BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Yingying Fan

eRA COMMONS USER NAME (credential, e.g., agency login): YINGYINGF

POSITION TITLE: Associate Professor of Data Sciences and Operations

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
University of Science and Technology of China	BS	07/2003	Statistics and Finance
Princeton University	MS	06/2006	Operations Research and Financial Engineering
Princeton University	PhD	06/2007	Operations Research and Financial Engineering

A. Personal Statement

This proposal aims at developing mathematically rigorous and computationally efficient approaches to deal with highly complex big data and applying these approaches to solve fundamental and important biological and biomedical problems. We will 1) theoretically investigate the power of the recently proposed model-free knockoffs (MFK) procedure, and theoretically justify the robustness of MFK with respect to the misspecification of covariate distribution; 2) develop deep learning approaches to predict viral contigs with higher accuracy, integrate our new algorithm with MFK to achieve FDR control for virus motif discovery; 3) take into account the virus-host motif interactions and adapt our algorithms and theories in 2) for predicting virus-host infectious interaction status; 4) apply the developed methods from the first three aims to analyze the shotgun metagenomics data sets in ExperimentHub to identify viruses and virus-host interactions associated with several diseases at some target FDR level.

Previously, I developed advanced statistical and computational methods for big data analyses, including variable selection, classification, interaction screening and selection, association network estimation, and false discovery rate control. For almost all these methods, I also theoretically justified their effectiveness and applied them to real data sets such as gene expression data to test their applicability. Recently, I have become increasingly interested in using my expertise in these areas to solve problems in microbiome study. In particular, I am in the process of integrating and adapting the novel model-free knockoffs framework that I recently proposed with the deep learning methods to analyze sequence data and predict viral contigs with higher accuracy and controlled FDR. My previous work and experience have well prepared me with the expertise, collaborations, leadership, and motivation necessary to successfully lead the proposed work.

a. Candès, E. J., Fan, Y., Janson, L. and Lv, J. (2017). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, to appear.
b. Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics* **45**, 897–922.
c. Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics* **44**, 2098–2126.
d. Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43**, 1243–1272.

B. Positions and Honors

Positions and Employment

2007 - 2008	Lecturer, Department of Statistics, Harvard University, Cambridge, MA
2008 - 2009	Visiting Assistant Professor, Information and Operations Management Department, University of Southern California, Los Angeles, CA
2009 - 2015	Assistant Professor, Data Sciences and Operations Department, University of Southern California, Los Angeles, CA
2015	Visiting Scholar, Department of Statistics (Host: Professor Peter Bickel), University of California, Berkeley, Berkeley, CA
2016-	Board Member, USC Machine Learning Center, University of Southern California, Los Angeles, CA
2017-	Associate Fellow, USC Dornsife Institute for New Economic Thinking (INET), University of Southern California, Los Angeles, CA
2015-	Associate Professor, Data Sciences and Operations Department, University of Southern California, Los Angeles, CA

Other Experience and Professional Memberships

2009 - 2011	Membership Committee of the International Chinese Statistical Association
2014	Invited review panelist for NSF Grant Proposals
2012 -	Associate Editor of Journal of the American Statistical Association (2014- Present); Associate Editor of Journal of Econometrics (2015 - Present); Associate Editor of The Econometrics Journal (2012 - Present); Associate Editor of Journal of Multivariate Analysis (2013 - 2016); Guest Associate Editor of Statistica Sinica (2013 - 2014)
Member	American Statistical Association (ASA), Institute of Mathematical Statistics (IMS), The Royal Statistical Society (RSS)

Honors

2009 – 2012	National Science Foundation Grant, PI
2010	USC Marshall Dean's Award for Research Excellence
2010 - 2011	Zumberge Individual Award from USC, PI
2013	Noether Young Scholar Award
2014	The Inaugural Dr. Douglas Basil Award from USC Marshall
2017	The Royal Statistical Society (RSS) Guy Medal in Bronze
2017	USC Marshall Dean's Award for Research Excellence
2016 - 2017	USC Marshall Outlier Research Grant, Co-PI
2012 - 2017	National Science Foundation (NSF) Faculty Early Career Development (CAREER) Award, PI
2017 - 2018	Lord Foundation Grant from USC Marshall, PI

C. Contributions to Science

Project on FDR Control with applications to Crohn's disease. Reproducibility is crucially important in many scientific discoveries. One research topic I have been working on is false discovery rate (FDR) control in high-dimensional variable selection, where one is interested in identifying important variables that contribute to certain outcome with controlled error rate on variable selection. Having controlled FDR is an effective way for enhancing reproducibility in many scientific studies. Most existing literature addresses the challenging issue of FDR control by consulting p-values calculated using some classical statistical theory. However, in Candès, Fan, Janson and Lv (2017), we discovered numerically that when moving away from linear model, p-value produced by classical theory becomes invalid, in the sense of having non-uniform distribution under null hypothesis when the dimension is non-negligible compared to sample size. As a result, the p-value based FDR control methods may have uncontrolled FDR. To overcome this difficulty, we proposed a new framework, named model-free knockoffs (MFK), for FDR control in general nonlinear models with arbitrary dimensionality.

Our method bypasses the use of p-value and has been proved to have exact FDR control in finite sample size and with arbitrary conditional dependence structure of response on covariates. We have applied our method to the genetic analysis of Crohn's disease, a 2007 case-control study by WTCCC. Compared to existing results, our methods made twice as many discoveries, with many new discoveries confirmed by a larger meta-analysis.

a. Candès, E. J., Fan, Y., Janson, L. and Lv, J. (2017). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, to appear.

Project on precision matrix estimation with applications to gene expression data. How to develop highly scalable methods for estimating the full conditional dependence structure among a large number of variables? Such problem of large-scale graphical network learning was addressed in Fan and Lv (2016), where we proposed a novel method ISEE which converts the problem of precision matrix estimation to that of large covariance matrix estimation through high-dimensional linear regressions. This new formulation enables us to utilize the recent developments in large covariance matrix estimation and high-dimension regression to the more difficult problem of graphical model estimation. We have applied our new method ISEE to a disease classification problem using gene expression data from a breast cancer study. Thanks to the scalability and efficiency of our new method, the classification accuracy was greatly improved compared to existing results in the literature. In the era of big data, another challenge that we often face is the problem of heterogeneity, where data are from different sources. To address such problem, we proposed a new tuning-free method THI (Ren, Kang, Fan and Lv, 2017) for heterogeneous inference with diverging number of large-scale networks. We applied this new method THI to gene expression data from triple-negative breast cancer study and uncovered the underlying connectivity pattern among genes. Compared to existing methods, our new method yields more interpretable results with controlled FDR.

a. Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics* **44**, 2098–2126.
b. Ren, Z., Kang, Y., Fan, Y. and Lv, J. (2017). Tuning-free heterogeneity pursuit in massive networks. *Journal of the American Statistical Association*, under 2nd review for revision.

Project on interaction screening and selection with application to disease classification data. How to effectively and efficiently identify interactions among vast number of variables that are important to certain outcome(s)? In Fan, Kong, Li and Zheng (2015), we proposed a novel method IIS for screening and selecting interactions in high-dimensional classification, where the main idea is to use the innovated transformation of the data. Then based on the identified interactions and all main effects, we proposed a new classification procedure IIS-QDA that automatically adapts between linear and quadratic classifiers. Then in Kong, Li, Fan and Lv (2017) and Fan, Kong, Li and Lv (2017), we proposed new methods IPDC and SPRING for interaction screening and selection in regression settings, where the former method is designed for scalar response and the latter method can handle diverging number of responses. These methods are highly efficient and particularly suitable for interaction pursuit in big data. These methods have been applied to gene expression data sets in both classification and regression settings, and have also been demonstrated to perform well compared to many state-of-the-art methods.

a. Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43**, 1243–1272.
b. Fan, Y., Kong, Y., Li, D. and Lv, J. (2017). Scalable principled interaction network learning. *The Annals of Statistics*, under review.
c. Kong, Y., Li, D., Fan, Y. and Lv, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics* **45**, 897–922.

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Research Support

USC Lord Foundation Grant, Yingying Fan (PI) 1/1/17-6/30/18
Scalable Heterogeneity Pursuit via Random Projection Ensemble

The goal of this study is to develop a novel framework for scalable heterogeneity pursuit with statistical guarantees to uncover the heterogeneity prevalent in the data, which is central to big data applications.

Role: PI

Completed Research Support (within the last three years)

NSF CAREER Award, Yingying Fan (PI) 8/1/12-7/31/17
High-Dimensional Variable Selection in Nonlinear Models and Classification with Correlated Data

The goal of this proposal is to develop new and effective methods for high-dimensional variable selection in ultra-high dimensional regression and classification. Methods are also developed for heavy-tailed data and when high correlations among covariates exist.

Role: PI