SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

Canyi Chen
Postdoctoral Research Fellow
Department of Biostatistics, University of Michigan
Ann Arbor, MI 48109, USA
canyic@umich.edu | (+86) 130-5151-7009
https://canyi-chen.github.io

November 22, 2025

Faculty Search Committee
Department of Statistics and Data Science
Tsinghua University
Haidian District, Beijing 100084, P. R. China

Dear Members of the Faculty Search Committee,

I am writing to apply for the tenure-track Assistant Professor position in the Department of Statistics and Data Science at Tsinghua University, beginning in the 2026–2027 academic year. I am currently a Postdoctoral Research Fellow at the University of Michigan, working with Dr. Peter Song, and I earned my Ph.D. in Statistics from Renmin University of China in June 2023, supervised by Dr. Liping Zhu. My research in *distributed statistical learning*, *causal mediation analysis*, and *GenAI-enhanced inference* develops statistically principled methodologies that ensure reliable inference in large-scale, heterogeneous, and complex data environments. These contributions I believe would be great fit with Tsinghua's strategic vision at the confluence of statistical methodology, optimization, AI, and interdisciplinary data science.

My research agenda centers on building statistically grounded frameworks that enable robustness, efficiency, and inferential validity for complex and large-scale data systems. To date, I have produced 30 manuscripts, including 16 peer-reviewed articles in leading statistics journals such as *JASA*, *JCGS*, and *Statistica Sinica*, with additional work currently under review in *Biometrika* and *Biometrics*. In *distributed statistical learning*, I demonstrate that robust inference, communication efficiency, and optimization scalability can be simultaneously achieved in decentralized and adversarial environments. Through the integration of kernel smoothing, robust estimation, and nonsmooth optimization, I have developed algorithms that preserve statistical guarantees while accommodating data heterogeneity and contamination. In *causal mediation analysis*, I have introduced calibrated testing procedures and dependence-aware FDR control methods that enable valid inference on mediation pathways under complex dependence structures.

I believe my research strongly resonates with Tsinghua University's mission to advance foundational statistics in support of national priorities in AI and data science. In particular, my emerging work on *GenAI-enhanced inference*—developing principled stopping criteria for inference safeguards and enhancement—would complement the department's growing strengths in the era of LLMs. My long-term vision is to advance optimization in statistics while developing the rigorous methodological toolbox needed for integrating GenAI into statistical science.

I am deeply committed to teaching and mentorship. I have served as a teaching assistant for Ph.D.-level courses in data science computing and asymptotic statistics, where I emphasized both theoretical foundations and algorithmic implementation. As a research mentor, I have guided junior scholars whose projects have culminated in publications in *JCGS* and *Statistica Sinica*. I am prepared to teach undergraduate courses in probability and mathematical statistics, as well as graduate seminars on distributed learning, causal inference, and high-dimensional inference.

I am committed to academic service through open-source software contributions (e.g., the `SIT` and `abima` R packages) and active referee work for journals in statistics and data science. Enclosed are my curriculum vitae, research statement, and teaching statement. I would be honored to further discuss how I can contribute to Tsinghua University's continued excellence in statistical science.

Sincerely,

Canyi Chen

Canyi Chen
Postdoctoral Research Fellow
University of Michigan

Branding Statement

Brand

I am a statistical methodology researcher dedicated to advancing statistical theory, methodology, algorithms, and software for complex data structures. My work focuses on causal mediation analysis, distributed data analytics, and GenAI-enhanced inference.

General Problem

Modern data infrastructures have evolved beyond centralized and transparent settings. Data are often geographically distributed, causal pathways are embedded in directed acyclic graphs, and an increasing portion of data is generated by opaque AI "black-box" systems. Classical statistical methods, originally developed for centralized, transparent, and static datasets, are therefore inadequate for delivering valid inference in these emerging environments.

Specific Problem

To ensure the reliability of modern data systems, statistical methods must be capable of:

- Learning from *distributed* and potentially adversarial data without centralized aggregation, while preserving communication efficiency.
- Uncovering complex *causal pathways* with rigorous control of type I error, particularly in settings where standard hypothesis tests fail to provide calibrated inference.
- Validating and leveraging *GenAI*-generated outputs (e.g., synthetic data) without compromising statistical integrity, thereby enabling safeguarded inference.

Achievement

I have developed an integrated set of statistical frameworks that address these challenges, resulting in *over 30 manuscripts*, including 16 peer-reviewed articles published in *JASA*, *JCGS*, and *Statistica Sinica*. My research establishes that *oracle-efficient inference* is attainable in distributed and adversarial environments, *calibrated type I error control* can be achieved for mediator discovery under composite null hypotheses, and *optimal stopping criteria* for GenAI training can be rigorously characterized to enhance statistical power when incorporating synthetic data. These advances bridge the longstanding gap between theoretical rigor and modern computational practice.

Vision

My long-term vision is to establish a *unified statistical foundation for trustworthy data science*. I aim to develop optimality theory for privacy-constrained distributed learning and to design rigorous *statistical safeguards* that enable the safe integration of Generative AI into high-stakes scientific and decision-making pipelines.

References

Chen, C., L. Zhou, and P. X.-K. Song. *Taking the Power of Distributed Systems: A Sequential Combination Method to Testing Joint Significance*. In preparation, 2025.

Research Statement

Keywords: Distributed Learning, Causal Mediation Analytics, and GenAI-Enhanced Inference.

My research is motivated by the inherent challenges of modern data infrastructures: datasets are massive and geographically distributed, dependence structures are complex, and increasingly, data are generated or transformed by opaque AI "black boxes." Classical statistical methods—developed for centralized and transparent environments—often break down under these conditions. *I develop statistical theory and algorithms that enable reliable inference in large-scale, decentralized, and AI-augmented systems.*

This research agenda has resulted in *over 30 manuscripts*, including *16 peer-reviewed publications* in leading journals such as *JASA*, *JCGS*, and *Statistica Sinica.* My work is organized around three interconnected themes that bridge foundational statistical theory with modern computational environments:

1. *Efficient Distributed Learning:* Addressing communication, statistical efficiency, and computation constraints in federated and multi-site massive data settings.
2. *Calibrated Mediation Analytics:* Developing valid methodologies for testing causal pathways under the composite null and controlling FDR within high-dimensional, complex dependence structures.
3. *GenAI-Enhanced Inference:* Establishing statistical safeguards and enhancement principles for inference using synthetic data and LLM-driven workflows.

R.1 Efficient Distributed Learning in Heterogeneous Environments

Modern collaborations in health and technology often collect data across multiple institutions where centralizing data is impossible due to privacy or bandwidth constraints. The fundamental challenge is designing estimators that approach the efficiency of centralized "oracle" procedures while minimizing communication.

Communication-Efficient Estimation via Linearization and Quadratic Approximation. A key insight of my work is that naive averaging in distributed networks introduces bias when local sample sizes or model specifications differ. The native meta approach is hence not optimal in large sytems. In work published in *JASA* and *Statistica Sinica*, I developed Linearization and Quadratic Approximation strategies that surrogate the impossible full-sample loss with feasible approximations. I demonstrated that, under mild regularity conditions, a finite round of communication can recover the first-order efficiency of a centralized estimator. (C. Chen et al., *Taking the Power of Distributed Systems: A Sequential Combination Method to Testing Joint Significance*, In preparation, 2025)

Robustness to Adversarial Contamination. Beyond efficiency, distributed systems are vulnerable to "Byzantine" failures or data corruption at specific nodes. Addressing "doubly nonsmooth" objectives, I introduced smoothing strategies for composite quantile regression and support vector machines. This approach, detailed in *Statistics and Computing* and *JASA*, enables gradient-based algorithms that achieve linear convergence and finite-sample error guarantees even under heavy tails, outliers and Byzantine attacks.

R.2 Calibrated Testing in High-Dimensional Mediation

While R.1 focuses on *how* to learn efficiently from data, R.2 investigates *what* can be inferred about underlying causal mechanisms. In modern scientific studies, such as those involving omics or digital biomarkers, discovering mediators within the directed acyclic graphs (DAGs) is a central objective. However, classical testing procedures are often overly conservative when addressing the *composite null* problem, defined as $H_{0j} : \alpha_j \beta_j = 0$, because failure of either the exposure–mediator or mediator–outcome pathway implies the absence of mediation. Moreover, strong dependence among candidate mediators further undermines the validity of standard FDR control procedures.

High-Sensitivity Joint Significance Testing. I have developed calibrated testing procedures that explicitly respect the composite null structure while maintaining computational efficiency. By employing cross-fitting strategies, these methods remain valid under weak signals and thereby achieve greater power than existing approaches. I further introduced new FDR control procedures that account for dependence among mediators, ensuring accurate FDR control in high-dimensional and general dependence settings. These works are currently under review at *Biometrika*.

Copula and Quantile-Based Frameworks. To handle the non-Gaussian data commonly encountered in digital health applications, I developed copula-based frameworks that decouple marginal distributions from dependence structures. This allows the modeling of heavy-tailed and skewed data, and extends mediation analysis beyond mean effects to quantile-level inference, thereby capturing heterogeneous mediation effects across the outcome distribution. These results are published in the *Canadian Journal of Statistics* or are under review at *Statistica Sinica*.

R.3 Vision: Statistical Safeguards and Enhancements for the GenAI Era

While my previous work has optimized inference for distributed and naturally occurring data, the next frontier lies in understanding and utilizing data generated by opaque algorithms. Large Language Models (LLMs) and other generative systems are increasingly used to produce synthetic data and predictive representations, yet they lack inherent uncertainty quantification and may introduce subtle biases. *My long-term vision is to* *establish the statistical foundations that enable trustworthy and effective use of GenAI in inference.*

Stopping Rules and Stability. In my first work, I established statistically principled stopping rules for GenAI training based on sequential monitoring. These rules prevent overfitting in high-capacity models and provide safeguards when generative models are integrated into high-stakes policy or decision-making pipelines. This work is published in *JASA*.

Principled Integration of Synthetic Data. A key question guiding my future research is: *Under what conditions can synthetic data enhance classical inference without sacrificing validity?* I am developing two-step estimators that leverage synthetic data either as control variates or for informative initialization. By treating synthetic outputs as auxiliary statistics, these methods deliver efficiency gains while rigorously preserving asymptotic normality and valid inferential guarantees.

The Unified Framework. Ultimately, I aim to connect these three research directions: building distributed systems (R.1) that support complex causal discovery (R.2), and augmenting them with statistically safeguarded generative modeling (R.3). This integrative vision aligns with the interdisciplinary strengths of Tsinghua's Department of Statistics and Data Science, where I look forward to collaborating across optimization, causal inference, and artificial intelligence to advance the foundations of trustworthy data science.

Selected References: [1] C. Chen et al., *JASA* (2025). [2] C. Chen et al., *Statistica Sinica* (2023). [3] N. Qiao, C. Chen et al., *Stats & Comp* (2025). [4] C. Chen et al., *Biometrika* (In Review).

References

Chen, C., L. Zhou, and P. X.-K. Song. *Taking the Power of Distributed Systems: A Sequential Combination Method to Testing Joint Significance.* In preparation, 2025.

Teaching Statement

My goal as an educator is to cultivate critical and independent thinkers who regard statistics not merely as a set of formulas, but as a principled framework for reasoning under uncertainty. At Tsinghua University, I aspire to contribute to a curriculum that prepares students to lead in an era defined by the convergence of statistics, optimization, and artificial intelligence. My teaching philosophy emphasizes that rigorous theoretical understanding must be complemented by 'from-scratch" algorithmic implementation, enabling students to peer inside the 'black box" of modern statistical and AI methodologies.

Teaching Philosophy: Bridging Theory, Code, and Critical Thinking

1. Intuition Before Rigor. Students develop lasting understanding when they grasp the *why* before the *how*. Abstract results become meaningful when anchored in statistical intuition. For example, when introducing likelihood-based inference, I begin with the information-theoretic rationale for modeling before transitioning to asymptotic theory. Presenting convergence not merely as a technical requirement, but as a guarantee of reliability, helps students recognize the purpose and value of formal proofs.

2. Integration of Theory and Computation. A core principle of my pedagogy is that statisticians must understand methods at the algorithmic level. In introductory courses, I deliberately avoid "black-box" libraries and require students to implement fundamental procedures, such as BFGS and Newton–Raphson, from scratch. This approach exposes subtle failure modes—such as instability due to poor initialization or sensitivity to outliers—that remain hidden when relying solely on pre-packaged software.

3. Critical Judgment in the GenAI Era. Rather than prohibiting LLMs, I position them as subjects of statistical evaluation. I emphasize that AI-generated code or analysis should be treated as a *hypothesis* requiring scrutiny. I design assignments in which students must audit AI-generated statistical outputs, identify flawed assumptions, and correct reasoning errors. This reinforces the principle that while AI may expedite computation, the responsibility for inference ultimately rests with the statistician.

Teaching and Mentoring Experience

My teaching philosophy has been shaped through service as a Teaching Assistant for Ph.D.-level courses in *Data Science Computing*, *Asymptotic Statistics*, and *Natural Language Processing*.

Case Study: Data Science Computing. In the Ph.D.-level course *Computer Skills in Data Science*, I noticed that while students were comfortable applying Python libraries, they struggled to diagnose convergence issues in non-standard models. To address this, I designed a lab module in which students implemented robust regression techniques from first principles. Using a dataset contaminated with heavy-tailed noise, I demonstrated the failure of ordinary least squares and guided students to implement the Huber loss and iteratively reweighted least squares. This exercise provided a concrete understanding of how robustness is achieved at the algorithmic level.

Case Study: Asymptotic Statistics. In *Asymptotic Statistics*, students often found it challenging to connect theoretical limit theorems with finite-sample behavior. I led discussion sessions where we used simulation to stress-test asymptotic guarantees. By generating data from both symmetric and skewed distributions, we visualized how slowly the Central Limit Theorem manifests under heavy skewness. These simulations helped students appreciate both the power and the limitations of the asymptotic theory they were proving.

Research Mentoring. Outside the classroom, I view mentoring as an individualized extension of teaching. I have supervised several junior researchers through the full research cycle—from formulating research questions and conducting literature reviews to designing simulations and developing theoretical arguments. These collaborations have resulted in publications in outlets such as *JCGS* and *Statistica Sinica*. I am committed to bringing this same mentorship-driven approach to supporting graduate students at Tsinghua.

Teaching Interests at Tsinghua

I am prepared to teach core undergraduate courses such as *Probability*, *Mathematical Statistics*, and *Regression Analysis*. At the graduate level, I am eager to develop advanced courses that align with both my research expertise and the department's strategic priorities:

- *Distributed Learning & Optimization:* A course on divide-and-conquer estimators, communication-efficient algorithms, and privacy-preserving inference, addressing the computational challenges of modern large-scale data.

- *High-Dimensional Mediation Analysis:* A seminar focused on causal mechanisms in complex systems, with emphasis on composite null testing, copula structural equation models, and applications in genomics and digital health.

- *Trustworthy Inference with GenAI:* A forward-looking course on the statistical foundations of synthetic data, optimal stopping rules for AI training, and safeguards for AI-augmented decision-making.

Conclusion

I am excited about the opportunity to join the Department of Statistics and Data Science at Tsinghua University. I look forward to fostering a learning environment in which students are challenged to become both rigorous theoreticians and capable computational scientists, equipped to address the data-driven challenges of the future.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.