Research Statement:
Mathematical Foundations of Data Science

Today’s data deluge is driving a transformation of the world-economy and even the modern way of life. Cutting edge data science challenges are pushing the boundaries of what can be learned from data, giving rise to an array of applications in artificial intelligence, medicine, and social sciences. My past work has provided several major contributions at the front of mathematical statistics and information theory, addressing some of the fundamental theoretical aspects of data science.

Among these contributions, my Ph.D. work addressed the effects of data compression on real-world analog data. It provided the first complete characterization of the optimal trade-offs between sampling rate, compression bitrate, and distortion in processing real-world signals digitally. My postdoc work focuses on challenging high-dimensional signal estimation and classification settings. The high-dimensionality here is typically the result of a vast number of features available for inference. These features can represent raw measurements, the result of compression or dimensionality reduction procedures, or the output of data-driven techniques like a neural network’s activations.

Arguably, the most pressing issues of modern data-driven inference techniques are their general inability to explain models’ decisions and to generalize to unexpected new situations. Feature-based inference techniques are perfectly suited for resolving this ‘interpretability’ and ‘generalizability’ crises. Interpretability is attained by crafting many meaningful features, while generalizability is attained by using optimal feature selection and inference rules for these features. For this reason, we are witnessing a shift in machine learning and artificial intelligence communities from end-to-end to feature-based approaches.

The theoretical tools to handle feature-based inference include multiple hypothesis testing and variable selection in statistics and geometric understanding of high-dimensional distributions from information theory. My research integrates these two disciplines with ambitious computations, validating theoretical results and assessing inference’s performance on real-world datasets.

In what follows, I will review in more detail my past achievements and describe my short- and long-term future research plans.

Past Research

Discriminating High-dimensional Distributions

Two-sample testing and the Higher Criticism

Testing whether two datasets are simply different realizations of the same underlying high-dimensional probabilistic mechanism is at the heart of many modern data science applications. Concrete examples include: detecting language changes in texts, detecting narrowband signals over a wide spectrum, and early detection of anomalous behavior in biological and economic systems. It is not surprising, therefore, that the aforementioned problem has been studied extensively in recent years across several fields. Especially in statistics, theoretical computer science, and information theory.

My work takes a novel approach to this problem that typically outperforms, both theoretically and empirically, all previously known ones. Specifically, my approach views the two
probabilistic mechanisms as two high-dimensional distributions over a vast set of features. It then tests for the significance of any feature individually, as opposed to classical approaches that directly operate over the entire data. Finally, my approach uses the Higher Criticism (HC) to combine the information from individual features to a global test against the null hypothesis stating that the two underlying distributions are identical. HC is a well-studied tool in multiple hypothesis testing and feature selection in statistics. The underlying mechanism driving the success of the method is the extreme sensitivity of HC to deviations from a Brownian Bridge process over \([0,1]\), as this process describes the empirical distribution of the P-values summarizing the test of each feature.

I analyzed the theoretical performance of the HC-based test in a particularly challenging setting where differences between the two distributions, if exist, are concentrated in a small and unknown subset of features. Meanwhile, no single feature can provide decisive evidence against the null hypothesis of identical distributions. In other words, discriminating evidence is rare and weak. One of the main results of this analysis is an asymptotic characterization of the test’s power throughout a phase space whose coordinates control the rarity/intensity of differences under the alternative hypothesis. As it turns out, the test displays a sharp phase transition. My work derived formulas for a curve partitioning the phase space such that on one side of this curve the test is asymptotically fully powered and asymptotically powerless on the other side. This is illustrated in Figure 1. This theoretic result was validated using a massive amount of computations.

Both empirical and theoretical analyses show that the phase transition curve of HC is superior to all previously studied tests in the rare/weak setting, such as Pearson’s chi-squared, Bonferroni, and inference based on false-discovery rate-controlling; see Figure 1. Furthermore, the method provides perfect interpretability in terms of the features it uses. Specifically, as a by-product, the HC calculation selects a subset of discriminating features that comprise the best evidence against the null hypothesis.

One model incorporating all rare/weak feature settings

Rare/weak feature models were studied extensively in the past two decades, providing an array of fruitful insights for the limitation of signal detection, feature selection, and classification. In recent work, I have shown that all previously studied rare/weak feature models share the same fundamental structure: under the alternative hypothesis, each feature’s P-value is approximately a sparse mixture of an exponential distribution and a scaled non-central log-chi-squared distribution. This discovery implies that the study of rare/weak feature models is reduced to characterizing only two parameters, i.e., the scaling and non-centrality. For example, appropriate scaling corresponds to heteroscedasticity, while the one- and two-sample
settings differ in their non-centrality parameter. Consequently, my work delivers asymptotic phase transition formulas for many new situations.

**Testing authorship without handcrafting or tuning**

Another aspect of my work used a series of ambitious computations involving real-world data to demonstrate the effectiveness of my HC-based test in text classification and, in particular, in authorship attribution challenges. Namely, considering two word-frequency tables associated with two documents, my work proposes that the test scores are high when the two documents are different in terms of authorship. In this setting, the evidence against the null hypothesis of homogeneous authorship is thought to lay within the subset of words identified by the aforementioned feature selection mechanism of HC.

Real-data performance evaluations show that the method is incredibly effective as an unsupervised, untrained discriminator. Specifically, we compared the performance of our method using the largest available copyright-free library of literary works. Our method outperformed all unsupervised discriminator we tried, as-well-as the most popular supervised classification techniques like naive Bayes, logistic regression, and neural networks.

The empirical success of our method is somewhat surprising, as topic structure within real-world textual data typically violates the idealized two-sample multinomial model. Fortunately, HC’s feature selection mechanism explains this success. Massive amounts of evaluations using real data imply that, in practice, the words HC weighs mostly have low variance across documents belonging to a corpus of homogeneous authorship. Since topic-related words are associated with large within-corpus variance, this empirical discovery says that our method is mostly affected by author-related words and is relatively unaffected by topic structure.

**Inference from Compressed Data**

Due to the disproportionate size of modern datasets, the bottleneck for reliable inference has shifted from the availability of raw data to the available number of bits to communicate or process the data. Consequently, a growing number of researches are dedicated to inference techniques that are applied to a compressed representation of the data or extracted features, rather than the data itself. These situations are different from classical settings in that the overall performance is also determined by the number of bits available to the estimator, and not only by the quality of the features (sample size and signal-to-noise ratio).

In this context, Galen Reeves and I provided exact asymptotic characterization of the performance of any well-behaved inference procedure from data undergoing lossy compression. Our method combines distribution approximation techniques from mathematical transport theory with classical concentration results of random points on spheres. As a consequence, our work delivers, for the first time, optimal tradeoffs between bitrate, noise, and expected quadratic risk in sparse regression and compressed sensing operating under bitrate constraints.

**Future Research**

My recent postdoctoral work considered various high-dimensional inference settings at the front of modern data science and information theory. Most notably, this work showed that adopting a feature-centered approach to the long-standing problem of discriminating distributions yields interpretable and useful tools for analyzing high-dimensional signals.
My long term plan is to provide a unified framework for discriminating high-dimensional datasets and signals that incorporates most important properties of the data. Such properties include rare/weak discriminating evidence, large or small sample, discrete or continuous data, and inter-feature dependency. The most striking result of this unified treatment is inference and feature selection procedures that can be applied to features obtained using existing data-driven techniques (e.g., neural networks, decision trees, language models). Such inference and feature selection procedures can alleviate the current ‘explainability’ and ‘generalizability’ issues of these data-driven approaches, as these approaches are interpretable and robust as the feature they rely upon.

Recent works of myself and others have already resolved many of the challenges in achieving the aforementioned vision. Below I describe several remaining gaps that I intend to pursue as short-term milestones.

**Testing the closeness of distributions with sparse data**

Consider two datasets sampled from possibly distinct discrete distributions over many features (categories). The case where the number of samples is small compared to the number of features is referred to as the *sparse* case, as many of the features have zero counts. Sparse data arise in a host of applications such as the analysis of store sales data, the spread of diseases, and single-photon detection using sensor arrays. Nevertheless, theoretical works in this area were largely focused on the non-sparse case; the fundamental limits of inference from sparse data are little understood.

This research direction aims to characterize the limitations of hypothesis testing and estimation in the sparse feature regime. As in the non-sparse case, it is anticipated that such understanding would lead to optimal discriminators and feature selection procedures.

**Multiple testing and selection of dependent features**

Feature independence is arguably one of the most limiting assumptions of multiple comparison techniques. The violation of this assumption in real-world data is the norm rather than the exception. Examples include dependencies between word occurrences in text analysis and dependencies between artificial neural networks’ activations. Furthermore, feature dependency may be present by design, as in certain communication schemes involving multiple receivers or in experiments where the number of features is larger than the size of the data ($p > n$).

The goal of this research direction is to extend feature selection methodologies to allow weak or strong feature dependency. Although this extension has been a long-standing challenge in many disciplines, my recent work on the Higher Criticism provides a promising new direction to address this challenge. Specifically, I plan to exploit recent advances in the study of empirical processes of correlated random variables to characterize the theoretical behavior of optimal feature selection rules. In practice, it is required to characterize the type of feature dependency present in specific types of data. I plan to carry an array of ambitious computations involving real data to achieve this characterization.

**Feature Selection for Transfer Learning**

Data-driven techniques like convolutional neural networks (CNN) have impressive track records in a host of challenges. They are frequently viewed as automatic machines for extracting features from data. When data are scarce, however, the end-to-end training approach for classification and regression used by CNN usually leads to poor results. Therefore, a by-now standard practice is to use the last layer of a CNN trained using training data for
one problem as a set of features for another problem, for which training data is scarce. Arguably, this “transfer learning” scenario is one of the most widely-used and little-understood practices.

The goal of this research project is to understand the limitations of transfer learning through an ideal high-dimensional model feature like the rare/weak feature model. This project would combine theoretical analysis of the proposed models with ambitious data science evaluations to verify their validity.

Summary

My past work provided several decisive contributions to mathematical statistics and information theory. Building on these contributions, I sought to provide a general framework for feature-centered high-dimensional inference that can be applied to most types of signals, including those resulting from data-driven techniques. This framework would be pivotal to the ongoing academic effort to formalize and develop data science as a subject of study and research activity.