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

March 30, 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 recommendation systems.

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, insurance industry, 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 science problems arising from science and engineering.

This proposal lays out an ambitious plan aiming to aid decision-making in insurance application 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

Machine Learning with Tensors: From Theory to Application

Miaoyan Wang, Department of Statistics, http://pages.stat.wisc.edu/~miaoyan/

I work at the intersection of **statistics**, **machine learning**, and **optimization**, 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 scientific discovery and decision making for data-intensive challenges.

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. One example of tensor data arising from my current collaboration is multi-modal imaging of brain connectivity (Figure 1b). 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.

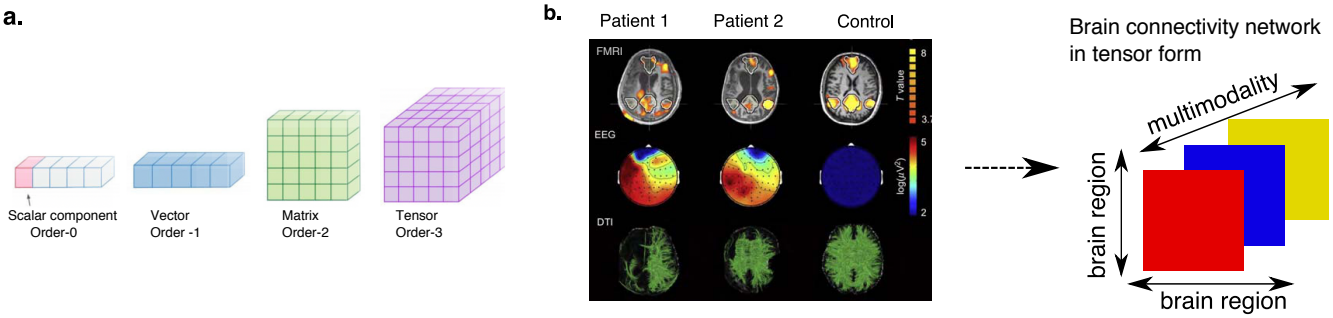


Figure 1: (a) Higher-order tensors are generalizations of matrices. (b) Tensor data from neuroimaging studies. Multiple measurements are used to construct brain connectivities within 1,200 individuals (Bruno et al., 2011).

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 and industrial applications can complement each other.

Project 1: Supervised learning with high-dimensional tensors. We consider prediction problems arising in machine learning, 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.

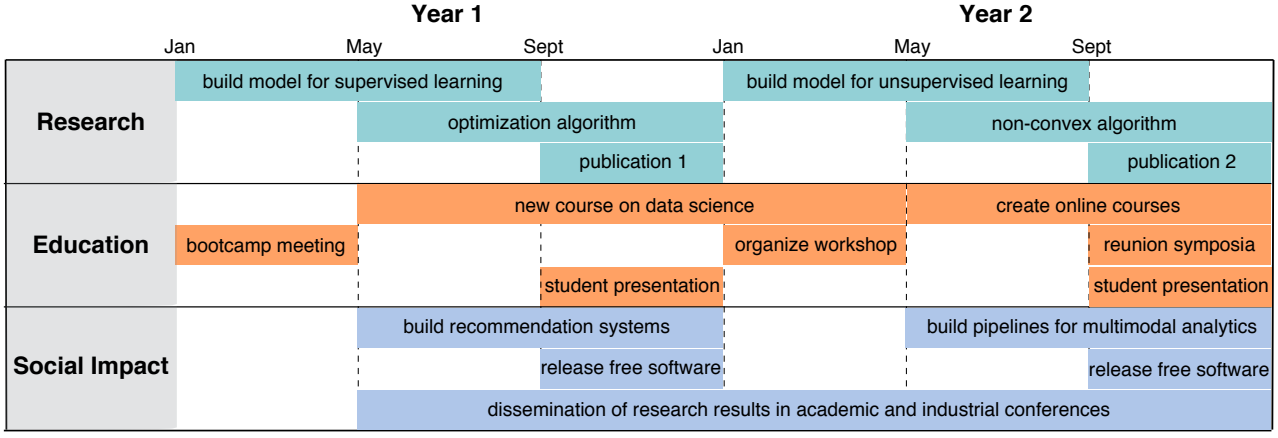
The learning problem is challenging because of the extremely high dimensionality in the tensor parameter space. The PI's previous extensive experience on tensor related work has shown that tensors sought in applications often possess special structures, such as (nearly) low-rankness, sparsity, non-negativity, 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 decision-making strategies in recommendation system for insurance industry.

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, and business applications including insurance 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 business decision making.
- Address real word challenges of data mining tasks from information gathering to market orders.

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, Korea, 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 complex 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
University of Wisconsin - Madison
1250B MSC, 1300 University Ave.
Wisconsin, WI 53703

Phone: 608-265-3990
Email: miaoyan.wang@wisc.edu

Research Interests

Machine learning applications to data fusion, recommendation system, and social network analysis

High-dimensional statistics, spectral methods for matrices and tensors

Employment

Assistant Professor of Statistics, 2018 - current
Department of Statistics, UW-Madison
Affiliated in Institute for Foundations of Data Science (IFDS), UW-Madison.

Postdoc at UC Berkeley (Computer Science) and at University of Pennsylvania (Math), 2016-2018.
Supervisor: Yun S. Song

Education

Ph.D., Statistics, The University of Chicago, 2015
Committee: Mary Sara McPeek (chair), Peter McCullagh, Dan Nicolae

B.S., Mathematics, Fudan University, 2010
2007 - 2010: switched to Math major (Rank: 1/172) after admitted into the China national program for talented students in fundamental sciences
2006 - 2007: majored in Computer Science (Rank: 1/105)

Selected Awards

UW-Madison nominee for Johnson and Johnson WiSTEM^{2}D scholar award in math discipline. Each University can nominate ONE candidate per discipline, 2019.

NeurIPS Travel Award, 2019.

Google Cloud Platform Education Grant. Free computing credit to students in class. 2019-2020.

Vice Chancellor for Research and Graduate Education (VCRGE) Travel Fund, UW-Madison, 2018.

Simons Math+X Postdoc Fellowship. Simons Foundation, 2015-2017.

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, 2014.

American Society of Human Genetics (ASHG) Charles J. Epstein Trainee Award for Excellence in Human Genetics Research – semifinalist (27 predoctoral recipients out of $\sim$ 550 candidates), 2014.

International Genetic Epidemiology Society (IGES) Williams Award for Best Platform Presentation by A Graduate Student – finalist (3 predoctoral recipients out of 156 candidates), 2013.

Paul Meier Fellowship, Department of Statistics, The University of Chicago. 2010 - 2013.

National Merit Scholarship $\times 3$ ($\sim$ top 1-2 per department per year), Ministry of Education of China. 2006 - 2007, 2007 - 2008, 2008 - 2009.

Honors Fellowship in Mathematics $\times$ 3, Fudan University, China. 2007 - 2008, 2008 - 2009, 2009 - 2010.

President's list, Top 10 role models out of $\sim 3,300$ freshmen, Fudan University, China, 2006 - 2007.

Publications

Graduate students under my supervision are underlined.

• C. Lee and **M. Wang**. Optimal tensor denoising and completion based on ordinal observation. Under review, `arXiv:2002.06524`, (2020).

• Z. Xu, J. Hu, and **M. Wang**. Exponential tensor regression with covariates on multiple modes. Under review. `arXiv:1910.09499`, (2020).

• **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,* `arXiv:1811.05076`, (2020).

• **M. Wang** and Y. Zeng. Multiway clustering via tensor block models. Neural Information Processing Systems 33 (NeurIPS 2019), 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,** 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).

Attention score in the top 5% of all research articles ever tracked by Altmetric; Higher than 89% of the research articles published in PNAS.

For this work, I was selected to present a **platform talk** *(21/156 submissions) at the 2nd Meeting for Probabilistic Modeling in Genomics, Cold Spring Harbor Laboratory, NY.*

• D. Jiang and **M. Wang**. Recent Developments in Statistical Methods for GWAS and High-throughput Sequencing Studies of Complex Traits. *Biostatistics and Epidemiology.* Vol. 2 (1), 132-159, (2018).

• **M. Wang** and Y. S. Song. Tensor Decomposition via Two-Mode Higher-Order SVD (HOSVD). *Journal of Machine Learning Research W&CP (AISTATS track)*, Vol 54, 614-622, (2017).

• **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,** 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, 446-60, (2016).

Highlighted as **Editor's Pick Paper** *of this issue.*

For part of this work, I was named a **semifinalist for the 2014 ASHG Charles J. Epstein Trainee Award** *(27 predoctoral recipients out of 550 candidates) for Excellence in Human Genetics Research at the Annual Meeting of American Society of Human Genetics; Also invited to give a platform talk (top 8%) in 2014 ASHG.*

For a different part of this work, I was named a **finalist for the 2013 IGES Williams Award** *(3 out of 156) for Best Platform Presentation by a Graduate Student at the Annual Meeting of the International Genetic Epidemiology Society; Also invited to give one of the six talks in Neels and Williams Awards Session in 2013 IGES.*

• 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, 3647-62, (2014).

Funding

Current (as sole PI):

National Science Foundation DMS-1915978, "Spectral methods for high-dimensional tensors", $179k, 2019-2022.

Wisconsin Alumni Research Foundation, Fall Research Competition, "Tensor theory and methods in Data Science", $39k, 2020-2021.

Pending:

National Science Foundation. TRIPODS: Collaborative: Institute for Foundations of Data Science, Role: senior personal (PI: Stephen J Wright). 2020-2025.

U.S. Department of Defense, ARMY. "Resolving the enigma of Factor Analysis", Role: co-I (PI: Karl Rohe). 2020-2023.

Mentoring

Thesis advisor

PhD student: Chanwoo Lee (2019 -), Jiaxin Hu (2019 -)

Alumni:

Yuchen Zeng (2018 - 2020). On to CS PhD in UW-Madison

Zhuoyan Xu (2018 - 2020). On to Industrial Engineering PhD in University of Washington.

PhD thesis committee

Luxi Cao (Advisor: Wei-Yin Loh), Ruosi Guo (Advisor: Zhengjun Zhang), Lili Zheng (Advisor: Garvesh Raskutti)

Selected Invited Talks

Department Seminars:

Columbia University, Department of Biostatistics, School of Public Health, 02/2020.

UW-Madison, Institute of Foundation of Data Science (IFDS), Brown Bag talk, 12/2019.

Fudan University, School of Data Science, Shanghai, China, 07/2019.

East China Normal University, Faculty of Economics and Management, Shanghai, China, 07/2019.

The University of Chicago, Department of Statistics, IL, 11/2018.

UW-Madison, Computation and Informatics in Biology and Medicine (CIBM) seminar, 11/2018.

UW-Madison, Systems, Information, Learning and Optimization (SILO) Seminar, 10/2018.

Stanford University, Department of Statistics, 08/2018.

UC Berkeley, Department of Biostatistics, 04/2018.

CMU, Department of Statistics and Data Science, 02/2018.

Columbia University, Department of Statistics, 02/2018.

University of Toronto, Department of Statistics, Canada, 02/2018.

UW-Madison, Department of Statistics, 01/2018.

Duke University, Department of Statistics, 01/2018.

Johns Hopkins University, Department of Biostatistics, 01/2018.

Queen's University, Department of Mathematics and Statistics, Canada, 01/2018.

University of Massachusetts Amherst, Department of Mathematics and Statistics,12/2017.

University of Pennsylvania, Song's Group, 09/2016.

Boston University, Department of Mathematics and Statistics, 08/2016.

Conference Talks:

Eastern North American Region (ENAR), International Biometric Society, 03/2020.

International Conference on Frontiers of Data Science, Hangzhou, China, 05/2019.

European Society for Evolutionary Biology - Coevolution Workshop, Munich, Germany, 03/2019.

International Conference on Data Science, Shanghai, China, 12/2018.

Institute of Mathematical Statistics (IMS) China Meeting, Dalian, China, 07/2019.

Society for Industrial and Applied Mathematics (SIAM) Annual Meeting, Portland, OR, 08/2018.

Joint Statistical Meetings, Chicago, IL. 08/2016.

Platform talk (**top 13% of all submissions)**, the 2nd Meeting for Probabilistic Modeling in Genomics, Cold Spring Harbor Laboratory, NY. 10/2015.

Platform talk **(top 8% of all submissions)**, American Society of Human Genetics (ASHG) Annual Meeting, San Diego, CA. 11/2014.

Neel and Williams Awards Talk, International Genetic Epidemiology Society (IGES) Annual Meeting, Chicago, IL. 09/2013.

Industry Research Lab Talks:

Bosch Center for Artificial Intelligence in North America, Sunnyvale, CA, 06/2018.

Global Analytic Group. Takeda Pharmaceuticals Inc., Deerfield, IL. 08/2014.

Poster Presentations (*presented by students I mentored):

*The 33rd Advances in Neural Information Processing Systems (NeurlPS), Vancouver, British Columbia, Canada, 12/2019.

*Computation and Informatics in Biology and Medicine (CIBM) Training Program and the Bio-Data Science (BDS) Program, Madison, 11/2019.

*2018 Berkeley Statistics Annual Research Symposium (BSTARS), Berkeley, CA, 03/2018.

The 20th International Conference on Artificial intelligence and Statistics (AISTATS), Florida, Fort Lauderdale, 04/2017.

The 7th Midwest Statistics Research Colloquium, Chicago, IL, 03/2014.

Teaching

Madison Teaching and Learning Excellence (MTLE) Fellow, 2019 -2020.

Lecturer, 2018 - current

Spring 2020. STAT 850: Theory and Application of Regression and Analysis of Variance-II (PhD core)

Fall 2019. STAT 849: Theory and Application of Regression and Analysis of Variance-I (PhD core)

Spring 2019. STAT 602: Statistical Methods–II (senior undergraduate)

Fall 2018. STAT 601: Statistical Methods–I (senior undergraduate)

Fall 2013. STAT 23400: Statistical Models and Methods (junior students in economics major)

Teaching Assistant at UChicago, 2011 - 2015

STAT 22000: Statistical Methods and Applications. Professor: Peter McCullagh

STAT 24400: Statistical Theory and Methods 1. Professor: Stephen Stigler and Debashis Mondal

STAT 24500: Statistical Theory and Methods 2. Professor: Debashis Mondal and Weibiao Wu

College Tutor at UChicago, Summer 2013: ECON 21000: Econometrics

Software

• **Tensor_ordinal**: A set of R tools for noise reduction and completion from ordinal tensor data with possibly missing values.
`https://cran.r-project.org/web/packages/tensorordinal/index.html`

• **Tensor_regress:** R program for generalized tensor regression with covariates on multiple modes.
`https://cran.r-project.org/web/packages/tensorregress/index.html`

• **Tensor_sparse**: R program for multiway clustering via tensor block models.
`https://cran.r-project.org/web/packages/tensorsparse/index.html`

• **Binary-Tensor**: R program for low-rank tensor estimation from binary observations.
`https://github.com/Miaoyanwang/Binary-Tensor`

• **Multi-Cluster**: Matlab/R programs for three-way clustering of gene expression tensorial data.
`https://github.com/Miaoyanwang/Multi-Cluster`

• **TM-HOSVD**: Matlab program for efficient decomposition of higher-order tensors.
`https://github.com/Miaoyanwang/Two-mode-HOSVD`

• **ATOMM**: C program for association **a**nalysis with a **t**wo-**o**rganism **m**ixed-effects **m**odel.
`https://github.com/Miaoyanwang/Two-way-MMA`

• **G-STRATEGY**: C program for optimal selection strategy of individuals for genotyping based on phenotypes and pedigrees. https://github.com/Miaoyanwang/G-STRATEGY

Services

Seminar Organizer:

European Society for Evolution Biology - Coevolution workshop, Munich, Germany, 2019

International Conference on Frontiers of Data Science, Hangzhou, China, 2019

Department Weekly Seminar Committee, Madison, WI, 2018 - 2019.

Grant Review Panel:

Research Grants Council (RGC) of Hong Kong, Physical Science Panel, 2020.

Natural Sciences & Engineering Research Council (NSERC, NSF equivalent in Canada), Discovery Grant Competition Panel, 2020.

University of Wisconsin Madison, Institute for Clinical and Translational Research (ICTR), Pilot Program, 2019.

Department Committee for PhD admission, PhD qualification exam, MS exam. 2018 - now.

Reviewer for Journal of the American Statistical Association (JASA), Journal of Machine Learning Research (JMLR), NIPS Comp Bio, Linear Algebra and Applications, Linear and Multilinear Algebra, Statistics and Probability Letters.

Member in IMS (Institute of Mathematical Statistics), SIAM (Society for Industrial and Applied Mathematics), ASHG (American Society of Human Genetics), IGES (International Genetic Epidemiology Society), Gooey Chocolate Cake Lunch (women faculty in the physical science in UW-Madison), Women in Probability (an NSF-funded organization for women active in probability research in North America).

Statistical Consultant, The University of Chicago, 2011 - 2015. Led consulting projects and supervised Masters and junior PhD students to provide statistical support for the larger university community.