Research Statement

Liyan Xie

November 8, 2025

My research lies at the intersection of statistics, optimization, and machine learning, with data as the central focus—its analysis, generation, and use for reliable decision-making. I am particularly interested in developing new models and algorithms for: (i) **analyzing data** in an online fashion, including *detecting distributional changes*, *hypothesis testing*, and *model fitting*, with an emphasis on computational efficiency, robustness, privacy preservation, and the ability to handle complex dependencies among data; and (ii) **generating data** via diffusion models that align with human preferences or constraints, including *controlled generation*, *missing value imputation*, and *task-aware data generation*. I aim to integrate data generation and analysis in a synergistic manner, enabling generated data to enhance the robustness, statistical efficiency, and empirical performance of downstream analytical tasks. My work finds pertinent applications in domains such as sensor networks, natural hazards, healthcare, and reliable AI.

Sequential Data Analysis

My work in sequential analysis focus on sequential change-point detection and hypothesis testing. In the field of change detection, my research proposed parametric [5, 18] and non-parametric [6] methods that achieve asymptotic optimality with a computational cost significantly lower than benchmarks. These methods have been extended to robust settings [3, 4, 17]. In the field of hypothesis testing, my research proposed robust hypothesis testing and calssificaiont methods [1, 22]. Applications span network/graph data [18, 19], point processes and neural science [2, 16], manufacturing, and pandemics [21, 20]. I aim to tackle the following four challenges in sequential statistical inference.

Computational efficiency. In sequential change detection—a common problem in statistics and signal processing—the objective is to quickly identify change points in streaming data [13]. Traditional generalized likelihood ratio (GLR) tests are asymptotically optimal but computationally expensive, with computational cost $O(|\log\alpha|^2)$ for each time update, where $\alpha \in (0,1)$ is the specified false alarm rate. This often prevents the GLR from being deployed in real time under stringent false-alarm requirements. My research introduces the *window-limited CUSUM test* [5], which retains asymptotic optimality while reducing computational complexity from $O(|\log\alpha|^2)$ to $O(\sqrt{|\log\alpha|})$. Beyond this computational efficiency, our findings indicate that this method outperforms many existing detection procedures when detecting small changes. This method has received much attention from the field and has led to much follow-up work. My work have applied it to seismic tremor detection [7], gesture tracking [8], and sensor networks [18].

Robustness. My research aims to design robust data-driven methods that work well in the presence of distributional uncertainties. These uncertainties can arise from non-stationary data distribution or our limited prior knowledge. Using tools from distributionally robust optimization, my research develops tractable robust hypothesis testing and change detection approaches [1, 3, 17, 22]. These works construct uncertainty sets around empirical distributions via the Wasserstein distance, and develop computationally efficient frameworks for solving the minimax problem, along with statistical insights into the optimization solutions.

Privacy. Traditional sequential detection or testing methods are typically non-private and lack adequate protection of sensitive information from individuals. In many applications, the data streams being monitored contain sensitive personal or organizational information, such as financial or medical records, whose disclosure could have severe consequences. My recent work [9] is among the first to study sequential detection under differential privacy, proposing computationally efficient algorithms and establishing provable guarantees on false-alarm rate, detection delay, and asymptotic optimality.

Spatio-temporal dependencies. Rapid advancements in sensing and computing technologies have led to a surge in "spatio-temporal event data," characterized by event times, locations, and additional marks. My research focuses on understanding complex dependencies among events. I have developed new models and algorithms for model fitting [2], uncertainty quantification [15], temporal dynamics over networks [19, 16], and their applications in epidemic modeling and prediction [21, 20].

Data Generation and Imputation

My research in data generation focuses on developing statistically principled and computationally efficient frameworks for generating and reconstructing data using diffusion models. My research seeks to integrate human preferences, physical constraints, and downstream tasks into the data generation process. This line of research spans controllable generation [12], training and imputation with missing values [11], and task-aware data generation [10], unified by the goal of enabling diffusion models to produce data that are not only realistic but also useful for inference, prediction, and decision-making.

Controllable generation. Diffusion models have achieved remarkable success in data generation across multiple modalities, including images, text, tabular, and time-series data. With the emergence of large-scale pretrained diffusion models trained on generic datasets, there is growing interest in how to steer and leverage these models to generate data that follow desired distributions or preferences with minimal computational effort. My research aims to address this challenge in a statistically principled manner by developing transfer-learning frameworks that adapt pretrained diffusion models without fine-tuning, enabling controllable and preference-aware generation while maintaining efficiency and theoretical guarantees [12, 14].

Missing value. Data incompleteness is ubiquitous in real-world applications—ranging from electronic health records to finance and sensor systems—and poses serious challenges for model training and inference. Diffusion models, however, are typically trained on complete data and are not directly applicable when missing values are present. My research leverages diffusion models to address missingness from two perspectives. First, we develop a training pipeline that directly trains diffusion models from partially observed data, without requiring prior imputation or discarding incomplete samples. Second, the trained diffusion model enables a generate-for-impute procedure, which can generate multiple plausible completions for the missing entries and naturally support uncertainty quantification in downstream tasks [11].

Task-aware data generation. While diffusion models are typically designed to replicate the training data distribution, many real-world applications require generated data to support specific decision-making tasks rather than merely mimic observed samples. In light of this, my research incorporates downstream objectives directly into the generation process to enable task-informed synthetic data generation. My research takes robust classification as a preliminary but representative downstream task—since it typically requires substantially more training data and benefits from synthetic data augmentation—and develops the Contrastive-Guided Diffusion Process [10]. This framework integrates contrastive learning into the diffusion process to produce synthetic data that are both realistic and optimized for the downstream task, establishing a principled bridge between data generation and downstream inference.

Future Directions

Looking ahead, I aim to advance the development of modern algorithms for *sequential data analysis* and *data generation*, which are foundational to machine learning and AI technologies. My future research will address key challenges in ensuring that these learning algorithms are *robust*, *privacy-perserving*, *interpretable*, and *effective in practice*.

Building upon my current work, I plan to (i) develop robust inference methods that guarantee stability under more broad and practical distributional uncertainties while being less conservative in the worst-case sense; (ii) establish a systematic framework for private change detection that accommodates diverse privacy requirements, distributional assumptions, and network structures; and (iii) apply these algorithmic developments to large-scale spatio-temporal data in real-world applications. In parallel, I aim to advance data generation techniques for multi-modal data—such as text, time series, and tabular data—and to improve both the fine-tuning and inference efficiency of synthetic data generation to better align with practical analytical and decision-making needs. Beyond methodological contributions, I am actively seeking applications of these developments in critical domains, including power systems for online reliability monitoring, spatio-temporal monitoring of natural hazards such as wildfires and earthquakes, and mobile health, where modeling and generating electronic health record data can enhance personalized and reliable medical decision-making.

References

[1] Rui Gao, **Liyan Xie**, Yao Xie, and Huan Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

[2] Anatoli Juditsky, Arkadi Nemirovski, **Liyan Xie**, and Yao Xie. Convex parameter recovery for interacting marked processes. *IEEE Journal on Selected Areas in Information Theory*, 1(3):799–813, 2020.

[3] **Liyan Xie**. Minimax robust quickest change detection using Wasserstein ambiguity sets. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1909–1914, 2022.

[4] **Liyan Xie**, Yuchen Liang, and Venugopal V Veeravalli. Distributionally robust quickest change detection using Wasserstein uncertainty sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1063–1071. PMLR, 2024.

[5] **Liyan Xie**, George V. Moustakides, and Yao Xie. Window-limited CUSUM for sequential change detection. *IEEE Transactions on Information Theory*, 69(9):5990–6005, 2023.

[6] **Liyan Xie** and Yao Xie. Sequential change detection by optimal weighted ℓ_2 divergence. *IEEE Journal on Selected Areas in Information Theory*, 2(2):747–761, 2021.

[7] **Liyan Xie**, Yao Xie, and George V. Moustakides. Asynchronous multi-sensor change-point detection for seismic tremors. In *IEEE International Symposium on Information Theory (ISIT)*, pages 787–791, 2019.

[8] **Liyan Xie**, Yao Xie, and George V. Moustakides. Sequential subspace change point detection. *Sequential Analysis*, 39(3):307–335, 2020.

[9] **Liyan Xie** and Ruizhi Zhang. Sequential change detection with differential privacy. *arXiv preprint arXiv:2509.02768*, 2025.

[10] Yidong Ouyang, **Liyan Xie**, and Guang Cheng. Improving adversarial robustness through the contrastive-guided diffusion process. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 26699–26723, 2023.

[11] Yidong Ouyang, **Liyan Xie**, Chongxuan Li, and Guang Cheng. Missdiff: Training diffusion models on tabular data with missing values. *arXiv preprint arXiv:2307.00467*, 2023.

[12] Yidong Ouyang, **Liyan Xie**, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. *Advances in Neural Information Processing Systems*, 37:136962–136989, 2024.

[13] **Liyan Xie**, Shaofeng Zou, Yao Xie, and Venugopal V Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

[14] Zhengyan Wan, Yidong Ouyang, **Liyan Xie**, Fang Fang, Hongyuan Zha, and Guang Cheng. Discrete guidance matching: Exact guidance for discrete flow matching, 2025.

[15] Haoyun Wang, **Liyan Xie**, Alex Cuozzo, Simon Mak, and Yao Xie. Uncertainty quantification for inferring Hawkes networks. *Advances in Neural Information Processing Systems*, 33:7125–7134, 2020.

[16] Haoyun Wang, **Liyan Xie**, Yao Xie, Alex Cuozzo, and Simon Mak. Sequential change-point detection for mutually exciting point processes. *Technometrics*, 65(1):44–56, 2023.

[17] Yiran Yang and **Liyan Xie**. Sequential Wasserstein uncertainty sets for minimax robust online change detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9491–9495, 2024.

[18] Minghe Zhang, **Liyan Xie**, and Yao Xie. Spectral CUSUM for online network structure change detection. *IEEE Transactions on Information Theory*, 69(7):4691–4707, 2023.

[19] Xiaojun Zheng, Simon Mak, **Liyan Xie**, and Yao Xie. PERCEPT: A new online change-point detection method using topological data analysis. *Technometrics*, 65(2):162–178, 2023.

[20] Shixiang Zhu, Alexander Bukharin, **Liyan Xie**, Mauricio Santillana, Shihao Yang, and Yao Xie. High-resolution spatio-temporal model for county-level COVID-19 activity in the us. *ACM Transactions on Management Information Systems (TMIS)*, 12(4):1–20, 2021.

[21] Shixiang Zhu, Alexander Bukharin, **Liyan Xie**, Khurram Yamin, Shihao Yang, Pinar Keskinocak, and Yao Xie. Early detection of COVID-19 hotspots using spatio-temporal data. *IEEE Journal of Selected Topics in Signal Processing*, 16(2):250–260, 2022.

[22] Shixiang Zhu, **Liyan Xie**, Minghe Zhang, Rui Gao, and Yao Xie. Distributionally robust weighted k-nearest neighbors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 29088–29100, 2022.