# Zhenfeng Lin

*1600 Southwest Pkwy, #220, College Station, TX 77840, USA*

☎ (+01) 443-808-3242 | ✉ zflin@stat.tamu.edu | ⌂ zflin.github.io | in zhenfeng-lin-3141a39b

## Summary

**Skills**: Machine Learning; Big Data; Novelty/Outlier Detection; Deep Learning; Text Mining

**Coding**: Python (inc. Scikit-Learn, TensorFlow), R (inc. Rcpp, Shiny), C/C++, Git, SAS, SQL, Scala, Spark, Hadoop, MATLAB

## Education

**Texas A&M University** *USA*

Ph.D. in Statistics *06/2015 - PRESENT*

**La Serena Winter School for Data Science** *Chile*

Visiting scholar program: Applied Tools for Data-driven Sciences *08/2017 - 08/2017*

**The Statistical and Applied Mathematical Sciences Institute (SAMSI)** *USA*

Graduate Fellow *08/2016 - 01/2017*

**Sun Yat-Sen University** *China*

M.S. in Probability & Mathematical Statistics *09/2011 - 07/2013*

**Sun Yat-Sen University** *China*

B.S. in Mathematics & Applied Mathematics *09/2007 - 07/2011*

## Working Experience

**Chevron** *Houston, USA*

Data Scientiest Intern *05/2016 - 08/2016*

- Designed a work-flow to predict production of shale oil wells using R. The work-flow involves standard machine learning techniques:
  - Data cleaning that deals with outliers and missing values
  - Feature engineering that creates and selects important predictors
  - Prediction modeling that involves random forest, gradient boosting, neural network, SVM, multiple general linear regression, and LASSO
  - Parameter tuning that uses grid search to optimize model performance
  - Cross validation that compares performance of different models
  - Data visualization that smartly shows data's natural properties and prediction performance

**OriginLab** *Guangzhou, China*

Statistics Researcher & Software Engineer *07/2013 - 05/2015*

- Led the research and development of statistics algorithms for a popular scientific graphical software (OriginLab®) using C/C++/R/MATLAB. Developed built-in algorithms including Bivariate Gaussian Kernel Density Estimation, Bootstrapping, Distribution Fitting, and Unbalanced Repeated Measure ANOVA

## Project Experience

**Discovery of RR-Lyraes in Big Data** *College Station, USA*

Research project *02/2017 - 07/2018*

- Fitted more than one million of multi-band light curves in DES catalog with templates
- Extracted near 60 features of stars from fitted curves
- Implemented the calculation using R and Python on clusters
- Discovered about 5000 RR Lyraes with Random Forest classifier

**Hierarchical Bayesian Approach for PLRs Calculation** *College Station, USA*

Research project *09/2017 - current*

- Developed a hierarchical multi-band Gaussian processes for semi-periodic Mira light curves
- Designed a procedure to simulate Mira light curves
- Implemented the model using PyStan with paralleled computation
- Achieved near 97% accuracy in period recovery, compared to 90% for existent best method

### A Flexible Procedure for Positive–Unlabeled Learning

*College Station, USA*

Research project *11/2016 - 07/2018*

- Developed a flexible procedure to solve PU learning easily: use classifier to reduce dimension to one, and then apply one–dimensional methods
- Proved consistency of our proposed estimators
- Validated the procedure in different settings

### Galactic Archeology: phylogenetic tree of Stellar Populations

*La Serena, Chile*

Winter school project *08/2017 - 08/2017*

- Reproduced results in a paper, which applied phylogenetic tree algorithm on 21 stellars with chemical features
- Applied phylogenetic tree to a larger data, which consists of near 3000 stars
- On the larger data, several techniques are used: missing values are imputed; PCA and t-SNE are used for dimensionality reduction; other clustering methods are used for comparison with phylogenetic tree

### 2016 Capital One Student Modeling National Competition: Development of an Optimal Credit Card Transaction Fraud Prevention Strategy

*College Station, USA*

2nd Place, Finalist to present at Capital One Financial Corporation, VA *04/2016 - 04/2016*

- Analyzed Big (about 10 millions observations with hundreds of features) and Dirty (many outliers, missing) data
- Extracted credit card profiles from real transactions and created Recency, Frequency, Monetary (RFM) features to predict fraudulent transaction
- Performed reasonable data segmentation and built Ensemble eXtreme Gradient Boosting (EXGB) models

### Robust Control of Contracting Discrete-time Markov Decision Processes (DTMDPs) with First Passage Expectation Criteria

*Guangzhou, China*

M.S. thesis *09/2011 - 06/2013*

- Integrated the first passage expectation criteria into DTMDPs to optimize the system performance
- Controlled the transition law in a fuzzy set with uncertainty

### 9th National Graduate Mathematical Contest in Modeling Gene Recognition Algorithm

*Guangzhou, China*

2nd Prize *10/2012 - 10/2012*

- Established indicator sequences from training DNA sequence (mitochondrial gene of human) and obtained DNA power spectrum sequence by using fast discrete fourier transform (FDFT)
- Created ROC to estimate a threshold value for human species and predicted exon regions for the target DNA sequences

### 8th National Graduate Mathematical Contest in Modeling Analysis and Evaluation of Lodging Resistance in Wheat

*Guangzhou, China*

1st Prize *10/2011 - 10/2011*

- Identified the most important predictors of wheat lodging resistance by correlation analysis
- Established multiple linear regression of lodging index using R

## Publications

1. W. Yuan, L.M. Macri, A. Javadi, **Zhenfeng Lin**, and J.Z. Huang. Near-infrared Mira Period-Luminosity Relations in M33. The Astronomical Journal. 2018. `https://arxiv.org/abs/1807.03544`

## Papers in Preparation

1. **Zhenfeng Lin**, S. He, W. Yuan, L.M. Macri, and J.Z. Huang. "Period Estimation for a Set of Irregularly Sampled Quasi-periodic Functions with Application to Mira Stars."
2. **Zhenfeng Lin** and James P. Long. Mixture Proportion Estimation for Positive-Unlabeled Learning via Classifier Dimension Reduction. `https://arxiv.org/abs/1801.09834`
3. K.M. Stringer, P. Ferguson, **Zhenfeng Lin**, J.P. Long, L.M. Macri, J.L. Marshall, C. Nielsen, F. Paz-Chinchon, and the DES Collaboration. Discovery of RR Lyraes in multiband, sparsely-sampled data from the Dark Energy Survey using template fitting and Random Forest classification

## Presentations

1. **Zhenfeng Lin**. Automatic outlier detection for light curve data from AAVSO. *Cook's Branch Workshop*. Montgomery, TX, April 4, 2018.
2. **Zhenfeng Lin** and James P. Long. Mixture Proportion Estimation via Dimension Reduction with Classifier. *Rice 2017 Data Science Conference*. Poster presentation. Houston, TX, October 9-10, 2017.
3. **Zhenfeng Lin**. Fitting Multi-band Gaussian Processes Mira Model with RSTAN (HMC). Course project presentation. College Station, TX, November 20, 2017.
4. **Zhenfeng Lin**. Learning From Noisy Labels via Modified Logistic Regression. *Southeast Texas Chapter of the American Statistical Association (SETCASA) Poster Session*. Poster presentation. College Station, TX, April 21, 2017.
5. **Zhenfeng Lin**. Probabilistic Prediction Calibration using Brier Score. Course project presentation. Durham, NC, November 30, 2016.

## Honors & Awards

| | | |
|---|---|---|
| 2017 | **Bronze Prize**, SETCASA Poster Competition, TAMU | *College Station, USA* |
| 2016 | **2nd Place**, Capital One Student Modeling National Competition | *McLean, USA* |
| 2015 | **College of Science Lechner Fellowship**, TAMU | *College Station, USA* |
| 2015 | **OGAPS Dean's Doctoral Fellowship**, TAMU | *College Station, USA* |
| 2015 | **Excellent Graduate Scholarship**, SYSU | *Guangzhou, China* |
| 2011 | **1st Prize**, 8th National Graduate Mathematical Contest in Modeling | *Guangzhou, China* |
| 2008-2010 | **1st Class Excellent Student Scholarship**, SYSU | *Guangzhou, China* |