LUYAO PENG, Ph.D.

Email: luyaopeng.cn@gmail.com Website: pengluyaoyao.github.io Phone: (619)-313-3133

Research engineer applied statistics, deep learning, educational testing and natural language processing with background in statistics (Ph.D.) and quantitative methodology in education (M.A.).

EDUCATION

University of California, Riverside 2019
Ph.D. in Applied Statistics.

University of California, Riverside 2014
M.A. in Education: Quantitative Research Methods

Beijing Language and Culture University 2012
M.A. in Linguistics: second language testing and acquisition

Capital Normal University 2009
B.A. in English Language and Literature

WORK EXPERIENCE

Senior Research Engineer, ByteDance
Feb 2021 — Present
- Develop research project in educational testing, adaptive learning and NLP
- Develop assessment algorithms for educational products

Data Scientist II, Artificial Intelligence & Machine Learning Team, ACT, Inc.
July 2019 — July 2020
- Developed and led research projects on exact Gaussian Processes deep learning model for ACT domestic and international essay scoring and general NLP tasks
- Applied Seq2seq model to content generation by keywords
- Explored different deep learning models, such as BERT, GPT2, RNN, LSTM and Variational Deep Gaussian Processes, for Natural Language Understanding tasks, such as named entity recognition, sentiment analysis, content generation and chatbot

Research Intern (Machine Learning), CTB McGraw-Hill Education
June 2015 — August 2015
- Conducted machine learning research on abnormal behavior detection and forensic analysis
- Developed Kernel Principal Component Analysis algorithm in R language
- Developed and deployed an online visualization application for fraudulent detection

Research Assistant, University of California, Riverside
June 2016 — September 2016
- Preprocessed survey data of Oakland Unified School District (missing value imputation)
- Conducted factor analysis to understand the impact of school racial climate on teachers of colors

Statistical Consultant, University of California, Riverside
September 2015 — November 2016
- Led statistics workshops with topics on data mining, programming, and reproducible research
- Provided consultations on empirical research methods in various applied disciplines

PUBLICATIONS

- Subir Ghosh, Li Guo, Luyao Peng. 2018. Variance Component Estimators OPE, NOPE and AOPE in Linear Mixed-effects Models. *Australian & New Zealand Journal of Statistics 60*(4), 481-505.
- Gregory Palardy, Luyao Peng. 2015. The Effects of Including Summer on Value-added Assessments of Teachers and Schools. *Education Policy Analysis Archives, 23(92), 1-26.*

CONFERENCE PAPERS

- Luyao Peng, Empirical Bayesian Neural Networks for Classifications. *International Conferences on Machine Learning Techniques and NLP (MLNLP), 2020*
- Luyao Peng, Sandip Sinharay. Using Linear Mixed-effects Models to Detect Fraudulent Erasures at an Aggregate Level. *National Council of Measurement in Education Annual Meeting, 2020.*
- Luyao Peng, Subir Ghosh. An Algorithmic Construction of All Unbiased Estimators of Variance Components in Linear Mixed Effects Models. *Joint Statistical Meeting by American Statistical Association. 2019.*
- Luyao Peng, Raghuveer Kanneganti. Statistical High-dimensional Outlier Detection Methods to Identify Abnormal Responses in Automated Scoring. *National Council of Measurement in Education Annual Meeting, 2016.*
- Luyao Peng. Deterministic, Gated IRT Model for Continuous Probability of Item Cheating. *National Council of Measurement in Education Annual Meeting, 2015.*

WORKING PAPERS

- Luyao Peng. Empirical Bayesian Neural Networks for Named Entity Recognition.
- Luyao Peng, Sandip Sinharay. Using Linear Mixed-effects Models to Detect Fraudulent Erasures at an Aggregate Level. *Educational and Psychological Measurement (submitted)*
- Subir Ghosh, Luyao Peng. Searching for Optimal Estimators of Variance Components in Linear Mixed-effects Model Using a Regularization Matrix. *Journal of Statistical Planning and Inference (submitted)*

OPEN-SOURCE PACKAGES AND PROJECTS

"GaussianProcessClassificationModel" [link]

- Exact and variational Gaussian Processes classification model using tensorflow-probability

"MMeM" (Multivariate Mixed-effects Model) [link]

- Developed and maintained an R package for modeling multivariate mixed-effects using REML and Henderson3 methods.
- Downloads: 428/month

Gaussian Process Deep-and-Wide Model [link]

- Developed scalable Gaussian process deep and wide model for prediction and classification
- Applied the model to the ASAP essay data, Stanford Treebank sentiment analysis dataset and large-scale MIMIC3 medical data

Fraudulent Response Detection in Automated Essay Scoring [link]

- Applied Kernel Principal Component Analysis (KPCA) and Support Vector Machine (SVM) to detect abnormal essays in automated essay scoring examination and cheating erasures in forensic analysis

"regrrr" (Toolkit for Compiling and Visualizing Regression Results) [link]

- Co-developed an R package for regression result reporting, hypothesis testing, and visualization.
- Downloads: 389/month

NON-ACADEMIC ACTIVITIES AND AWARDS

- Co-founding Vice President, Data Science Club, University of California, Riverside, 2016-2019.
- Innovation and Entrepreneurship Award, UCR Office of Technology Partnerships, 2019.
- Excellent Ph.D. Student Fellowship, Graduate Division, University of California, Riverside, 2013.
- Excellent Paper Award, Second Language Acquisition Forum, Peking University. 2012.

TECHNICAL SKILLS

Certificates:

- Databricks Certification for Apache Spark, 2020
- Fellow in Data Science, The Data Incubator (TDI), 2018
  Trained with skills in machine learning toolkit, web scraping, SQL, mapreduce, Natural Language Processing, Spark and Tensorflow
- R and Spark Certification: Tools for Data Science Workflows, NISS, 2017.

Programming Language:

- Python, R, Shell, SQL, Git

Machine/deep learning toolkits:

- scikit-learn, TensorFlow, PyTorch, Transformers, fairseq, ParlAI

Big data processing:

- PySpark, HPCC, AWS

Visualization:

- Bokeh, Seaborn, Matplotlib, Shiny, ggplot2

REFERENCES

Subir Ghosh (Dissertation Chair)
Professor of Statistics, University of California, Riverside
subir.ghosh@ucr.edu

Sandip Sinharay
Distinguished Presidential Appointee, ETS
ssinharay@ets.org

Alina Von Davier
Chief of Assessment, Duolingo
avondavier@duolingo.com