Research Statement
Yao Xie

Change-point detection

Change-point detection is a long-standing challenge in statistics and signal processing. Classic change-point detection usually considers univariate data and makes strong modeling assumptions such as independent and identically distributed (i.i.d.) data, and thus does not capture the complex nature of big data. My works along this line develop new approaches for change-point detection in big data, which tackle (1) structured change (such as sparsity, low-rank, network structures) in high-dimensional data, (2) non-stationary change and non-i.i.d. data, (3) spatial and temporally correlated data, as well as (4) extending to distribution-free methods. The impact of her research in this area is high since change-point detection is a fundamental question in scientific discovery (where change-point represents novelty) and monitoring complex systems and critical infrastructures (where change-point represents an anomaly or threat). My work in this area was supported by multiple NSF grants, including an NSF CAREER Award in 2017.

For parametric change-point detection, one of my highly cited papers presented a new statistical method for detecting sparse changes using multi-sensor observations. The proposed statistical method, the mixture procedure, incorporates a sparse structure into the detection statistic through a latent variable model; my paper leads to a series of works on sparse change detection in the literature. I have also provided a theoretical foundation for analyzing the procedure’s false alarm rate and detection delay of the detection procedure through novel stopping time analysis for dependent stochastic processes.

Modern complex and high-dimensional data tend to have complicated distributions that are hard to capture through parametric distributions. My work developed distribution-free and non-parametric methods for change-point detection. The proposed methods are based on the kernel minimum mean discrepancy (MMD) statistic, empirical histogram, and topological structures of data and consider non-ideal observation models, such as missing data, and indirect observations, such as sketching. In these papers, I present a rigorous theoretical analysis to achieve false alarm control for the detection statistics. In particular, I tackle the notorious technical challenges for analyzing non-parametric procedures, leveraging techniques from change-of-measure for correlated Gaussian random fields.

I have also worked on incorporating change-point detection for decision-making in non-stationary environments (e.g., non-stationary bandits, pricing) by developing a Joint Change-Point Detection and Time-adjusted Upper Confidence Bound (CU) algorithm, which consists of two components: the change-point detection component and the exploration-exploitation component. We theoretically show that our CU algorithm achieves the regret lower bounds (up to logarithmic factors), and our numerical study shows that our policy performs well in various market environments.

Selected Publications

• Spectral CUSUM for online network structure change detection
  Minghe Zhang, Liyan Xie, and Yao Xie.
Modern hypothesis tests

Hypothesis testing is a fundamental problem in statistics and a crucial building block for machine learning, such as classification and anomaly detection. However, traditional methods often fall short when it comes to modern challenges involving high-dimensional data with complex distributions, missing and imbalanced data, and low-dimensional structures. To tackle these challenges, I have developed various new hypothesis tests.

One of my recent contributions is a new non-parametric minimax hypothesis test that assumes the distributions under each hypothesis belong to two disjoint “uncertainty sets” constructed using the Wasserstein distance. These uncertainty sets contain all distributions close to the empirical distributions formed by the training samples in Wasserstein distance, making the approach more robust in small-sample-size regimes when we cannot accurately estimate the true data-generating distributions. Our approach has the advantage of computational tractability and explicit characterization of the optimal test. The optimal test is based on a pair of least favorable distributions (LFD) from the uncertainty sets, which is a reminiscence of Huber’s robust test. However, our LFDs are computationally tractable in general, despite the infinite-dimensional optimization problem we face when finding the minimax test. To tackle this challenge, we make a connection to recent advances in distributionally robust optimization. We also characterize the radii choice of the uncertainty sets and prove a theoretical upper bound for the sufficient radii based on the so-called profile function, which is defined as the minimum Wasserstein distance between the empirical distributions and distributions that yield the same test as the oracle one.

In another work, I develop a general theory for the goodness-of-fit test to non-linear models. Our main result shows that the residual of the model fit, by solving a non-linear least-square problem, follows a possibly noncentral $\chi^2$ distribution. We present a method to select the model orders sequentially, which has broad applications in machine learning and signal processing, such as determining the rank of low-rank matrices and tensors from noisy, partial, or indirect observations, determining the number of sources in signal demixing, and potential applications in determining the number of hidden nodes in neural networks.

Recently, my research has focused on bridging hypothesis tests with deep learning, which will provide a statistical foundation for trustworthy machine learning. The research effort is funded
by an NSF grant starting in 2022, for which I serve as a Lead PI. With the lack of systematic tools to provide theoretical performance guarantees and quantify the uncertainties, we have not yet seen wide usage of neural network methods for hypothesis testing, making notable discoveries in science, engineering, and medicine, or drawing conclusions from data. My research aims to address such critical challenges and enable machine learning-based classification algorithms to be applicable and trustworthy for making discoveries from data, akin to the role that hypothesis testing has played in past decades. By enabling comparisons between models and hypotheses in the face of randomness, statistical hypothesis testing has great potential to provide an essential piece of the puzzle in establishing verifiable and reliable Machine Learning algorithms.

Selected Publications

- **Goodness-of-fit tests on manifolds.**
  Alexander Shapiro, Yao Xie, and Rui Zhang.

- **Neural tangent kernel maximum mean discrepancy.**
  Xiuyuan Cheng, Yao Xie.
  NeurIPS 2021.

- **Robust hypothesis testing with Wasserstein uncertainty sets.**
  Rui Gao, Liyan Xie, Yao Xie.

Recovery of generalized linear models via convex optimization

I have worked on recovering generalized linear models (GLMs) through computationally efficient optimization techniques by leveraging their underlying structures. GLMs are a type of statistical model that can effectively model non-normally distributed data, such as binary observations, count data, and more. This flexible framework can be used to construct many types of regression models, including linear regression, logistic regression, and Poisson regression.

My research has focused on studying low-rank matrix completion and recovery for count data, including cases of incomplete observation, referred to as the "Poisson matrix completion and recovery problem." This problem is a crucial aspect of modern data analysis when there is missing data, and it is often utilized in applications such as recommender systems. To this end, I have investigated the fundamental limits of low-rank matrix completion and recovery using regularized maximum likelihood estimator for discrete data, which differs from existing approaches that assume continuous observations. My work demonstrates that the regularization likelihood approach for count data is nearly optimal, despite the fact that the Poisson distribution is only locally sub-Gaussian. This is theoretically significant, as it expands the techniques that can be used to perform analysis in such cases. Furthermore, I have explored the non-convex formulation of matrix completion and general low-dimensional manifold problems through a differential geometry perspective, providing new insights into the topic.

In addition, I have recently introduced a new computational framework for estimating generalized generalized linear models (GGLMs), a class of models that extends GLMs to account for dependencies among observations in spatio-temporal data. This approach uses a monotone
operator-based variational inequality method to overcome non-convexity in parameter estimation and provide guarantees for parameter recovery. These results apply to both GLMs and GGLMs, with a particular focus on spatio-temporal models. Additionally, I have presented model recovery guarantees and online instance-based bounds using martingale concentration inequalities.

Selected Publications

• **Convex parameter recovery for interacting marked processes.**
  Anatoil Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie.

• **Matrix completion with deterministic pattern - a geometric perspective.**
  Alexander Shapiro, Yao Xie, and Rui Zhang.

• **Poisson matrix recovery and completion**
  Yang Cao and Yao Xie.

Conformal prediction

The accuracy and reliability of prediction algorithms are essential for statistical and machine learning models, especially when it comes to high-stakes domains like finance, energy systems, and healthcare. However, predicting future outcomes based on past observations is not always straightforward, and it is essential to quantify the uncertainty associated with these predictions to make informed decisions. Prediction intervals, which provide an estimate of the range of values within which the true outcome is likely to fall, are of particular interest for high-stakes domains where accurate predictions are critical.

Traditionally, classic approaches for prediction intervals have relied on strong parametric assumptions of time-series models such as autoregressive and moving average (ARMA) models, which impose strong distribution assumptions on the data-generating process. These approaches are not always suitable for complex prediction models such as random forests and neural networks, which can make it challenging to quantify uncertainty accurately. To address this issue, recently, in the statistics and machine learning community, the conformal prediction has become a popular distribution-free technique for uncertainty quantification in machine learning. However, conformal prediction for time series is challenging due to the lack of exchangeability assumption in conformal inference. This means that the order of the data points matters in time series data, and there is a temporal correlation between the data points. This correlation makes it challenging to perform uncertainty quantification accurately and reliably.

To overcome this challenge, I propose a conformal inference framework for time series with scalar outputs that utilize the feedback structure of prediction residuals in sequential prediction problems. By exploiting the serial dependence across prediction residuals, we can obtain good instantaneous coverage for future prediction intervals. In other words, the most recent past residuals contain information about the immediate future ones, allowing for accurate and reliable predictions. Theoretically, I have shown that their algorithm enjoys the same finite-sample and distribution-free marginal coverage guarantee as traditional conformal prediction methods.
but with more suitable assumptions for time series, including stationarity and strongly mixing conditions.

Our proposed algorithm has been widely adopted in the industry and is incorporated in several popular libraries and frameworks. For example, it has been incorporated into the MAPIE module for scikit-learn in Python and implemented in Amazon Web Service (AWS)'s Library for Uncertainty Quantification Fortuna. We are currently working with Meta/Facebook to incorporate their algorithm into Kats, a general framework for performing time series analysis. As a Meta/Facebook team collaborator stated, "Good prediction intervals are one of the main pain points for the people who do forecasting here and our end consumers." Therefore, the proposed algorithm has the potential to make a significant impact in industries where accurate and reliable predictions are essential.

Selected Publications

- **Conformal prediction for time series**
  Chen Xu, and Yao Xie.
  In Revision, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2023. Conference version published in ICML 2021 (Long presentation, top 3%)

- **Sequential predictive conformal inference for time series**
  Chen Xu, Yao Xie.

Spatio-temporal point process models

Advances in sensing technology and computing have led to the generation of large amounts of discrete event data at unprecedented rates and scales. This type of data contains event times, locations (or node index if there is an underlying network structure), and potentially additional information in the form of "marks." Understanding the complex dependencies within the data, such as the latent influence of historical events on future events, can enable the prediction of future events and insight into underlying dynamics. I have worked extensively to develop new models and algorithms to address these challenges, including capturing complex dependencies, recovering network topology, performing causal inference, understanding temporal dynamics over networks, making predictions, and detecting anomalies and distribution shifts.

One of my particular contributions is a deep learning model for point process data. Although recurrent neural network (RNN) models are powerful for this type of data, they may not capture non-stationary dependencies effectively due to their recurrent structure. Our proposed model represents the influence kernel (rather than the intensity function) with neural networks. The influence kernel is approximated using a novel low-rank decomposition that enables efficient representation through deep neural networks, better performance, and computational efficiency. We also introduce a log-barrier penalty to maintain the non-negativity constraint of the conditional intensity. We demonstrate our proposed method’s performance and computational efficiency compared to the state-of-the-art on simulated and real data.

I have also used these models to study various datasets, including a high-resolution COVID-19 dataset in Cali, Colombia. This dataset provides precise location and time information for ev-
ery confirmed case, enabling vital insights into spatio-temporal interactions and disease spread in a metropolis. Using our non-stationary spatio-temporal point process, we assume that previously infected cases trigger newly confirmed cases and introduce a neural network-based kernel to capture spatially varying triggering effects. Our kernel is designed to enhance expressiveness while maintaining interpretability, and we also incorporate exogenous influences from city landmarks. Numerical results on real data demonstrate our method’s predictive performance for hot-spot prediction compared to the state-of-the-art, as well as its interpretable findings.

Selected Publications

- **Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data**
  Zheng Dong, Shixiang Zhu, Yao Xie, Jorge Mateu, Francisco J. Rodriguez-Cortes.

- **Spatio-temporal point processes with deep non-stationary kernels**
  Zheng Dong, Xiuyuan Cheng, Yao Xie.
  ICLR 2023.

Optimizing police operations through data-driven modeling

In large urban areas, police departments often divide the geographical region of a city into multiple patrol areas called zones (or precincts), and each zone is further divided into smaller areas called beats (or sectors). The design of patrol zones affects both the demand and the capacity for police services in each zone and beat, as well as the travel time of patrol units. Together, these factors determine the police’s response time to emergency calls and crime events, making the design of patrol zones critical to the efficiency of police operations.

To address this issue, I have developed a data-driven optimization framework for redesigning police patrol zones in an urban environment. Our objectives are to rebalance police workload among geographical areas and to reduce response time to emergency calls. We develop a stochastic model for police emergency response by integrating multiple data sources, including police incident reports, demographic surveys, and traffic data. Our framework was implemented by the Atlanta Police Department in March 2019, and we show that the new design has reduced the response time to high-priority 911 calls by 5.8% and the imbalance of police workload among different zones by 43%. Our work on Atlanta Police re-districting has been recognized for its practical applications and clear and intelligible writing, including being selected as one of four finalists for the prestigious INFORMS Wagner Prize in 2021.

Additionally, we propose a new statistical modeling framework for spatio-temporal-textual data and demonstrate its usage in crime linkage detection. Our approach captures linkages of crime incidents via multivariate marked spatio-temporal Hawkes processes and treats embedding vectors of the free text as marks of the incident, inspired by the notion of modus operandi (M.O.) in crime analysis. Numerical results using real data demonstrate the good performance of our method as well as reveal interesting patterns in crime data.

The crime linkage detection algorithm has been incorporated into the AWARE system used by the Atlanta Police Department, demonstrating its potential to enhance police operations. The project won the Smart 50 Award at the annual Smart Cities Connect and Expo in 2018 and was
showcased in the City of Atlanta’s "Experience the Smart City" event for the general audience in 2017. The paper was selected to be one of the three papers presented in "The Best of AOAS" Session at the Joint Statistical Meetings 2022.

Selected Publications

- Data-driven optimization for police zone design.
  Shixiang Zhu, He Wang, and Yao Xie.

- Crime linkage detection by spatial-temporal-textual point processes.
  Shixiang Zhu, and Yao Xie.

Summary and future directions

My research interests encompass both theory and application and are focused on the intersection of statistics, optimization, and machine learning. I am passionate about leveraging data to solve real-world problems with significant societal impact.

Moving forward, my research will continue to address the challenges posed by high-dimensional, incomplete, contaminated, and spatio-temporally dependent data that violate traditional statistical models’ assumptions. To do so, I will develop new models, methods, and theories that integrate cutting-edge techniques from statistics and optimization while taking data structure and geometry into account.

In addition to addressing these challenges, I am also interested in quantifying the uncertainty in predictions, estimation, and testing to enable robust decision-making and trustworthy machine learning. My work aims to go beyond traditional statistical inference and establish causal inference and more interpretable and explainable predictions.

While my ultimate goal is to impact practical applications, I employ rigorous mathematical tools to understand the theoretical properties of the methods I develop and to establish guarantees such as statistical recovery error bounds, sample complexity, detection delay, false alarm guarantees, and uncertainty quantification. I strive for optimality whenever possible.