KAIJUN WANG

Fred Hutchinson Cancer Research Center
Office: Robert M. Arnold Building, M2-C816, Seattle, WA 98109
☎ 267.261.1836 ◇ ✉: kwang2@fredhutch.org
Homepage: https://kaijunwang19.github.io/main/

ACADEMIC EDUCATION

Fred Hutchinson Cancer Research Center Sep 2019 – Now
PostDoc, Vaccine and Infectious Disease Division
(https://www.fredhutch.org/en/research/divisions/vaccine-infectious-disease-division.html)
Project: High Dimensional Flow Cytometry Data Analysis
Supervisor: Prof. Ollivier Hyrien

Temple University, Fox School of Business Aug 2015 – Aug 2019
Ph.D., Department of Statistical Science
(https://www.fox.temple.edu/departments/statistical-science/)
Thesis Topic: Graph-based Modern Nonparametrics For High-dimensional Data
Advisor: Prof. Subhadeep (DEEP) Mukhopadhyay

University of Southern California Sep 2012 – June 2014
M.S., Mathematical Finance
(https://dornsife.usc.edu/mathematical-finance/)
Research Topic: Trading Strategy based on Modified RSI signal

Central University of Finance and Economics Sep 2008 – June 2012
B.S., Applied Mathematics (Currently School of Statistics)
(http://en.stat.cufe.edu.cn/about/index.html)
Thesis Topic: Clustering Based Cash-flow Anomaly Detection

RESEARCH INTERESTS

Nonparametric Statistical Learning, Graph Data Science, and Large-scale Inference, Biomedical Data Science.

HONORS & AWARDS

- Winner of school-wide PhD research paper competition 2017
  Temple University, Fox School of Business
- George Carides Memorial Award 2017
  Temple University, Fox School of Business
- Meritorious award, Interdisciplinary Contest in Modeling 2011
  Central University of Finance and Economics

PUBLICATIONS

Journal Publications:

1. Mukhopadhyay, S. and **Kaijun Wang** (2020) "A Nonparametric Approach to High-dimensional Ksample Comparison Problem". *Biometrika*, **107(3)**, page 555–572

Abstract: High-dimensional k-sample comparison is a common applied problem. We construct a class of easy-to-implement distribution-free tests based on new nonparametric tools and unexplored connections with spectral graph theory. The test is shown to possess various desirable properties along with a characteristic exploratory flavor that has practical consequences. The numerical examples show that our method works surprisingly well under a broad range of realistic situations.

2. Mukhopadhyay, S. and **Kaijun Wang** (2020) "Spectral Graph Analysis: A Unified Explanation and Modern Perspectives" arXiv:1901.07090 (*Accepted, Nature Scientific Reports*)

Abstract: Spectral graph analysis is undoubtedly the most favored technique for graph data analysis, both in theory and practice. The achievement of this paper is to take a step towards discovering a simple, yet universal statistical logic of spectral graph analysis, which has dual significance: (i) Theoretical side: The prescribed viewpoint accomplishes the miracle of unifying and generalizing the existing paradigm as a consequence of just a single formalism and algorithm. (ii) Practical side: it opens up several possibilities for constructing specially-designed more efficient spectral learning algorithms for complex networks; an example is given where we achieve more than a 350x speedup over conventional methods.

3. Zhu, Jiaqi, **Kaijun Wang**, Yunkun Wu, Zhongyi Hu, and Hongan Wang. (2016) "Mining User-Aware Rare Sequential Topic Patterns in Document Streams." *IEEE Transactions on Knowledge and Data Engineering*, **28**, page 1790–1804.

Abstract: We proposed Sequential Topic Patterns (STPs) and formulated the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet for characterizing and detecting personalized and abnormal behaviors of Internet users. The method can be applied in many real-life scenarios, such as real-time monitoring on abnormal user behaviors. The algorithms to solve problem involves three phases: preprocessing text data and identify sessions for different users, pattern-growth, and selecting URSTPs. Experiments on both real (Twitter) and simulated datasets show that our approach can indeed discover special users and interpretable URSTPs effectively and efficiently.

Under Review

4. Mukhopadhyay, S. and **Kaijun Wang** (2020) "On The Problem of Relevance in Statistical Inference".

Abstract: Given a large cohort of "similar" cases one can construct an efficient statistical inference procedure by learning from the experience of others (also known as "borrowing strength" from the ensemble). But what if, instead, we were given a massive database of *heterogeneous* cases? It's not obvious how to go about gathering strength when each piece of information is fuzzy. This paper develops a *new model* of large-scale inference to tackle some of the unsettled issues that surround the relevance problem. The central role is played by the concept of LASER, which are specially-designed Artificial RELevant Samples, and a vital piece to enable individualized custom-tailored inference. Through examples, we demonstrate how this new statistical perspective answers previously unanswerable questions in a realistic and feasible way.

5. Mukhopadhyay, S. and **Kaijun Wang** (2020) "Statistical Machine Learning: An Integrated Approach"

Abstract: This paper introduces a new data analysis framework, called Integrated Statistical

Learning (ISL) theory, which for the *first time*, offers solutions to blend the parametric statistical modeling and algorithmic machine learning into a coherent whole by establishing a link between them. Tthe challenges and potential gains of this new *integrated* statistical thinking are illustrated through various examples.

Working Papers

6. **Kaijun Wang** and Mukhopadhyay, S. (2020) "Integrated Statistical Solution to Beta Regression Problem".

   Description: Beta regression have been the state of the art method for modelling a response variable over $(0, 1)$ with high dimensional covariates. In this paper, we propose a different perspective to this problem by utilizing the modern LP nonparmetric statistics, which provides an elegant nonparametric answer for estimating these conditional distributions of the response variable.

7. **Kaijun Wang** and Ollivier Hyrien (2020) "Analyzing Dependency Structures of Responses to Stimuli in Flow Cytometry Data".

   Description: Flow cytometry data are usually used in modern study of immune-related diseases. So far, various model have been proposed on how to process the signal intensity to identify cell subsets of different marker combinations, yet little has been said what to do with the resulting count data. This is not a trivial matter as the resulting data set consists of sparse count data of high dimensionality, and analyzing it using traditional means are challenging. We believe that an important step to understanding this data set is finding out which subsets have similar responses to the stimuli. To this end, I'm developing a framework using mixture model and graph theory that can quickly cluster together marker combinations with similar reaction to stimuli for a given time period.

SOFTWARE

1. **Kaijun Wang**, Mukhopadhyay, S. (2017). LPKsample: LP Nonparametric High Dimension K-Sample Comparison. URL: https://cran.r-project.org/package=LPKsample.
2. **Kaijun Wang**, Mukhopadhyay, S. (2018). LPGraph: Nonparametric Smoothing of Laplacian Graph Spectra. URL: https://cran.r-project.org/package=LPGraph.
3. **Kaijun Wang**, Mukhopadhyay, S. (2018). LPMachineLearning: Integrated Nonparametric Statistical Machine Learning. URL: https://github.com/LPStat-Hub/LPMachineLearning.

TEACHING EXPERIENCE

STAT 1001, Quantitative Methods for Business I	Fall 2017 and Spring 2018, 2019
STAT 1102, Quantitative Methods for Business II	Fall 2016

RESEARCH PROJECTS

PhD level: **IQVIA** July 2018 – August 2018
Nonparametric Modeling for Advanced Targeting of Medical Service

Industry Supervisor: Prof. Yong Cai;
Academic Supervisor: Prof. Subhadeep Mukhopadhyay

We worked with a large high-dimensional page-view data to build a model that can differentiate between patients and normal subjects based on the pattern in their browsing history. Structural Topics Model and Latent Dirichlet Allocation are used to study the raw web browser logs, and high dimensional multiple comparisons are performed to learn the difference. The computation was performed in in `R` and `Python`.

Temple University: Deep Learning Project Aug 2017 – Dec 2017
Deep Neural Network Model for Seizure EEG data
Instructor: Prof. Joseph Picone

In this project, I built a deep learning model for identifying seizure from EEG data. For each EEG record the features vectors were obtained using cepstral transformation technique. I studied the effect of different setups of layers, optimizers, batch size and loss function on the model training and prediction. The deep learning models are implemented and trained using `Keras` package under `Python`.
(https://www.isip.piconepress.com/wang_kaijun.pdf)

Institute of Software, Chinese Academy of Science June 2015 – Dec 2015
Mining User-Aware Rare Sequential Topic Patterns in Document Streams
Joint work with J. Zhu, Y. Wu, Z. Hu and H. Wang

We proposed a method for mining user specific rare sequential topic pattern. The method includes steps that pre-process the input into user specific sessions, grow sequential patterns and their support, as well as mining the user-aware sequential topic pattern. I worked with Dr. Zhu and his group for developing the framework and running it on different data sets for demonstration. Our papers are published at *IEEE Transactions on Knowledge and Data Engineering.* (http://ieeexplore.ieee.org/document/7431989/)

Institute of Software, Chinese Academy of Science Aug 2014 – Jan 2015
Pecuniary Loss Computation for MH370 Victims
Leader: Xiaoming Deng

Our team developed a software that can compute the pecuniary compensation for MH370 victims. This is a joint project between Institute of Software and Chinese Ministry of Justice. This software is utilized for the negotiation of compensations.

Master level:

University of Southern California Jan 2013 – June 2014
Research Project: Trading Strategy based on Modified RSI signal
Advisor: Jin Ma

We designed a trading strategy that improves upon an existing method. This strategy is based on a trade volume adjusted RSI signal, and the performance is studied via S&P 500 historical data. The method is implemented using `Matlab`.

Undergraduate:

Interdisciplinary Contest in Modeling, COMAP Feb 2011
Environment Cost Model for Comparing Conventional and Electrical Fuel
Advisor: Xiaoming Fan

In this competition, my group compared the environmental and economic soundness of conventional fueled and electric fueled vehicles. The algorithm was based on a cost model where I combined both economic and environmental consideration. We implemented the method using `Matlab`. The project was awarded Meritorious Winner.

Online copy: https://www.comap.com/ICM_2011_Results.pdf

RELEVANT COURSE WORKS

- **Statistics**: Nonparametric Statistics, Advanced Statistical Inference, Multivariate Analysis, Survival Analysis, Time Series Analysis, Generalized Linear Models, Categorical Data.
- **Computer Science:** Machine Learning, Deep Learning, Data Mining, Neural Networks.
- **Mathematics:** Real Analysis, Complex Analysis, Graph, Topology, Stochastic Process.
- **Programming Skills:** Proficient with `R`, `Python`, `Matlab`.
  - **Data Analysis Tool:** `Spark`, `Pandas`.
  - **Large-scale Machine Learning Tool:** `Keras`.

PRESENTATIONS

1. Graph-Based Compressive High-Dimensional Testing, JSM Conference 2020
2. Graph-based Modern Nonparametrics For High-Dimensional Data 2018
3. Nonparametric High dimensional K-sample Comparison 2018
4. Analysis of Small Social Networks: Karate Club and Sampson Monk Network Data 2017
5. Time Series Analysis on Asian Stock Indices 2017
6. Nonparametric generalized linear model and Comparison with COM-Poisson 2016
7. Nonparametric Analyses of Brain Connectivity Data 2016

REFERENCE

Dr. Subhadeep(DEEP) Mukhopadhyay
Assistant Professor of Statistical Science
Temple University, Fox School of Business
Philadelphia, PA 19122-6083
Email: deep@unitedstatalgo.com
http://unitedstatalgo.com/
Tel: 979-587-9957

Dr. Ollivier Hyrien
Vaccine and Infectious Disease Division,
Fred Hutchinson Cancer Reserach Center
Robert M. Arnold Building, M2-C200, Seattle, WA 98109
Email: ohyrien@fredhutch.org
Tel: 206-667-2809

Dr. Pallavi Chitturi
Deputy Chair, Statistical Science Dept.
Director, Center for Statistical Analysis
Temple University, Fox School of Business
Philadelphia, PA 19122
Email: chitturi@temple.edu
Tel: 215-204-5070

Kuang-Yao Lee
Assistant Professor of Statistical Science
Temple University, Fox School of Business
1801 Liacouras Walk,
Philadelphia, PA 19122-6083
Email: kuang-yao.lee@temple.edu
https://sites.google.com/site/kuangyaolee1217