Harnessing the Power of Latent Structure Models and Modern Big Data learning

Jiahua Chen (University of British Columbia),
Yingying Fan (University of Souther California),
Tracy Ke (Princeton Unversity),
Stanislav Volgushev (University of Toronto)

12/10/2023 – 12/15/2023

1 Content and topics covered

The aim of the workshop was to bring together participants from different areas related to latent structure models and more broadly, big data learning and inference. To this end, we invited participants from a wide variety of backgrounds who presented work on a range of topics related to latent structure models and big data learning.

Below, we highlight some of the main themes that appeared throughout the workshop. Additional details on the talks in each theme are given in the last section of this document.

1.1 Latent structures and variables

An important class of questions around latent variables deals with cases where the main interest is unobserved measure of quality or innate ability. This could include the ability of a worker, the quality of a scientific publication, a product or a movie. an important objective is to provide rankings based from observational data and quantify the surrounding uncertainty. A classical model in this setting is the Bradley-Terry-Luce model and extensions thereof. The observed data correspond to the top choices in sets of $M \geq 2$ randomly selected items, where the choice probability is governed by latent quality parameters of each unit. The classical version of this model corresponds to pair-wise comparisons ($M = 2$). In such models, **Yuxin Tao** focused on cases where the unobserved latent quality distribution falls into several clusters. A central problem is to identify which units are from the same cluster and rank the clusters. The problem of estimating the latent parameters with strong theoretical guarantees and inference on rankings in multiway comparisons where $M \geq 2$ was considered in the talk by **Jianqing Fan**. An overview of results in such models was provided by **Wenguang Sun**.

A completely different type of latent model was considered by **Ji Zhu** who proposed a new class of latent models for hyper-graphs. In contrast to classical graphs, which have so far been the main focus of statisticians, hyper-edges in hyper-graphs link multiple nodes. This allows to model more complex relationships between units but is a far less understood problem and comes with new theoretical challenges.

Another large class of models that can be considered as having latent structures in a wide sense is given by change-point models. Here, observations are ordered by time and the latent structure comes into play through assuming that population quantities are constant over certain time intervals. Key questions are detecting

whether there is more than one latent group, or, in classical time series language, whether there is a change-point. **Lixing Zhu** presented novel methodology for solving this problem when data are best modeled by tensors but the slices of the tensors can be heterogeneous.

1.2 Mixture models

While mixture models could be viewed as a class of models for population made of some latent groups, the workshop had a particularly rich set of talks on this topic which warrants a separate discussion. Within the class of mixture models, several themes appeared repeatedly.

Due to their non-standard statistical properties, even the most basic class of *parametric finite mixture models* still poses interesting open questions. One classical question in this direction pertains to obtaining estimators in general finite mixture models that have optimal rates of convergence. **Yun Wei** presented a new general class of estimators that incorporates existing ones which are known to have optimal convergence rates (such as method of moments and minimum KS-distance estimators) as special cases. This talk also discussed finite mixture models with repeated measurements on each component.

A topic that has recently gained substantial attention in other areas of statistics is the problem of aggregating estimators from several sources to obtain an overall model. This could be due to computational bottlenecks for large scale data sets or due to decentralized storage of different portions of the data. For finite mixture models, this problem remained open until recently and a first systematic approach with theoretical guarantees was presented by **Qiong Zhang**.

An application of finite mixture models for classification and outlier detection under covariate and label shifts was discussed by **Yukun Liu**.

Finite Gaussian mixture models also have connections to discriminant analysis in classification, where the working model for the observed data (X, Y) is that the distribution of features X given class k for the discrete outcome Y are multivariate normal. Minimax-optimal classification in such models under additional assumptions on the underlying Gaussians was discussed by **Marten Wegkamp**.

Florentina Bunea provided an axiomatic justification for using the Wasserstein distance to measure discrepancy between mixing distributions and discussed optimal estimation and inference in finite multinomial mixtures with applications to topic models.

One interesting extension of classical finite parametric mixtures aims at including *non-parametric components*. Without additional structure such models are not identifiable, but there are many types of additional assumptions that make identification and estimation possible. One such structure was presented by **Pengfei Li** who discussed methodology for non-parametrically estimating multivariate mixture models with a finite number of components under the additional assumption that the components of all vectors are independent. **Hajo Holzmann** reviewed results on mixture models with non-parametric components and discussed extensions of this setting to mixtures of regressions. Non-parametric identification and estimation of mixtures of regression models was also discussed by **Weixin Yao** who provided identification and estimation results for models where different components of the mixture distribution correspond to scaled versions of the same unspecified distribution.

A generalization of mixtures of regression models is given by the class of *mixtures of experts models* where the class weights can also depend on the predictors (in contrast, in mixtures of regressions, the class weights are independent of the covariates). **Nhat Ho** discussed new theoretical results regarding a special class of mixtures of experts models - Gaussian mixtures of experts with softmax gating functions. Sparse mixtures of experts with a possibly over-specified number of components were discussed by **Abbas Khalili** who focused on the computational and methodological challenges related to estimation and feature selection in such models.

1.3 Statistics involving complex objects in big data

An important challenge in big data learning is develop methodology for setting when the data are not Euclidean and classical approaches developed for R^d do not apply.

Paromita Dubey considered a classical statistical question, two-sample testing, in a framework where the *observations are random elements in metric spaces*. The talk highlighted the usefulness of depth notions in metric spaces and the challenges that surround their theoretical analysis. Observations in general spaces were

also the topic discussed in the talk by **Axel Munk** who considered extensions for the notion of dependency from random vectors to random variables taking values in general Polish spaces. The key technical tool in constructing scalar dependency measure for general objects is shown to be related to the notion of optimal transport in general spaces.

Nikolaos Ignatiadis cast the problem of multiple testing in an empirical Bayes framework. The main challenge here is that a part of the model is given by a prior which is estimated non-parametrically; this corresponds to a functional parameter.

Jiashun Jin considered citation counts and modeling of such data through constructing *co-citation networks* and modeling research interests of authors by estimating their memberships in those networks while allowing for time variation.

Tracy Ke considered statistical inference and for *text data*. Her talk highlighted the utility of multinomial distributions in understanding text data through modeling word counts. As specific application, two sample testing problems were considered.

The use of manifold methods and specifically utilizing the structure of the Stiefel manifold in the setting of high-dimensional multitask learning via sparse singular value decomposition of the regression coefficients was presented by **Jinchi Lv**. Manifolds also played a central role in the talk by **Zhigang Yao** who proposed a novel approach to learning latent manifold representations from noisy observations.

Ping Ma presented approaches to deal with *graph structured data*. The talk focused on subsampling methodology for large graphs that allows to retain information on smaller communities that are typically overlooked by existing approaches.

Xuening Zhu discussed approaches to modeling *matrix-valued time series* data with possibly missing entries. A key difference in the analysis of this talk to existing approaches lies in utilizing observed network structures for a more precise analysis.

1.4 Further Topics in Big data learning, inference, and beyond

Staying true to our objective of bringing together researchers from very different areas related to latent models and big data learning, we also had a number of talks that span a wide range of topics without a clearly being part of the three big themes covered above. The brief overview of additional topics covered during the workshop highlights the diverse backgrounds and interests of our participants.

Emre Demirkaya revisited a popular regression approach based on distributional nearest neighbours and proposed modifications that lead to reduced bias and improved convergence rates for smooth regression functions. Applications of latent models to studying the adoption of renewable energy in rural Nepal were given by **Jyoti Devkota**. A theoretical analysis of approximate knockoff procedures where the feature distribution of the classical knockoff approach is miss-specified or estimated was presented by **Lan Gao**. Connections between identification of structural parameters in dynamic logit models and truncated moment problems in mathematics were made by **Jiaying Gu**. **Adel Javanmard** studied tradeoffs between adversarial robustness and accuracy in latent structure models. Novel ensemble approaches to testing global null hypotheses for multivariate problems in whole genome sequencing were presented by **Xihong Lin**. **Wei Lin** discussed over-parametrization in two-layer networks and presented new results solidifying our understanding of double decent and the benefits of over-parametrization for prediction risk in such models.

Differential privacy was considered by **Po-Ling Loh**, who discussed the usage of noisy gradient descent algorithms to achieve differentially private inference in high-dimensional settings and by **Weijie Su** who demonstrated that statistical accuracy of existing approaches can be improved substantially by a more refined analysis of f-differential privacy under composition.

Computational aspects of Bayesian Additive regression trees and the amount of computation required to reach regions of high posterior densities were discussed by **Yan Shuo Tan**.

Xin Tong presented a new framework for optimal classification rules under equal opportunity constraints.

Estimation and inference for the problem of selecting predictor regions where the regression function exceeds minimal thresholds, with applications to health data and with minimax optimality guarantees was considered by **Richard Samworth**.

Geoffrey McLachlan discussed semi-supervised learning and provided theoretical explanations for the surprising phenomenon where estimators based on partially labeled data can be more efficient than estimators with access to full labels if the missing label mechanism is modeled sufficiently precisely.

Inference in complex survey data utilizing empirical likelihood and Neyman orthogonalization methodology was discussed by **Puying Zhao**.

Liping Zhu presented new methodology for constructing powerful tests for covariate effects in high-dimensional models where different covariates can exhibit varying degrees of variation.

2 Format and participant feedback

The format of the workshop was hybrid. All participants commented that everything worked perfectly on the technical side. Some online participants mentioned that it was difficult for them to get the most out of the workshop because they were not able to attend many of the talks due to the time difference. The positive side of the hybrid format, as pointed out by anther participant, was that it allowed participation from several leading experts in Europe and North America who would not have been able to attend this meeting in person. While the time difference remains an unavoidable challenge, the organizers agree that allowing for online participation enabled us to reach a wider range of participants and considers this format as a success, even in a world where international travel restrictions have been lifted.

Several participants, including both junior and senior researchers, commented that they were surprised by the breadth of topics related to latent structure models. Several participants mentioned explicitly that the workshop broadened their view of the field and gave them ideas for future research directions. Junior participants appreciated the chance to present their research to senior researchers, get insightful feedback and build their research network.

3 Abstracts of presented talks in alphabetical order of authors

Florentina Bunea *Inference for the Wasserstein distance between mixing measures in topic models*

The Wasserstein distance between mixing measures has come to occupy a central place in the statistical analysis of mixture models. We give the first axiomatic justification of its usage as a canonical measure of discrepancy between any mixture distributions. Inference for the Wasserstein distance between mixing measures is gener- ally difficult in high dimensions. Specializing to discrete mixtures arising from topic models, we offer the first minimax lower bound on estimating the distance between pairs of mixing measures in this model class. This reveals regimes under which fast estimation of the distance between mixing measures can be expected, even if the ambient dimension of the mixture distributions is large. In such regimes, we develop fully data-driven inferential tools that allow us to obtain the first asymptotically valid confidence intervals for the Wasserstein distance between mixing measures, in topic models. Our results apply to potentially sparse mixtures of potentially sparse high- dimensional discrete probability distributions, and are illustrated via an example on movie reviews from the IMDb data set.

Emre Demirkaya *Optimal Nonparametric Inference with Two-Scale Distributional Nearest Neighbors*

The weighted nearest neighbors (WNN) estimator has been popularly used as a flexible and easy-to-implement nonparametric tool for mean regression estimation. The bagging technique is an elegant way to form WNN estimators with weights automatically generated to the nearest neighbors (Steele, 2009; Biau et al., 2010); we name the resulting estimator as the distributional nearest neighbors (DNN) for easy reference. Yet, there is a lack of distributional results for such estimator, limiting its application to statistical inference. Moreover, when the mean regression function has higher-order smoothness, DNN does not achieve the optimal nonparametric convergence rate, mainly because of the bias issue. In this work, we provide an in-depth technical analysis of the DNN, based on which we suggest a bias reduction approach for the DNN estimator by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. The two-scale DNN estimator has an equivalent representation of WNN with weights admitting explicit forms and some being negative. We prove that, thanks to the use of negative weights, the two-scale DNN estimator enjoys the optimal nonparametric rate of convergence in estimating the regression function under the fourth-order smoothness condition. We further go beyond estimation and

establish that the DNN and two-scale DNN are both asymptotically normal as the subsampling scales and sample size diverge to infinity. For the practical implementation, we also provide variance estimators and a distribution estimator using the jackknife and bootstrap techniques for the two-scale DNN. These estimators can be exploited for constructing valid confidence intervals for nonparametric inference of the regression function. The theoretical results and appealing finite-sample performance of the suggested two-scale DNN method are illustrated with several simulation examples and a real data application.

Jyoti U. Devkota *Identification of latent structures in qualitative variables Examples from Renewable Energy users of Nepal*

This study is based on data collected from two sample surveys. They are namely survey of 300 households of national grid energy users and 400 households of biogas users. It was conducted in three different rural settings of Nepal. The responses to questions were classified into multiple choice options. This generated categorical data and reduced ambiguity and confusion between interviewer and interviewee. Such data were classified into ordinal scale and modeled. As the dependent variable had more than two categories, polytomous and not dichotomous models are developed and fitted. Ten different hypotheses assessing and measuring the energy consumption dynamics are tested. Values of parameters of these model and odds ratio are used in quantifying the impact of change with respect to energy consumption. The variables considered were namely time spent in the collection of firewood, type of house, amount of firewood saved, time saved, employer and school located within 15 min distance. Such data-based studies are very crucial for country like Nepal which lacks a strong backbone of accurate and regularly updated official records. These studies can be generalized to other countries of Asia and Africa. The results obtained can provide guidelines to policy makers and planners regarding formulation of realistic energy policies for such countries.

Paromita Dubey *Two Sample Inference for Random Objects using Depth Profiles*

In this talk I will describe a novel framework for two sample inference to distinguish between populations of random objects. The test statistic is based on the differences in the depth profiles of the observations with respect to their own population and with respect to a potentially different population. I will describe the asymptotic behavior of the test statistic under the null hypothesis of no differences across the populations and also under contiguous alternatives close to the null. A theoretically justified permutation approach is used to approximate the critical value in practice. The utility of the new test will be illustrated using a wide range of simulations for a large variety of metric spaces and multiple real data applications.

Jianqing Fan *Ranking Inferences Based on the Top Choice of Multiway Comparisons*

This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to $M = 2$. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both ℓ_2-norm and ℓ_∞-norm, with the mini- mum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multi- plier bootstrap. The estimated distributions are then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simula- tion studies lend further support to our theoretical results. A real data application illustrates the usefulness of the proposed methods convincingly. (Joint work with Zhipeng Lou, Weichen Wang, and Mengxin Yu)

Lan Gao *Robust Knockoffs Inference with Coupling*

We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family wise error rate (FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former

uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework.

Jiaying Gu *Identification of Dynamic Panel Logit Models with Fixed Effects*

We show that the identification problem for a class of dynamic panel logit models with fixed effects has a connection to the truncated moment problem in mathematics. We use this connection to show that the sharp identified set of the structural parameters is characterized by a set of moment equality and inequality conditions. This result provides sharp bounds in models where moment equality conditions do not exist or do not point identify the parameters. We also show that the sharp identified set of the non-parametric latent distribution of the fixed effects is characterized by a vector of its generalized moments, and that the number of moments grows linearly in T. This final result lets us point identify, or sharply bound specific classes of functionals, without solving an optimization problem with respect to the latent distribution. We illustrate our identification result with several examples, and an empirical application on modeling childrens respiratory conditions.

Nhat Ho *Demystifying Softmax Gating Gaussian Mixture of Experts*

Understanding parameter estimation of softmax gating Gaussian mixture of experts has remained a long-standing open problem in the literature. It is mainly due to three fundamental theoretical challenges associated with the softmax gating: (i) the identifiability only up to the translation of the parameters; (ii) the intrinsic interaction via partial differential equation between the softmax gating and the expert functions in Gaussian distribution; (iii) the complex dependence between the numerator and denominator of conditional density of softmax gating Gaussian mixture of experts. We resolve these challenges by proposing novel Vononoi loss functions among parameters and establishing the convergence rates of maximum likelihood estimator (MLE) for solving parameter estimation in these models. When the number of experts is unknown and overspecified, our findings show a connection between the rate of MLE and a solvability problem of a system of polynomial equations.

Hajo Holzmann *Mixture models and mixtures of regressions with nonparametric components*

Recently there has been some interest in mixture models and mixtures of regressions in which at least some components are not parametrically, but rather semi or nonparametrically specified.

In the first part of the talk we give an overview of the literature and in particular of some recent contributions to this subject.

Then we investigate in detail a flexible two-component semiparametric mixture of regressions model, in which one of the conditional component distributions of the response given the covariate is unknown but assumed symmetric about a location parameter, while the other is specified up to a scale parameter. The location and scale parameters together with the proportion are allowed to depend nonparametrically on covariates.

After settling identifiability, we provide local M-estimators for these parameters which converge in the sup-norm at the optimal rates over Hlder-smoothness classes. We also introduce an adaptive version of the estimators based on the Lepski-method.

We investigate the finite-sample behaviour of our method in a simulation study, and give an illustration to a real data set from bioinformatics.

Nikolaos Ignatiadis *Empirical partially Bayes multiple testing and compound χ^2 decisions*

We study multiple testing in the normal means problem with estimated variances that are shrunk through empirical Bayes methods. The situation is asymmetric in that a prior is posited for the nuisance parameters (variances) but not the primary parameters (means). If the prior were known, one could proceed by computing p-values conditional on sample variances; a strategy called partially Bayes inference by Sir David Cox. These conditional p-values satisfy a Tweedie-type formula and are approximated at nearly-parametric rates when the prior is estimated by nonparametric maximum likelihood. If the variances are in fact fixed, the approach retains type-I error guarantees.

Adel Javamand *Adversarial robustness for latent models: Revisiting the robust-standard accuracies tradeoff*

Over the past few years, several adversarial training methods have been proposed to improve the robustness of machine learning models against adversarial perturbations in the input. Despite remarkable progress in this regard, adversarial training is often observed to drop the standard test accuracy. This phenomenon has intrigued the research community to investigate the potential tradeoff between standard accuracy (a.k.a generalization) and robust accuracy (a.k.a robust generalization) as two performance measures. In this talk, we will revisit this tradeoff for latent models and argue that this tradeoff is mitigated when the data enjoys a low-dimensional structure. In particular, we consider binary classification under two data generative models, namely Gaussian mixture model and generalized linear model, where the features data lie on a low-dimensional manifold. We develop a theory to show that the low-dimensional manifold structure allows one to obtain models that are nearly optimal with respect to both, the standard accuracy and the robust accuracy measures. We further corroborate our theory with several numerical experiments, including the Mixture of Factor Analyzers (MFA) model trained on the MNIST dataset.

Jiashun Jin *Learning and ranking of research topics*

In his Fisher's Lecture in 1996, Efron suggested that there is a philosophical triangle in statistics with "Bayesian", "Fisherian", and "Frequentist" being the three vertices, and many representative statistical methods can be viewed as a convex linear combination of the three philosophies. We collected and cleaned a data set consisting of the citation and bibtex (e.g., title, abstract, author information) data of 83,331 papers published in 36 journals in statistics and related fields, spanning 41 years. Using the data set, we constructed 21 co-citation networks, each for a time window between 1990 and 2015. We propose a dynamic Degree-Corrected Mixed- Membership (dynamic-DCMM) model, where we model the research interests of an author by a low-dimensional weight vector (called the network memberships) that evolves slowly over time. We propose dynamic-SCORE as a new approach to estimating the memberships. We discover a triangle in the spectral domain which we call the Statistical Triangle, and use it to visualize the research trajectories of individual authors. We interpret the three vertices of the triangle as the three primary research areas in statistics: Bayes, Biostatistics and Nonparametrics. The Statistical Triangle further splits into 15 sub- regions, which we interpret as the 15 representative sub-areas in statistics. These results provide useful insights over the research trend and behavior of statisticians.

Abbas Khalili *Estimation and Sparsity in overfitted mixture-of-experts (MOE) models*

In this talk, we will first give an overview of the recent developments on estimation and sparsity in MOE models. We then discuss overfitted sparse MOE models and challenges when performing both estimation and feature selection.

Y Tracy Ke *Testing High-dimensional Multinomials with Applications to Text Analysis*

Motivated by applications in text mining and discrete distribution inference, we test for equality of probability mass functions of K groups of high-dimensional multinomial distributions. Special cases of this problem include global testing for topic models, two-sample testing in authorship attribution, and closeness testing for discrete distributions. A test statistic, which is shown to have an asymptotic standard normal distribution under the null hypothesis, is proposed. This parameter-free limiting null distribution holds true without requiring identical multinomial parameters within each group or equal group sizes. The optimal detection boundary for this testing problem is established, and the proposed test is shown to achieve this optimal detection boundary across the entire parameter space of interest. The proposed method is demonstrated in simulation studies and applied to analyze two real-world datasets to examine, respectively, variation among customer reviews of Amazon movies and the diversity of statistical paper abstracts. This is based on joint work with Tony Cai and Paxton Turner.

Pengfei Li *Maximum binomial likelihood for multivariate mixture data*

Multivariate mixture data analysis presents numerous challenges and constitutes a vital area of interest in the fields of statistics and data science. Research into multivariate mixture structures holds relevance across diverse application domains and plays a pivotal role in the advancement of artificial intelligence (AI) and machine learning. In this paper, we focus on nonparametric estimation techniques for multivariate mixture data. Specifically, we assume a known number of subpopulations and propose a binomial likelihood method,

along with an efficient numerical algorithm, to estimate the mixing proportions and cumulative distribution functions of these subpopulations without relying on parametric assumptions. Through extensive numerical experiments, we demonstrate three key advantages of our approach: (1) Our method eliminates the need for tuning parameters. (2) It does not require the assumption of continuous component density functions. (3) Our method consistently delivers stable performance. Under mild regularity conditions, we provide theoretical proofs for the L_2 convergence and uniform convergence of our estimators. To illustrate the practical performance of our method, we include a real-data example.

Xihong Lin *Ensemble Testing of Global Null Hypotheses with Applications to Whole Genome Sequencing Studies*

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact of no uniformly most powerful test, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power, e.g., the class of alternatives in the scenario of the same effect sign or signal sparsity. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. The key component of the framework is to introduce a certain random procedure in the construction of base tests. We then apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations are conducted to demonstrate type I error control and power gain of the proposed ensemble tests. In an analysis of the WGS data from the Atherosclerosis Risk in Communities (ARIC) study, the ensemble tests demonstrate substantial and consistent power improvement compared to other existing tests. This is a joint work with Yaowu Liu (Southwestern Finanicial University) and Zhonghua Liu (Columbia University)

Wei Lin *Nonasymptotic theory for two-layer neural networks: Beyond the bias-variance trade-off*

Large neural networks have proved remarkably effective in modern deep learning practice, even in the overparametrized regime where the number of active parameters is large relative to the sample size. This contradicts the classical perspective that a machine learning model must trade off bias and variance for optimal generalization. To resolve this conflict, we present a nonasymptotic generalization theory for two-layer neural networks with ReLU activation function by incorporating scaled variation regularization. Interestingly, the regularizer is equivalent to ridge regression from the angle of gradient-based optimization, but plays a similar role to the group lasso in controlling the model complexity. By exploiting this ridge-lasso duality, we obtain new prediction bounds for all network widths, which reproduce the double descent phenomenon. Moreover, the overparametrized minimum risk is lower than its underparametrized counterpart when the signal is strong, and is nearly minimax optimal over a suitable class of functions. By contrast, we show that overparametrized random feature models suffer from the curse of dimensionality and thus are suboptimal.

Yukun Liu *Classification and outlier detection with semi-parametric empirical likelihood under density ratio model*

The goal of classification is to assign categorical labels to unlabelled test data based on patterns and relationships learned from a labeled training dataset.Yet this task become challenging when the training data and the test data exhibit distributional mismatches. The unlabelled test data follow a finite mixture model, which is not identifiable without any model assumptions. In this paper, we propose to model the test data by a finite semiparametric mixture model under density ratio model, and construct a semiparametric empirical likelihood prediction set (SELPS) for the labels in the test data. Our approach tries to optimize the out-of-sample performance, aiming to include the correct class and to detect outliers as often as possible. It has the potential to enhance the robustness and effectiveness of classification models when dealing with varying distributions between training and test data. Our method circumvents a stringent separation assumption between training data and outliers, which is required by Guan and Tibshirani (2022) but is often violated by commonly-used distributions. We prove asymptotic consistency and normalities of our parameter estimators

and asymptotic optimality of the proposed SELPS. We illustrate our methods by analyzing four real-world datasets.

Po-Ling Loh *Differentially private penalized M-estimation via noisy optimization*

We present a noisy composite gradient descent algorithm for differentially private statistical estimation in high dimensions. We begin by providing general rates of convergence for the parameter error of successive iterates under assumptions of local restricted strong convexity and local restricted smoothness. Our analysis is local, in that it ensures a linear rate of convergence when the initial iterate lies within a constant-radius region of the true parameter. At each iterate, multivariate Gaussian noise is added to the gradient in order to guarantee that the output satisfies Gaussian differential privacy. We then derive consequences of our theory for linear regression and mean estimation. Motivated by M-estimators used in robust statistics, we study loss functions which downweight the contribution of individual data points in such a way that the sensitivity of function gradients is guaranteed to be bounded, even without the usual assumption that our data lie in a bounded domain. We then show how the private estimators obtained by noisy composite gradient descent may be used to obtain differentially private confidence intervals for regression coefficients, by leveraging work in Lasso debiasing proposed in high-dimensional statistics.

Jinchi Lv *SOFARI: High-Dimensional Manifold-Based Inference*

Multi-task learning is a widely used technique for harnessing information from various tasks. Recently, the sparse orthogonal factor regression (SOFAR) framework, based on the sparse singular value decomposition (SVD) within the coefficient matrix, was introduced for interpretable multi-task learning, enabling the discovery of meaningful latent feature-response association networks across different layers. However, conducting precise inference on the latent factor matrices has remained challenging due to orthogonality constraints inherited from the sparse SVD constraint. In this paper, we suggest a novel approach called high-dimensional manifold-based SOFAR inference (SOFARI), drawing on the Neyman near-orthogonality inference while incorporating the Stiefel manifold structure imposed by the SVD constraints. By leveraging the underlying Stiefel manifold structure, SOFARI provides bias-corrected estimators for both latent left factor vectors and singular values, for which we show to enjoy the asymptotic mean-zero normal distributions with estimable variances. We introduce two SOFARI variants to handle strongly and weakly orthogonal latent factors, where the latter covers a broader range of applications. We illustrate the effectiveness of SOFARI and justify our theoretical results through simulation examples and a real data application in economic forecasting. This is a joint work with Yingying Fan, Zemin Zheng and Xin Zhou.

Ping Ma *Subsampling in Large Graphs via Ricci Curvature*

In the past decades, many large graphs with millions of nodes have been collected/constructed. The high computational cost and significant visualization difficulty hinder the analysis of large graphs. Researchers have developed many graph subsampling approaches to provide a rough sketch that preserves global properties. By selecting representative nodes, these graph subsampling methods can help researchers estimate the graph statistics, e.g., the number of communities, of the large graph from the subsample. However, the available subsampling methods, e.g., degree node sampler and random walk sampler, tend to leave out minority communities because nodes with high degrees are more likely to be sampled. In this talk, I present a novel subsampling method via an analog of Ricci curvature in manifolds, i.e., Ollivier Ricci curvature.

Geoffrey McLachlan *An Apparent Paradox in Semi-Supervised Learning*

With the considerable interest on machine learning these days, there is increasing attention being given to a semi-supervised learning (SSL) approach to constructing a classifier. As is well known, the (Fisher) information in an unclassified feature with unknown class label is less (considerably less for weakly separated classes) than that of a classified feature which has known class label. Hence in the case where the absence of class labels does not depend on the data, the expected error rate of a classifier formed from the classified and unclassified features in a partially classified sample can be relatively much greater than that if the sample were completely classified. On treating the labels of the unclassified features as missing data and adopting a framework for their missingness, it is shown that the performance of the Bayes classifier can be improved to an extent where the SSL rule so produced can outperform the rule based on the sample if it were completely classified. This is a most surprising result. It can occur in situations where the unclassified features tend to fall in overlapping regions of the classes in the feature space; that is, for features that are difficult to classify.

Such features tend to have relatively high entropy and so it is proposed that the probability a class label is missing be modelled as a function of the entropy of the associated feature vector. This is joint work with Daniel Ahfock.

Axel Munk *Transport Dependency: Optimal transport based depedency measures*

Finding meaningful ways to determine the dependency between two random variables ξ and ζ is a timeless statistical endeavor with vast practical relevance. In recent years, several concepts that aim to extend classical means (such as the Pearson correlation or rank-based coefficients like Spearmans ρ) to more general spaces have been introduced and popularized, a well-known example being the distance correlation. In this talk, we propose and study an alternative framework for measuring statistical dependency, the transport dependency $\tau \geq 0$ (TD), which relies on the notion of optimal transport and is applicable in general Polish spaces. It can be estimated via the corresponding empirical measure, is versatile and adaptable to various scenarios by proper choices of the cost function. It intrinsically respects metric and geometric properties of the ground spaces. Notably, statistical independence is characterized by $\tau = 0$, while large values of τ indicate highly regular relations between ξ and ζ . Based on sharp upper bounds, we exploit three distinct dependency coefficients with values in $[0, 1]$, each of which emphasizes different functional relations: These transport correlations attain the value 1 if and only if $\zeta = \phi(\xi)$, where ϕ is a) a Lipschitz function, b) a measurable function, c) a multiple of an isometry, which all can be undestood as a latent, unknown dependency structure. Besides a conceptual discussion of transport dependency, we address numerical issues and its ability to adapt automatically to the potentially low intrinsic dimension of the ground space. Monte Carlo results suggest that TD is a robust quantity that efficiently discerns dependency structure from noise for data sets with complex internal metric geometry. The use of TD for inferential tasks is illustrated for independence testing on a data set of trees from cancer genetics.

This is joint work with Giacomo Nies and Thomas Staudt.

Yumou Qiu *Optimal signal detection in covariance and precision matrices*

This talk considers testing for high dimensional covariance and precision matrices by deriving the detection boundaries as a function of the signal sparsity and signal strength. It first shows that the optimal detection boundary for testing sparse means is the minimax detection lower boundary for testing covariance and precision matrices. Multi-level thresholding tests are proposed and are shown to be able to attain the detection lower boundaries over a substantial range of the sparsity parameter, implying that the multi-level thresholding tests are sharp optimal in the minimax sense over the range. The asymptotic distributions of the multi-level thresholding statistic for covariance and precision matrices are derived by developing a novel U-statistic decomposition to handle the complex dependence among the elements of the estimated covariance and precision matrices. The superiority in the detection boundary of the multi-level thresholding test over the existing tests are also demonstrated.

Richard Samworth *Isotonic subgroup selection*

Given a sample of covariate-response pairs, we consider the subgroup selection problem of identifying a subset of the covariate domain where the regression function exceeds a pre-determined threshold. We introduce a computationally-feasible approach for subgroup selection in the context of multivariate isotonic regression based on martingale tests and multiple testing procedures for logically-structured hypotheses. Our proposed procedure satisfies a non-asymptotic, uniform Type I error rate guarantee with power that attains the minimax optimal rate up to poly-logarithmic factors. Extensions cover classification, isotonic quantile regression and heterogeneous treatment effect settings. Numerical studies on both simulated and real data confirm the practical effectiveness of our proposal, which is implemented in the R package ISS.

Weijie Su *Gaussian Differential Privacy and How to Enhance Census Data Privacy for FREE!*

In the 2020 Decennial Census, differential privacy (DP) was used to protect the privacy of sensitive census data. While being a mathematically rigorous approach to privacy-preserving data analysis, DP often results in a poor tradeoff between privacy guarantees and statistical efficiency, particularly under composition. In this talk, we demonstrate how to achieve a better tradeoff using f-DP, through an application to decennial census data. Our results are based on a refined analysis of privacy losses under composition using Edgeworth expansions. Experimental results indicate that our new approach can achieve a roughly 10% decrease in the

MSE of census data, while maintaining the same privacy guarantees as the method employed in the 2020 Decennial Census.

Wenguang Sun *Ranking and Selection in Large-Scale Inference of Heteroscedastic Units*

The allocation of limited resources to a large number of potential candidates presents a pervasive challenge. In the context of ranking and selecting top candidates from heteroscedastic units, conventional methods often result in over-representations of subpopulations, and this issue is further exacerbated in large-scale settings where thousands of candidates are considered simultaneously. To address this challenge, we propose a new multiple comparison framework that incorporates a modified power notion to prioritize the selection of important effects and employs a novel ranking metric to assess the relative importance of units. We develop both oracle and data-driven algorithms, and demonstrate their effectiveness in controlling the error rates and achieving optimality. We evaluate the numerical performance of our proposed method using simulated and real data. The results show that our framework enables a more balanced selection of effects that are both statistically significant and practically important, and results in an objective and relevant ranking scheme that is well-suited to practical scenarios.

Yan Shuo Tan *The Computational Curse of Big Data for Bayesian Additive Regression Trees: A Hitting Time Analysis*

Bayesian Additive Regression Trees (BART) is a popular Bayesian non-parametric regression algorithm that is commonly used in causal inference. Its posterior is a distribution over sums of decision trees, and Markov Chain Monte Carlo (MCMC) is performed on this space to obtain approximate posterior samples. While the inferential properties of the BART posterior has been well-studied, there has been little theoretical work on the computational properties of BART. This is unfortunate because the BART sampler is notoriously often slow to mix. Prior work in this direction focused on mixing time lower bounds on tree structures, but these are unidentifiable parameters of the model. In this talk, we introduce a new method of quantifying the computational effectiveness of the BART sampler via hitting time lower bounds for the highest density posterior region, which conveniently captures all tree structures with the smallest bias and complexity. Across a range of different data generating models, we show theoretically that the hitting times grow with the training sample size, and we further illustrate this phenomenon with an extensive simulation study. Our results yield insights on why the BART sampler may experience computational issues and how to overcome these problems, such as via adjusting the temperature of the sampler.

Yuxin Tao *Homogeneity pursuit in ranking inferences based on pairwise comparison data*

The Bradley-Terry-Luce (BTL) model is one of the most celebrated models for ranking inferences based on pairwise comparison data, which associates individuals with latent preference scores and produces ranks. An important question that arises is the uncertainty quantification for ranks. It is natural to think that ranks for two individuals are not trustworthy if there is only a subtle difference in their preference scores. In this paper, we explore the homogeneity of scores in the BTL model, which assumes that individuals cluster into groups with the same preference scores. We introduce the clustering algorithm in regression via data-driven segmentation (CARDS) penalty into the likelihood function, which can rigorously and effectively separate parameters and uncover group structure. Statistical properties of two versions of CARDS are analyzed. As a result, we achieve a faster convergence rate and sharper confidence intervals for the maximum likelihood estimation (MLE) of preference scores, providing insight into the power of exploring low-dimensional structure in a high-dimensional setting. Real data, including NBA basketball ranking and netflix movie ranking, are analyzed, which demonstrate the improved prediction performance and interpretation ability of our method.

Xin Tong *Neyman-Pearson and equal opportunity: when efficiency meets fairness in classification*

Organizations often rely on statistical algorithms to make socially and economically impactful deci- sions. We must address the fairness issues in these important automated decisions. On the other hand, economic efficiency remains instrumental in organizations survival and success. Therefore, a proper dual focus on fairness and efficiency is essential in promoting fairness in real-world data science solutions. Among the first efforts towards this dual focus, we incorporate the equal op- portunity (EO) constraint into the Neyman-Pearson (NP) classification paradigm. Under this new NP-EO framework, we derive the oracle classifier,

propose finite-sample based classifiers that sat- isfy population-level fairness and efficiency constraints with high probability, and demonstrate the statistical and social effectiveness of our algorithms on simulated and real datasets.

Wanjie Wang *Network-Guided Covariate Selection and Downstream Applications*

Nowadays, it is frequently seen that the data set often contains network information and covariates. Studies have shown that the covariates will be helpful in uncovering the network structure. The opposite direction should also work, that the network information helps to denoise the covariates and improve the statistical inference. In this talk, I will present a covariate selection method based on the spectral information of the adjacency matrix. By the eigenvectors, we design a testing statistic of the covariates and select them by Higher-Criticism statistic. We prove the optimality of this method and the effect of it in the regression and clustering problems.

Marten Wegkamp *Discriminant Analysis in High-Dimensional Gaussian Mixtures*

We consider binary classification of high-dimensional features under a postulated model with a low-dimensional latent Gaussian mixture structure and non-vanishing noise. We propose a computationally efficient classifier that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. We derive explicit rates of convergence of the excess risk of the proposed PC-based classifier and we prove that the obtained rates are optimal, up to some logarithmic factor, in the minimax sense. In the second part of the talk, we retain all PCs to estimate the direction of the optimal separating hyperplane. The estimated hyperplane is shown to interpolate on the training data. While the direction vector can be consistently estimated as could be expected from recent results in linear regression, a naive plug-in estimate fails to consistently estimate the intercept. A simple correction, that requires an independent hold-out sample, renders the procedure consistent and even minimax optimal in many scenarios. The interpolation property of the latter procedure can be retained, but surprisingly depends on the way we encode the labels.

This is joint work with Xin Bing, University of Toronto

Yun Wei *Parameter Estimations in Finite Mixture Models*

Finite mixture models are among the most widely used models to address data heterogeneity. Recent work [Henrich & Kahn 2018] establishes the optimal uniform convergence rate (in the minimax sense) using minimum Kolmogorov-Smirnov distance estimators. Subsequently, the work [Wu & Yang 2020] proves that the method of moments also achieves the optimal uniform convergence rate for Gaussian mixture models. We propose a general framework and estimator that includes the two previous methods as special cases and establish the convergence rate under the general framework. The general framework can generate novel methods, and one instance is based on the maximum mean discrepancy, which is also shown to achieve optimal uniform convergence rate. Pointwise convergence rates are also established under the general framework.

In the second part of the talk, we consider the finite mixture of product distribution with the special structure that the product distributions in each mixture component are also identically distributed. In this setup, each mixture component consists of samples from repeated measurements, making such data exchangeable sequences. Applications of the model include psychological studies and topic modeling. We show that with sufficient repeated measurements, a model that is not originally identifiable becomes identifiable. The posterior contraction rate for parameter estimation is also obtained, showing that repeated measurements are beneficial for estimating parameters in each mixture component. These results hold for general probability kernels, including all regular exponential families, and can be applied to hierarchical models.

Based on joint work with Xuanlong Nguyen and Sayan Mukherjee.

Weixin Yao *Semiparametric Mixture of Regression with Unspecified Error Distributions*

In the fitting of mixture of linear regression models, the normal assumption has been traditionally used for the error term and then the regression parameters are estimated by the maximum likelihood estimate (MLE). Unlike the least squares estimate (LSE) for linear regression model, the validity of the MLE for mixtures of regression depends on the normal assumption. In order to relax the strong parametric assumption about the error density, in this article, we propose a mixture of linear regression model with unknown error density. We prove the identifiability of our proposed model and provide the asymptotic properties of the proposed estimates. In addition, we will propose an EM-type algorithm which uses a kernel density estimator for the

unknown error when calculating the classification probabilities in the E step. Using a Monte Carlo simulation study, we demonstrate that our method works comparably to the traditional MLE when the error is normal. In addition, we demonstrate the success of our new estimation procedure when the error is not normal. An empirical analysis of tone perception data is illustrated for the proposed methodology.

Zhigang Yao *Manifold Fitting with CycleGAN*

Manifold fitting, which offers substantial potential for efficient and accurate modeling, poses a critical challenge in non-linear data analysis. This study presents a novel approach that employs neural networks to fit the latent manifold. Leveraging the generative adversarial framework, this method learns smooth mappings between low-dimensional latent space and high-dimensional ambient space, echoing the Riemannian exponential and logarithmic maps. The well-trained neural networks provide estimations for the latent manifold, facilitate data projection onto the manifold, and even generate data points that reside directly within the manifold. Through an extensive series of simulation studies and real data experiments, we demonstrate the effectiveness and accuracy of our approach in capturing the inherent structure of the underlying manifold within the ambient space data. Notably, our method exceeds the computational efficiency limitations of previous approaches and offers control over the dimensionality and smoothness of the resulting manifold. This advancement holds significant potential in the fields of statistics and computer science. The seamless integration of powerful neural network architectures with generative adversarial techniques unlocks new possibilities for manifold fitting, thereby enhancing data analysis. The implications of our findings span diverse applications, from dimensionality reduction and data visualization to generating authentic data. Collectively, our research paves the way for future advancements in non-linear data analysis and offers a beacon for subsequent scholarly pursuits.

Puying Zhao *Augmented two-step estimating equations with nuisance functionals and complex survey data*

Statistical inference in the presence of nuisance functionals with complex survey data is an important topic in social and economic studies. The Gini index, Lorenz curves and quantile shares are among the commonly encountered examples. The nuisance functionals are usually handled by a plug-in nonparametric estimator and the main inferential procedure can be carried out through a two-step generalized empirical likelihood method. Unfortunately, the resulting inference is not efficient and the nonparametric version of the Wilks' theorem breaks down even under simple random sampling. We propose an augmented estimating equations method with nuisance functionals and complex surveys. The second-step augmented estimating functions obey the Neyman orthogonality condition and automatically handle the impact of the first-step plug-in estimator, and the resulting estimator of the main parameters of interest is invariant to the first step method. More importantly, the generalized empirical likelihood based Wilks' theorem holds for the main parameters of interest under the design-based framework for commonly used survey designs, and the maximum generalized empirical likelihood estimators achieve the semiparametric efficiency bound. Performances of the proposed methods are demonstrated through simulation studies and an application using the dataset from the New York City Social Indicators Survey. This is joint work with Changbao Wu

Qiong Zhang *Distributed learning of finite mixture models*

Advances in information technology have led to extremely large datasets that are often kept in different storage centers. Existing statistical methods must be adapted to overcome the resulting computational obstacles while retaining statistical validity and efficiency. In this situation, the split-and-conquer strategy is among the most effective solutions to many statistical problems, including quantile processes, regression analysis, principal eigenspaces, and exponential families. This paper applies this strategy to develop a distributed learning procedure of finite Gaussian mixtures. We recommend a reduction strategy and invent an effective majorization-minimization algorithm. The new estimator is consistent and retains root-n consistency under some general conditions. Experiments based on simulated and real-world datasets show that the proposed estimator has comparable statistical performance with the global estimator based on the full dataset, if the latter is feasible. It can even outperform the global estimator for the purpose of clustering if the model assumption does not fully match the real-world data. It also has better statistical and computational performance than some existing split-and-conquer approaches.

Ji Zhu *A Latent Space Model for Hypergraphs with Diversity and Heterogeneous Popularity*

While relations among individuals make an important part of data with scientific and business interests, existing statistical modeling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. This work addresses the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, we propose a new latent space model for hypergraphs using determinantal point processes, which is driven by the diversity within hyperedges and each node's popularity. This model mechanism is in contrast to existing hypergraph models, which are predominantly driven by similarity rather than diversity. Additionally, the proposed model accommodates broad types of hypergraphs, with no restriction on the cardinality and multiplicity of hyperedges. Consistency and asymptotic normality of the maximum likelihood estimates of the model parameters have been established. Simulation studies and an application to the What's Cooking data show the effectiveness of the proposed model.

Liping Zhu *Testing high-dimensional covariate effects in the presence of covariate heterogeneity*

In this talk, I introduce several tests for the mean effects of high-dimensional covariates on the response. In many applications, different components of covariates usually exhibit various levels of variation, which is ubiquitous in high-dimensional data. To simultaneously accommodate such heteroscedasticity and high dimensionality, we propose a novel test based on an aggregation of the marginal cumulative covariances, requiring no prior information on the specific form of regression models. Our proposed test statistic is scale-invariance, tuning-free and convenient to implement. The asymptotic normality of the proposed statistic is established under the null hypothesis. We further study the asymptotic relative efficiency of our proposed test with respect to the state-of-art universal tests in two different settings: one is designed for high-dimensional linear model and the other is introduced in a completely model-free setting. A remarkable finding reveals that, thanks to the scale-invariance property, even under the high-dimensional linear models, our proposed test is asymptotically much more powerful than existing competitors for the covariates with heterogeneous variances while maintaining high efficiency for the homoscedastic ones.

Lixing Zhu *Change point detection for tensors with heterogeneous slices*

In many applications, tensor data may consist of heterogeneous slices according to a categorical mode, and independent but not identically distributed error tensors over time. To detect change structures in such tensor data, we define a mode-based signal-screening Frobenius distance for the moving sums of slices to handle both dense and sparse model structures of the tensors. Based on this distance, we construct a mode-based signal statistic using a sequence of ratios with mode-based adaptive-to-change ridge functions. The number of changes and their locations can be consistently estimated in certain senses, and the confidence intervals of the locations of change points are constructed when the standardized error tensors are homogenous. The results hold when the size of the tensor and the number of change points diverge at certain rates, respectively. Numerical studies are conducted to examine the finite sample performances of the proposed method. We also analyze two real data examples for illustration.

Xuening Zhu *Network autoregression for incomplete matrix-valued time series*

We study the dynamics of matrix-valued time series with observed network structures by proposing a matrix network autoregression model with row and column networks of the subjects. We incorporate covariate information and a low rank intercept matrix. We allow incomplete observations in the matrices and the missing mechanism can be covariate dependent. To estimate the model, a two-step estimation procedure is proposed. The first step aims to estimate the network autoregression coefficients, and the second step aims to estimate the regression parameters, which are matrices themselves. Theoretically, we first separately establish the asymptotic properties of the autoregression coefficients and the error bounds of the regression parameters. Subsequently, a bias reduction procedure is proposed to reduce the asymptotic bias and the theoretical property of the debiased estimator is studied. Lastly, we illustrate the usefulness of the proposed method through a number of numerical studies and an analysis of a Yelp data set.