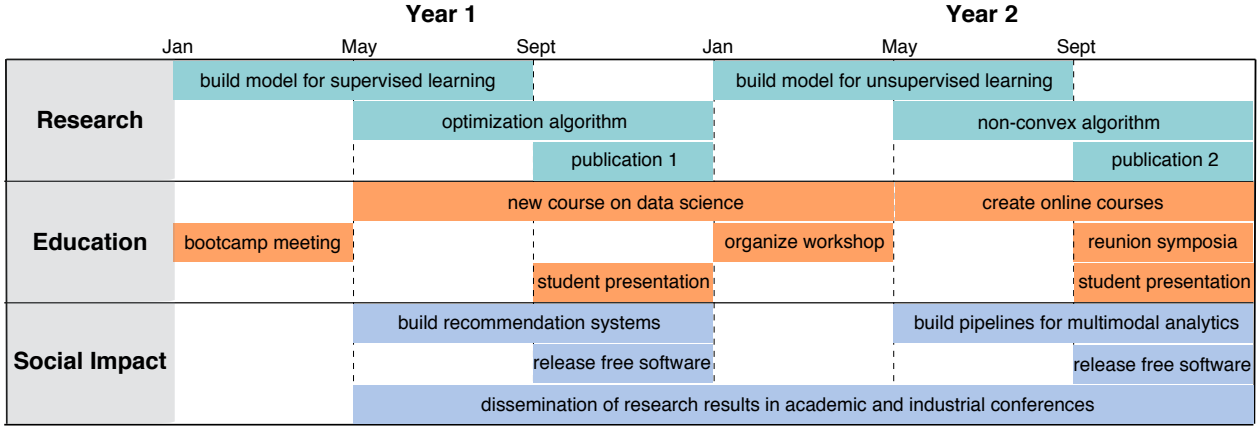
The learning problem is challenging because of the extremely high dimensionality in the tensor parameter space. The PI's previous extensive experience on tensor related work has shown that tensors sought in applications often possess special structures, such as (nearly) low-rankness, sparsity, non-negativity, or orthogonal decomposability. We will leverage the formalisms of *intrinsic dimension* to develop efficient statistical methods for analyzing these high-dimensional datasets. We will further develop adaptive, semi-supervised methods that incorporate practical constraints in insurance applications, such as incomplete observation with missing labels, corrupted distributions, and computation with time and memory constraints.

Impact and Significance:
- Develop efficient prediction tools for tensor classification, tensor regression, and deep tensor neural network in the presence of domain constraints.
- Build data-driven prototypes that integrate machine learning into the data-to-decision process.
- Improve efficiency in biomedical discovery using powerful multi-modal brain imaging analyses.

Project 2: Unsupervised learning with high-dimensional tensors. My goal in this direction is to develop unsupervised learning tools, i.e., clustering, denoising, and dimension reduction, for high-dimensional tensors, where the learner has little or no label information during training. This is an area where modeling and validation are notably difficult due to missing information. My group has recently developed an efficient tensor decomposition method to successfully estimate salient blocks in the multi-relational network data. Our clustering algorithm discovers community structure with higher accuracy, while being $11\times$ faster, than the competing methods. We are currently generalizing the methods for integrative analysis of multimodality networks. The work will unlock several ambitious directions that lead to robust decision-making in science, engineering, biomedical applications including health industry.

Impact and Significance:
- Develop interactive data exploration and visualization for unsupervised learning tasks such as tensor clustering, density estimation, and dimension reduction.
- Design robust tools to analyze tensor data and extract hidden information to aid biomedical decision making.
- Address data mining challenges from information gathering to pattern recognition.

Deliverables and Milestones. I am one of the few young faculty members in Statistics. My group consists of three PhD students and two Master/undergraduates, of which three are females (me included) and three are males. I have been actively contributed to several multi-institutional consortiums, in collaboration with researchers in United States, France, and China. This grant will support me to further create a diverse working group. The following table summarizes the planned milestones for my proposal.



The project will create deliverables in three aspects: **research**, **education**, and **social good**. I plan to devote one year for each of the aforementioned research problems. Regarding teaching, I will create a new course on *data science* in both traditional and online formats. The new course will introduce students to the real word challenges in implementing statistical machine learning approaches to decision making. Open-source software will be released, as the fruit of the research, that facilitates academia, industry, and society to analyze complicated tensor data. The PI will organize a series of workshops on the scientific applications of tensor data analysis, with the aim to encourage interdisciplinary collaborations.

Miaoyan Wang
Assistant Professor of Statistics
1300 University Avenue, UW-Madison
miaoyan.wang@wisc.edu

(a) Professional preparation

Fudan University	China	Mathematics	B.S. 2010
		Computer Science	2006-2007
University of Chicago	USA	Statistics	PhD, 2015
University of Pennsylvania	USA	Mathematics and Biology	Simons Math+X Postdoc, 2017
UC Berkeley	USA	Computer Science	Postdoc 2018

(b) Appointment

2018 - Present, Assistant Professor, Department of Statistics, University of Wisconsin–Madison

(c) Ten Selective Publications (Students under my supervision are underlined)

(i) C. Lee and **M. Wang.** Tensor denoising and completion based on ordinal observations. International Conference on Machine Learning (ICML). In press, 2020.

M. Wang and Y. Zeng. Multiway clustering via tensor block models. Advances in Neural Information Processing Systems 32 (NeurIPS), 715-725, 2019

M. Wang, J. Fischer, and Y. S. Song. Three-way Clustering of Multi-tissue Gene Expression Data Using Semi-Nonnegative Tensor Decomposition. Annals of Applied Statistics. Vol. 13, No. 2, 1103-1127, (2019).

M. Wang, K. Dao Duc, J. Fischer, and Y.S. Song. Operator Norm Inequalities Between Tensor Unfoldings on the Partition Lattice. Linear Algebra and its Applications, Vol.520, 44-66, (2017).

M. Wang and Y. S. Song. Tensor Decomposition via Two-Mode Higher-Order SVD (HOSVD). Proceeding of Machine Learning Research, Vol 54, 614-622, (2017).

(ii) M. Wang, F. Roux, C. Bartoli, C. H.-Chauveau, C. Meyer, H. Lee, D. Roby, M. S. McPeek, and J. Bergelson. Two-Way Mixed-Effects Methods for Joint Association Analyses Using Both Host and Pathogen Genomes. Proc. Natl. Acad. Sci. (direct submission), Vol. 115 (24), E5440-E5449, (2018).

D. Jiang and **M. Wang**. Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Studies of Complex Traits. Biostatistics & Epidemiology. Vol. 40, No. 6, 446–460, (2018).

M. Wang, J. Jakobsdottir, A. V. Smith, and M. S. McPeek. G-STRATEGY: Optimal Selection of Individuals for Sequencing in Genetic Association Studies. Genetic Epidemiology, Vol. 40, No. 6, (2016) 446-60. **Highlighted as Editor's Pick Paper of this issue.**

M. Wang and L. Li. Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and Its Statistical Optimality. Revision submitted to Journal of Machine Learning Research.

B. W. Engelmann, Y. Kim, **M. Wang**, B. Peters, R. S. Rock, and P. D. Nash. The Development and Application of a Quantitative Peptide Microarray Platform to Protein Interaction Domain Specificity Space. Molecular and Cellular Proteomics, Vol. 13, No. 12 (2014) 3647-62.

(d) Synergistic Activities

• Member in Women in Probability, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, American Society of Human Genetics. 2014 – now.
• Organizer for European Society for Evolution Biology workshop, International Conference on Frontier of Data Science, 2019.
• Reviewer for Journal of the American Statistical Association (JASA), NeurIPS, and Linear Algebra and application, and other applied math/statistics/genetics journals, 2014 – now.
• Statistical Consultant. Provided statistical support for the larger university community at the University of Chicago. 2012-2015.

(e) Current PhD students

Chanwoo Lee (2019 -): BS in Mathematics and Statistics, Seoul National University, 2018.
Jiaxin Hu (2020 -): BS/MS in Statistics, Wuhan University, 2020.
Yuchen Zeng (2019 -): current a PhD student in CS at UW-Madison.
Zhuoyan Xu (2019 -): current a PhD student in Statistics at UW-Madison.

(f) Recent Talks

Department Seminars: Columbia University, Stanford University, UC Berkeley, University of Chicago, CMU, Columbia University, University of Toronto, Fudan University, East China Normal University, Duke University, Johns Hopkins University, Queen's University, University of Massachusetts Amherst, University of Pennsylvania, Boston University.

Conference and Industrial Talks: Eastern North American Region (ENAR), International Conference on Frontiers of Data Science, European Society for Evolutionary Biology, Institute of Mathematical Statistics (IMS), Society for Industrial and Applied Mathematics (SIAM), Joint Statistical Meeting (JSM), American Society of Human Genetics (ASHG), International Genetic Epidemiology Society (IGES).

Industrial Research Lab Talks: Bosch Center for Artificial Intelligence, Takeda Pharmaceutical.

(g) Research Impact and Outreach

• Developed 8 open-source software packages for analyzing tensor datasets in genomics and neuroimaging.
• Faculty feature article ``Women in STEM: 5 Thoughtful Ways to Recruit and Retain Them'' in Course Hero.
• Won Charles J. Epstein Trainee Award for Excellence in Human Genetics Research –semifinalist (27 protectoral recipients out of 550 candidates), 2014
• Won Williams Award for Best Platform Presentation by Graduate Students – finalist (3 out of 156) in International Genetic Epidemiology Society (IGES), 2013.
• Runner-up for Department of Statistics Consulting Award (ranked as #2 among all PhD students in the departmental vote of 2014). Department of Statistics, The University of Chicago.
• Madison Teaching and Learning Excellence (MTLE) Fellow, 2019 -2020.