# PERSONAL STATEMENT

Xiaodong Lin

I obtained my Ph.D. in Statistics in 2003, was an Assistant Professor at the University of Cincinnati from 2004-2008, and a tenured Associate Professor from 2008-2009. I joined Rutgers University as an Associate Professor in 2009 and was tenured in 2012. Before and since arriving at Rutgers, I have continually strived to achieve excellence in the three components of my academic responsibilities: namely, research scholarship, teaching scholarship, and service.

My research is motivated by the increasing demand in our society to extract critical and useful information from massive datasets in business, science, and industry. It focuses on the areas of Statistical Data Mining, Privacy Preserving Computing, and their applications. My research has fundamentally advanced methodological development in machine learning and substantially impacted practical applications across a wide range of fields. My recent industry experience has served as a bridge for the application of my existing research to practical settings and brought along new areas of scientific significance worthy of future exploration. My results have been adopted by companies such as Google and Lufax and used to serve millions of users.

Throughout my career in higher education, I have created and taught courses at virtually all levels. Thousands of students have benefited from these courses and my continuous mentorship after their graduation. I have supervised dozens of talented students towards their Master's and PhD degrees in cutting-edge areas of modern statistical science. My active participation on a number of committees demonstrates my dedication to serve the institutions I am a part of. Moreover, my advisory and leadership roles on different academic organizations have enabled me to impact the broader community, and I am honored to give back to my profession.

## RESEARCH

My research focuses on two major areas: statistical machine learning and data mining; and privacy protection in data processing, sharing, and mining. I have published extensively in flagship journals such as *SIAM Journal on Optimization*, *Journal of Machine Learning Research*, *Annals of Applied Statistics*, *IEEE Transactions on Information Theory*, and *ACM Transactions on Sensor Networks*. My research has also appeared in prestigious conferences including the *ACM SIGKDD*, *ACM IPSN* and *IEEE ICDM*, where there is a strict anonymous referee process and the acceptance rates are typically less than 20%. My Google Scholar Citations count as of August 2020 is 3149, with an h-index of 20 and i10-index of 28, which are clear indications of the impact of my research.

**Statistical machine learning and data mining**

The emergence of large scale and high dimensional datasets in Science, Business, and Engineering, creates unique theoretical and practical challenges in machine learning. My research in this area has concentrated on developing scalable tools for analyzing massive data with complex structures; in particular, efficient regularization techniques for independent and time dependent data models, and more flexible statistical methods for analyzing both static and dynamic networks.

My first major contribution is the development of a system of optimization techniques for large scale machine learning problems. Specifically, I developed an alternating linearization framework for solving general structured regularization problems including high dimensional fused lasso, group/graph lasso, and total variation regularization. This approach converges monotonically at a fast rate, and is capable of efficiently handling massive data with hundreds of thousands of dimensions. The following reference provides greater details on this contribution.

X. Lin, M. Pham, and A. Ruszczynski (2014), "Alternating Linearization for Structured Regularization Problems". *Journal of Machine Learning Research.* 15(Oct), 3447-3481.

As a natural extension of alternating linearization for two blocks, I proposed the first operator-splitting type method which is globally convergent for an arbitrary number of operators, without artificial duplication of variables. Such convergence results are achieved by selecting the optimal order of splitting based on the values of the functions minimized. This method is proven effective in optimizing the sum of multiple component functions as described in

Y. Du, X. Lin, and A. Ruszczynski (2017), "Selective Linearization For Multi-Block Convex Optimization". *SIAM Journal on Optimization.* 27(2), 1102-1117.

I am currently working on the application of this optimization framework to non-convex and stochastic settings, and this cohort of work has resulted in the successful completion of three Ph.D. dissertations. Along the line of the previous two papers, the following manuscripts have been prepared.

- M. Pham, Y. Du, X. Lin, and A. Ruszczynski, "An Outer-inner Linearization Method for Non-convex and Nondifferentiable Composite Regularization Problems", *Journal of Global Optimization. Revised.*
- K. Wu, X. Lin, and A. Ruszczynski, "Stochastic Alternating Linearization", *In preparation.*

Secondly, in the context of statistical network models, my research centers on flexible models

that extract structural patterns from both static and dynamic graphs. My first major achievement is a multi-way mixed membership stochastic block model designed to quantify relational information among objects across two or more populations. This model has broad applications to social, financial, and biological networks and has been successfully applied to large scale gene-metabolite and drug-drug interaction studies as published in

E. Airoldi, X. Wang, and X. Lin (2013), "Multi-way Blockmodels for Analyzing Coordinated High-dimensional Responses". *Annals of Applied Statistics*. 7(4), 2431-2457.

I am currently extending the multi-way block model to study time evolving interactions across multiple populations.

More recently, my attention has shifted towards dynamic network models. I have developed a novel framework that unifies dynamic graph clustering with latent network structure learning. The latent network is inferred from cascade data, and the model can incorporate time-dependent attribute information for identifying dynamic changes on the graph structure. This model has been successfully applied in anomaly detection in dynamic attributed graphs. My current work involves models that enable effective vertex representations for diffusion networks and dynamic networks. Working with economists from the Federal Reserve, I am also developing a dynamic network model to study the evolution of trading relationships in the corporate bond market. Such results enable us to identify anomalous trading behaviors and quantify market illiquidity.

My third achievement is the development of a series of machine learning tools that incorporate physical restrictions and domain knowledge for structured knowledge discovery. Notably, I proposed one of the first regularization frameworks for stationary time series based on penalized likelihood method, and developed a multivariate GARCH model that achieves sparse and efficient parameter estimation. This work extends the rich literature on statistical regularization for independent data models to time dependent settings, and has been successfully applied in analyzing high dimensional financial time series. I have also developed a novel technique for variable selection in Support Vector Machines by applying a nonconvex penalty. An efficient iterative procedure by successive quadratic approximation of the objective function is designed to solve the underlying optimization problem. This technique has been widely used by researchers from different fields as one of the most effective solutions for simultaneous feature selection and classification.

**Privacy preserving computing and cryptography**

Motivated by consumer privacy regulations such as GBDR and CCPA, privacy preserving computing is an emerging research area that aims to develop theory, methods, and protocols for

large scale statistical analysis without violating privacy constraints. I am one of the pioneers in studying privacy issues on distributed statistical computing. My first accomplishment is the development of a privacy preserving clustering method using finite mixture models and distributed EM algorithm; this is also one of the first attempts to provide privacy preserving solutions for iterative procedures. Secondly, I have developed comprehensive solutions for privacy preserving regression, model based clustering and maximum likelihood estimation, including procedures for model fitting, model diagnosis and prediction. Thirdly, I have proposed a hierarchical model for understanding and assessing information and privacy as well as a general framework to determine the optimal balance between privacy and data utility for synthetic and randomized data. These results helped shape the way researchers approach critical issues in data privacy and confidentiality and led to a series of articles in leading journals. In particular, the papers *Privacy Preserving Clustering with Distributed EM Mixture Modeling* and *Tools for Privacy Preserving Distributed Data Mining* are among the most cited ones in this field.

My work in this area has also evolved to the field of cryptography, where I have developed the first provable secure learning with error problems (LWE) & Ring LWE-based Diffie-Hellman type key exchange protocol. My paper entitled *A simple provably secure key exchange scheme based on the learning with errors problem* provide greater details. Our original work has inspired the building of many LWE-based key exchanges including the "New Hope" that was deployed in Google Chrome Canary. The efficacy and practical implication of our protocol is reflected in Google's claim that its deployment of the "New Hope" instantiation was a success. "New Hope" is also now a candidate in the second round selection of the post-quantum cryptographic standard by the National Institute of Standards of Technology.

I am currently working on three projects in this area. I am developing a decision theoretic framework that guides the evaluation of risk and utility trade off, thus providing practitioners with the methods and tools needed for making decisions regarding statistical disclosure limitation. In collaboration with colleagues from Duke University and HKUST, I am building a federated learning system for multi-party modeling of consumer credit risk. Working with colleagues at the University of Cincinnati and Tsinghua University, we are building new quantum resistant private set intersection protocols by extending our work in LWE key exchange.

**Applied work and industrial collaboration**

In addition to my scholarship on statistical methodologies, I have collaborated extensively with researchers from other disciplines. With electrical engineers, I applied advanced machine learning algorithms to indoor source localization and passive gesture recognition. Together, we designed and implemented one of the first device-free passive localization methods based

on probabilistic classification. We also developed a highly robust system for in-bed movement detection and classification using low-end load cells. This system is capable of continuous and accurate sleep monitoring and abnormal movement detection with privacy preservation. I have worked with computer scientists on topics related to active learning for massive streaming data. This has led to the development of an active learning framework which selectively labels instances from data streams to build classifiers adaptively. My collaborations with biologists from Cincinnati Children's Hospital have focused on developing quantitative models for describing native Bicoid gradient and applying classification and feature selection techniques to identify essential genes using partial label information. These applied contributions address significant multidisciplinary problems and have appeared in prestigious journals including *ACM Transactions on Sensor Networks*, *IEEE Transcations on SMC*, *Developmental Cell*, and *Nucleic Acids Research*.

During 2018-2019, I spent a two years leave of absence in industry, where I led a team of Machine Learning researchers and engineers to transform an online wealth management platform. I worked on three major areas: data driven risk management system, intelligent customer segmentation and marketing, and a wealth management chatbot. We developed a Know Your Customer (KYC) system that utilizes user demographic and behavior information to characterize risk tolerance and risk preferences. We applied a system of network models to analyze the transaction network and user behaviors in real time to generate personalized marketing strategies and product recommendations. Using a large amount of de-identified dialogs between users and financial advisors, we were able to build a dialog management system that can undertake a myriad of wealth management tasks. All these systems are currently online and serve millions of customers. This experience not only solidifies the strong demand for my research in the practical landscape, but also generates new research directions in the emerging field of financial technology.

My research career has focused on developing theoretical foundations and their applications in my areas of expertise. I have built up a strong multidisciplinary research program where I collaborate with colleagues in statistics, business, medicine, and other disciplines. My expertise has attracted many graduate students whose Ph.D. research is fully integrated into the program. Furthermore, my industrial ties have refined and strengthened my ability to transfer methodological advances to the practical domain. Overall, I have consistently produced high quality research, and I am confident that I will continue to make significant contributions to my discipline and the university.

**TEACHING**

I am a dedicated teacher whose goal is to cultivate my students' ability to observe and analyze the dynamics of the business world from a statistics perspective. To achieve this, I focus my teaching energies on carefully designing a cohesive curriculum, actively involving the students in the classroom, and maintaining personal contact with students outside of the classroom. When serving as a mentor and advisor, I place great value on clear objectives and constant communication. As a result, I am highly regarded by my students and have successfully supervised many M.S. and Ph.D. students at Rutgers University and the University of Cincinnati.

Throughout my academic career, I have taught a wide range of graduate and undergraduate classes. I am one of the core faculty for the Quantitative Finance program and Business Analytics program at Rutgers Business School. I created a new course *Financial Time Series Analysis* that is tailored to the needs of the Ph.D. and Master of Quantitative Finance program and crafted a curriculum that balanced both theory and practical applications. I also developed the course *Time Series Models for Business* that introduces predictive models and forecasting techniques to business students. These courses are widely successful as it has attracted hundred of students and, as evidenced by my overall course evaluation rating, is extremely well received. My teaching experience also includes short courses and seminars on machine learning and forecasting techniques to MBA and other professional students. I am currently developing the course *Machine Learning and Financial Technology*, using my practical experiences and knowledge in this domain to prepare students for the future of the financial industry.

At the University of Cincinnati, I taught *Statistical Inference*, *Applied Regression Analysis*, *Mathematical Statistics* and *Analysis of Variance* at the graduate level, and *Probability & Statistics* and *Introduction to Statistics* at the undergraduate level. Additionally, I developed a Ph.D. course *Statistical Machine Learning and Data Mining*, for which I also designed a specific curriculum and course materials. Overall, my classes at the University of Cincinnati garnered much positive feedback from students and faculty members alike and consistently achieved high course evaluations.

In addition to classroom teaching, I am actively advising graduate students. I have supervised four Ph.D. students, and am currently the thesis advisor for two more Ph.D. candidates. I have also served on dozens of Ph.D. dissertation committees and supervised multiple Capstone projects for MITA students.

I believe that I have demonstrated my ability to be a successful teacher. My past teaching experiences have been tremendously rewarding, and I am confident that I will fulfill my goal of

being an even more outstanding instructor and mentor at Rutgers University.

**SERVICE**

At Rutgers, my service has consisted of contributing to various departmental and university committees. As a member of the Curriculum 2030 committee, I am tasked to reform the curriculum to prepare for the challenges and opportunities facing future business education at RBS. I served on the BAIT major planning committee and subsequently the BAIT curriculum committee to develop one of the most sought after majors at RBS. As the department representative, I served on the special program policy committee for approving new programs. I have also served on the recruitment committees for both faculty members and graduate students. At the University of Cincinnati, I played a critical role in establishing the computational science institute. I also served on the college's faculty advisory committee among other service engagements.

My service to the community includes refereeing for dozens of journals and conferences including JASA and Information Systems Research. I have served as an Associate Editor for several academic journals and on the program committees of top conferences such as SIGKDD and ICDM. I was an organizer and advisor for the International Conference on High Frequency Finance and Data Analytics and a founding member of the International Academy of Information Technology and Quantitative Management. I played and continue to play an active role in the Risk Analysis and Statistical Learning and Data Science sections of ASA.

I have been honored to play a leadership role in developing the direction of business statistics at RBS and having a voice in the broader academic community through my participation in a vast number of committees, journals, and conferences. I look forward to continuing my active engagement in the community and shaping the business landscape with all my expertise and experience.

I am delighted by the prospect of continuing to implement the transformative leadership I have consistently provided throughout my career in research scholarship, teaching scholarship and service to both my profession and my university.