JONGHYUN YUN

Data Scientist, PhD in Statistics

@ **jonghyun.yun@gmail.com** jyun.netlify.app github.com/jonghyun-yun in jongyun-yun

- **Summary** PhD in statistics with 10+ working experience in industry and academia. Proficiency in advanced statistical modeling, (un)supervised learning, visualization, Bayesian inference, and big-data analytic tools including R, Python, SQL. Capable of developing innovative approaches, applying existing data mining algorithms, overcoming granularity and scalability issues, and managing trainees and assistants.

EMPLOYEMENT HISTORY

Data Scientist

Institute of Statistical Data Intelligence

09/2019 – Present Mansfield, TX, USA

- Develop ML methods for prediction modeling, ensemble methods, time series, causal inference, segmentation for big data. Apply NLP and survival model to analyze timestamped log data using TensorFlow. Processing, cleansing and validating the integrity of data using SQL and Pandas.
- Develop novel network modeling frameworks to discover dynamic interaction b/w customers and goods. Parallel programming using C/C++ for complex Bayesian inference. Present analysis and visualization using R and Python, and developing software packages.

Assistant Professor of Statistics

Department of Mathematics, University of Texas at Arlington

09/2016 – 08/2019 Arlington, TX, USA

- Responsible for bringing innovative machine learning approaches to studies broadly related to statistics, engineering, business, and biomedical fields, and continuously developing, growing and sustaining research lab infrastructure.
- Designed data science courses including data mining and regression analysis. Created hands-on examples for R and Python programming. Mentored and trained junior scholars. Managed staff of teaching assistants.

Assistant Professor of Statistics

Department of Mathematical Sciences, University of Texas at El Paso

08/2015 – 06/2016 El Paso, TX, USA

- Responsible for developing statistical methods in biomedical research, translating meaningful findings back to the community, supporting researchers in Border Biomedical Research Center.

Postdoctoral Researcher

Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center

09/2012 – 07/2015 Dallas, TX, USA

- Developed innovative statistical methods to detect genomic markers by using multiple sequencing data sources. Collaborated with scientists to design a method for cancer genomic research. Presented research outcomes to all levels of audience.

EDUCATIONAL HISTORY

PhD in Statistics

Department of Statistics, University of Illinois at Urbana-Champaign

09/2006 – 08/2012 Champaign, IL, USA

- Research in Monte Carlo methods for high-dimensional models with focus on solar weather prediction, target tracking, time series, and data assimilation. Dissertation on *Ensemble Filtering of State Space Models*. Advised by Yuguo Chen.

MA in Applied Statistics

Department of Applied Statistics, Yonsei University

03/2004 – 02/2006 Seoul, South Korea

- Research in high-dimensional prediction models with applications in smart wearable and word frequency. Thesis on *Bandwidth Selection in Dimension Reduction Regression*. Advised by Hakbae Lee.

BA in Business Administration and Applied Statistics

Yonsei University

03/1997 – 02/2004 Seoul, South Korea

- Related studies in economics, finance, marketing, and accounting. Minor in mathematics

STRENGTHS

General skills Project leadership, Interdisciplinary collaboration, Presentaton, Mentorship

Data science skills Advanced statistical modeling, Data mining, Data analysis, Machine learning, Predictive modeling, Dimension reduction, Data visualization, Time Series, Hidden Markov model, Natural language processing, Bayesian inference, Monte Carlo method, Algorithm design, Causal inference, Multiple hypothesis testing

Areas of experience Biostatistics, Bioinformatics, Economics, Genomic data analysis, Smart infrastructure, Item response model, Network model, log data analysis

Technical skills R, Python, C/C++, MATLAB, SPSS, SAS, Lisp, Bash, Spark, TensorFlow, SQL, Git, Linux, OSX, Hugo, MS Office, LaTeX, Markdown, Jupyter, Cloud/Parallel computing

FEATURED ON-GOING PROJECTS

Customer behavioral analysis (code)

- Developing a novel machine learning approach to analyze timestamped sequence of action data (log data) leveraging natural language processing and survival models.
- Identifying behavioral differences in consumer buying decision-making processes.
- Presenting analysis and visualization using R and python, and develop software packages

Network dependence analysis using time to events (code)

- Developing Cox model equipped with latent space to discover patterns between connection time and outcome in bipartite network models.
- Inference on test-taker's proficiency using accuracy and response times in educational assessments such as Duolingo.

- Identified relationship between customer shopping time and decision.

PUBLISHED INTELLECTUAL CONTRIBUTIONS

Refereed Journal Articles

1. Yun, J., Ryu, K. R. & Ham, S. Spatial Analysis Leveraging Machine Learning and GIS of Socio-Geographic Factors Affecting Cost Overrun Occurrence in Roadway Projects. *Automation in Construction* **133,** 104007 (2022).
2. Yun, J., Kang, S., Tehrani, A. D. & Ham, S. Image Analysis and Functional Data Clustering for Random Shape Aggregate Models. *Mathematics* **8,** 1903 (2020).
3. Yun, J., Shin, M., Jin, I. H. & Liang, F. Stochastic Approximation Hamiltonian Monte Carlo. *Journal of Statistical Computation and Simulation* **90,** 3135–3156 (2020).
4. Nam, J. H., Yun, J., Jin, I. H. & Chung, D. hubViz: A Novel Tool for Hub-Centric Visualization. *Chemometrics and Intelligent Laboratory Systems* **203,** 104071 (2020).
5. Cai, L., Li, Q., Du, Y., Yun, J., Xie, Y., DeBerardinis, R. J. & Xiao, G. Genomic Regression Analysis of Coordinated Expression. *Nat Commun* **8,** 2187 (2017).
6. Yun, J., Yang, F. & Chen, Y. Augmented Particle Filters. *Journal of the American Statistical Association* **112,** 300–313 (2017).
7. Chen, B., Yun, J., Kim, M. S., Mendell, J. T. & Xie, Y. PIPE-CLIP: A Comprehensive Online Tool for CLIP-seq Data Analysis. *Genome Biol* **15,** R18 (2014).
8. Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., Wang, T., Kim, J., Yun, J., Xie, Y. & McKnight, S. L. Poly-Dipeptides Encoded by the C9orf72 Repeats Bind Nucleoli, Impede RNA Biogenesis, and Kill Cells. *Science* **345,** 1139–45 (2014).
9. Yun, J., Wang, T. & Xiao, G. Bayesian Hidden Markov Models to Identify RNA-Protein Interaction Sites in PAR-CLIP. *Biometrics* **70,** 430–440 (2014).

Non-Refereed Articles

1. Yun, J. & Chen, Y. Comments on "Particle Markov Chain Monte Carlo Methods" by C. Andrieu, A. Doucet, and R. Hollenstein. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **72,** 332–333 (2010).

Book Sections

1. Wang, T., Yun, J., Xie, Y. & Xiao, G. in *Hidden Markov Models* 177–184 (Humana Press, New York, NY, 2017).

Software

1. Yun, J. *Statistical Data Intelligence Tools for Cost-Overrun Analysis of Roadway Construction Projects* 2021. github.com/jonghyun-yun/dico.
2. Yun, J. *TEMPEST: Latent Space Competing Risk Model for Accuarcy and Reponse Time Data* https://github.com/Jonghyun-Yun/TEMPEST.
3. Yun, J. *Process Data Modeling for PIACC Data* 2021+. https://github.com/Jonghyun-Yun/proda.
4. Alvarez, H. & Yun, J. *Baseball Statistics Collecting Functions from HTML Tables* 2017. https://github.com/jonghyun-yun/brscrap.git.
5. Yun, J. *A MATLAB Toolbox to Identify RNA-protein Binding Sites in HITS-CLIP* 2013. https://qbrc.swmed.edu/labs/xiaoxie/download/README1.pdf.
6. Yun, J. *R Package for PAR-CLIP Analysis* 2013. https://qbrc.swmed.edu/labs/xiaoxie/download/README2.pdf.

Working Papers

1. Jin, I. H., Jeon, M., Yun, J., Schweinberger, M. & Lin, L. Hierarchical Network Item Response Modeling for Discovering Differences Between Innovation and Regular School Systems in Korea. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2020+). Invited for revision.
2. Yun, J., Jin, I. H. & Jeon, M. Latent Space Competing Risk Modeling for Accuacy and Response Time Based on Tests. *Journal of the American Statistical Association* (2021+). To be submitted.
3. Yun, J., Ick Hoon, J. & Minjeong, J. Analysis of Time-Stamped Action Sequences (2021+).
4. Yun, J., Wang, T., Wang, X. & Xiao, G. Identification of RNA-protein Binding Sites in HITS-CLIP Using Heterogeneous Logit Models via Semi-Supervised Learning (2021+).
5. Yun, J. & Chen, Y. Localized Agumented Particle Filters (2021+).
6. Yun, J., Wang, T., Wang, X. & Xiao, G. The Identification of Differential Binding Sites in CLIP-seq.

PRESENTATIONS

Invited Talks

11/2021 "Latent Space Accumulator Model for Analyzing Bipartite Networks with Connection Times and Its Applications to Item Response Data", *Autumn annual conference of the Korean statistical society*, virtual.

02/2017 "Integrative modeling approaches for next-generation sequencing data", *Colloquim Series*, Texas A&M University-Commerce.

06/2016 "Model based identification of RNA-protein binding sites", Bioinformatics Session, *International Workshop on Applied Probability*, Toronto, ON, Canada.

10/2015 "Comparative analysis of CLIP-seq under multiple experimental conditions", *Border Biomedical Research Center Seminar*, *UT El Paso*, El Paso, TX, USA.

08/2014 "Statistical strategies for identification of the RNA-protein binding site in CLIP-seq", Biometrics Section, *2014 Joint Statistical Meetings*, Boston, NY, USA.

10/2014 "Statistical models to identify RNA-protein binding sites from CLIP experiments", *Computational and Systems Biology Seminar*, UT Southwestern, Dallas, TX, USA.

10/2011 "Augmented particle filters", *Robert Bohrer Student Workshop in Statistics*, University of Illinois at Urbana-Champaign, Champaign, IL, USA.

Poster Presentation

02/2014 "Identification for RNA-protein binding sites in CLIP-seq", *7th Annual Bayesian Biostatistics and Bioinformatics Conference*, Houston, TX, USA.

PROFESSIONAL AND UNIVERSITY SERVICE

Professional Service

06/2016 Co-chair, Bioinformatics session at *2016 International Workshop on Applied Probability* at Toronto, ON, Canada.

University Service (UTA)

09/2017 – 08/2019 Department advisory committee.

09/2016 – 08/2019 Math preliminary exam B subcommittees.

01/2017 – 05/2017 Undergraduate affairs committee.

01/2019 – 08/2019 College of Science Data science working group.

04/2018 Judge, College of Science Aces Research Symposium.

University Service (UTEP)

Spring 2016 Math Club Zero committee

Referee/Reviewer Work (Journals)

- *Journal of the American Statistical Association*, *Journal of Computational and Graphical Statistics*, *Computational and Mathematical Methods in Medicine*, *Journal of Statistical Software*, *Journal of Probability and Statistics*, *Bayesian Analysis*, *International Journal of Data Science*, *Genes*, *Mathematics*, *International Journal of Environment Research and Public Health*, *Antibiotics*, *Axioms*, *Healthcare*

TEACHING ACTIVITIES

University of Texas at Arlington

Spring 2019 MATH6312 - Data Mining (10 students)
Fall 2018 MATH3316 - Statistical Inference (57 students)
Spring 2018 MATH5358 - Regression Analysis (13 students)
Fall 2017 MATH5312 - Mathematical Statistics I (12 students)
Spring 2017 MATH5392 - Selected Topics in Mathematics (Data Mining) (12 students)
MATH5313 - Mathematical Statistics II (6 students)
Fall 2016 MATH5312 - Mathematical Statistics I (14 students)

University of Texas at El Paso

Spring 2016 STAT5474 - Introduction to Data Mining (14 students)
Fall 2015 STAT5354 - Post-genomic Analysis (5 students)
BINF5113 - Math Seminar for Bioinformatics (4 students)

University of Illinois at Urbana–Champaign

Spring 2012 STAT200 - Statistical Analysis (51 students)
Summer 2011 STAT100 - Statistics (30 students)
01/2010 – 05/2011 STAT400-Statistics and Probability I (Discussion Section Leader)
Spring 2010 (59 students), Fall 2010 (60 students), and Spring 2011 (93 students)
08/2006 – 12/2009 Teaching Assistant: STAT100-Statistics, STAT400-Statistics and Probability I, STAT410-Statistics and Probability II, STAT424-Analysis of Variance, STAT429-Time Series Analysis, STAT510- Mathematical Statistics I, and STAT511-Mathematical Statistics II.

Yonsei University

12/2005 Preliminary Calculus

03/2005 – 12/2005 Discussion Section Leader: STA2101-Calculus (65 students) and STA2102-Linear Algebra (67 students).

03/2004 – 12/2004 Teaching Assistant: STA1001-Introductory Statistics, STA1001-Introductory Statistics, STA3102-Multivariate Statistical Analysis, and BC682-Statistical Methods for Behavioral Sciences.

DIRECTED STUDENT LEARNING

Graduate Supervised Research

09/2017 – 09/2019 Anthony Thomas (*Statistics*, UT Arlington)

Project: *Bayesian hierarchical dynamic factor models*

09/2017 – 12/2017 Mario Garza (M.S. *Statistics*, UT Arlington)

Project: *Forecasting sales using a finite-state HMM: an inventory control exercise*

5 M.S. Student Committees

09/2016 – 08/2019 Daniel Sang Le, Nidhi Kiran Dawda, Zachary Loucks, Hongbo Yu

Statistics, UT Arlington

09/2015 – 08/2016 Tun-Lee Ng

Statistics, UT El Paso

6 Ph.D. Student Committees

09/2016 – 08/2019 Souad Sosa, Izzet Sozucok, Geoffrey Schuette, Yi Liu, Mahmoud Jawad, Piyachart Wiang-nak

Statistics, UT Arlington

Undergraduate Supervised Research

Spring 2018 Henry Alvarez (*Mathematics*, UT Arlington)

Project: *Developing a software package to collect baseball statistics*