# Abstracts of lnvited Talks

(Order of Talks)

## 1. Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression

Jianqing Fan

**Abstract:** This talk introduces a Factor Augmented Sparse Throughput (FAST) model that utilizes latent factors and sparse idiosyncratic components for nonparametric regression and addresses a stylized feature in high-dimensional data. The FAST model bridges factor models (dense in inputs) on one end and sparse nonparametric models on the other end. It encompasses structured nonparametric models such as factor augmented additive model and sparse low-dimensional nonparametric interaction models as specific examples. Via diversified projections as estimation of latent factor space, we employ truncated Deep ReLU networks to nonparametric factor regression without explicit regularization and to more general FAST model using nonconvex regularization, resulting in factor augmented regression using neural network (FAR-NN) and FAST-NN estimators respectively. We show that FAR-NN and FAST-NN estimators adapt to unknown low-dimensional structure using hierarchical composition models in the sense of nonasymptotic minimax rates. We also study statistical learning for the factor augmented additive model using a more specific neural network architecture that facilitates the implementations. Our results are applicable to the weak dependent cases where the covariates do not admit factor structure. On the way of proving our main technical result for FAST-NN, we establish new deep ReLU network approximation result that contributes to the foundation of neural network theory. (Joint work with Yihong Gu)

## 2. SURE-tuned Lasso

Cun-Hui Zhang

**Abstract:** In sparse linear regression, the Lasso estimator requires a proper penalty to achieve the optimal rate in prediction error. Theory suggests that a proper penalty is proportional to the noise level of the regression model, which is usually unknown and often treated as a nuisance parameter in theoretical studies. The scaled Lasso eliminates the dependence of the unknown noise level in its scale-free penalty via an alternating minimization scheme. It essentially reduces the tuning parameter to a constant factor within a narrow band. Stein's unbiased risk estimation or SURE is a common criterion to select an estimator with minimal prediction error among a collection of candidates. We propose a SURE-tuned scaled Lasso method to fine-tune the constant factor in the scale-free penalty by SURE criterion. We prove an oracle inequality for the proposed estimator, which provides a theoretical guarantee that up to a higher-order term, our method achieves the minimal prediction error within an interval of penalty levels. Simulation studies under broad settings demonstrate its good performance in supporting our theory.

## 3. Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling

Annie Qu

**Abstract:** Crowdsourcing has emerged as an alternative solution for collecting large scale labels. However, the majority of recruited workers are not domain experts, so their contributed labels could be noisy. In this paper, we propose a two-stage model to predict the true labels for multicategory classification tasks in crowdsourcing. In the first stage, we fit the observed labels with a latent factor model and incorporate subgroup structures for both

tasks and workers through a multi-centroid grouping penalty. Group-specific rotations are introduced to align workers with different task categories to solve multicategory crowdsourcing tasks. In the second stage, we propose an anglebased approach to identify high-quality worker subgroups who are relied upon to assign labels to tasks. In theory, we show the estimation consistency of the latent factors and the prediction consistency of the proposed method. The simulation studies show that the proposed method outperforms the existing competitive methods, assuming the subgroup structures within tasks and workers. We also demonstrate the application of the proposed method to real world problems and show its superiority.

## 4. Repro samples method: A general framework for performance guaranteed finite- and large-sample frequentist inferences in data science

Mingge Xie

**Abstract:** Statistical inference on machine learning methods has been hampered by the slow progresses to address uncertainty issues on irregular inference problems to which the large sample central limit theorem does not apply. Particularly, methodology developments to make inference for discrete or non-numerical parameters, problems involving non-numerical data, and other irregular inference are lacking behind. In this talk, we present a novel, wide-reaching, and effective simulation-inspired approach, called `repro samples method,' to conduct statistical inference for these irregular problems plus more. We systemically develop both exact and approximate (asymptotic) theories to support the development. An attractive feature of the development is that it doesn't need to rely on a likelihood or use the large sample central limit theorem, and thus is especially effective for complicated and irregular inference problems often encountered in machine learning and data science. The effectiveness of the proposed approach is illustrated through a number of examples, including two case study examples of ``highly non-trivial" on: (a) normal mixture model where we construct a finite sample confidence set for the unknown number of components while controlling varying length of nuisance parameters; and (b) high dimensional regression where we construct both finite and large sample confidence sets for both unknown model and model coefficients while fully accounting for model selection uncertainty. Comparisons and numerical studies demonstrate that the proposed methods have far superior performance to existing Bayesian, frequentist, and fiducial approaches. Although the case studies pertain to the traditional statistics models, the framework also has direct extensions to more complex machine learning models. (Joint work with Peng Wang and Linjun Zhang).

## 5. Gene-environment interaction analysis assisted by multi-level hierarchical prior information

Shuangge Ma

**Abstract:** High-dimensional data analysis often suffers from a lack of information. This can be especially true for gene-environment interaction analysis, which has higher data dimensionality and weaker signals. In a seminal 2016 JASA article, Yuan Jiang, Yunxiao He, and Heping Zhang developed a prior Lasso approach, which can accommodate prior information on variables' importance and adaptively balance between prior information and observed data. In this work, we further develop this approach by incorporating multi-level prior information, which is generated by a sequence of query terms that have a hierarchical structure. The proposed approach can not only more effectively accommodate prior information but also suggest the "usefulness" of different information levels.

## 6. Clusters disagreements between training and testing data

Ofer Harel

**Abstract:** We introduce a disagreement problem that may be an issue when classification is implemented on estimated classes of the population.

A disagreement problem denotes the case in which a sample fails to cover a specific class that a testing observation belongs to. Or in other words, the training and testing data do not agree on the number of classes in the population. These disagreement problems may occur due to various reasons such as sampling errors, selection bias, or emerging classes of the population. Once the disagreement problem occurs, a testing observation will be misclassified, because a classification rule based on the sample cannot capture a class not observed in the training data (sample). To overcome such issues, we suggest a two-stage classification method that can ameliorate a disagreement problem in classification. Our proposed method tests whether a testing observation is sampled from the observed or possibly unobserved class, then classifies it based on the test result. We suggest a test for identification of the disagreement problem and demonstrate the performance of the two-stage classification via numerical studies.

## 7. A statistical learning method for simultaneous copy number estimation and subclone clustering with single-cell-sequencing data

Feifei Xiao

**Abstract:** The availability of single cell sequencing (SCS) enables us to assess intra-tumor heterogeneity and identify cellular subclones without the confounding effect from mixed cells. Copy number aberrations (CNAs) have been commonly used to identify subclones in SCS data since cells comprising a subpopulation are found to share genetic profile. However, currently available methods may generate spurious results (e.g., falsely identified CNAs) in the procedure of CNA detection, hence diminishing the accuracy of subclones identification from a large complex cell population. In this study, we developed a CNA detection method based on a fused lasso model, referred to as FLCNA, which can simultaneously identify subclones in single cell DNA sequencing (scDNA-seq) data. Spike-in simulations and real data analyses were conducted to evaluate the clustering purity of FLCNA benchmarking to existing copy number estimation methods (SCOPE, HMMcopy) in combination with commonly used clustering methods. In conclusion, FLCNA provided superior performance in subclone identification with scDNA-seq data.

## 8. Equivariant Variance Estimation for Multiple Change-point Model

Yue Niu

**Abstract:** The variance of noise plays an important role in many change-point detection procedures and the associated inferences. Most commonly used variance estimators require strong assumptions on the true mean structure or normality of the error distribution, which may not hold in applications. More importantly, the qualities of these estimators have not been discussed systematically in the literature. In this talk, we introduce a framework of equivariant variance estimation for multiple change-point models. In particular, we characterize the set of all equivariant unbiased quadratic variance estimators for a family of change-point model classes, and develop a minimax theory for such estimators.

## 9. Understanding high-dimensional sparsity-free prediction using approximate message passing with genetic applications

Bingxin Zhao

**Abstract:** Numerous statistical models have been proposed for genetic prediction using high-dimensional data from genome-wide association studies. The relative performance of these methods varies with the dataset characteristics and underlying genetic architecture of the targeted traits/diseases, and typically no method is dominant across all applications. Motivated by these empirical observations, we present a unified analysis of popular genetic prediction methods in a high-dimensional sparsity-free setting, where we are allowed to have few to many true signals. We show that the relative performance of the L1-type and L2-type regularization estimators depends on the model sparsity, the signal strength, the feature covariance structure, and the ratio of dimension over sample size. The analysis is based on recent advances in approximate message passing (AMP).

## 10. Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

Dehan Kong

**Abstract:** Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method, which constructs pseudo variables to remove irrelevant candidate instruments having spurious correlations with the exposure. Synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

## 11. Identification, Amplification and Measurement: A bridge to Gaussian Differential Privacy

Linglong Kong

**Abstract:** Gaussian differential privacy (GDP) is a single-parameter family of privacy notions that provides coherent guarantees to avoid the exposure of sensitive individual information. Despite the extra interpretability and tighter bounds under composition GDP provides, many widely used mechanisms (e.g., the Laplace mechanism) inherently provide GDP guarantees but often fail to take advantage of this new framework because their privacy guarantees were derived under a different background. In this paper, we study the asymptotic properties of privacy profiles and develop a simple criterion to identify algorithms with GDP properties. We propose an efficient method for GDP algorithms to narrow down possible values of an optimal privacy measurement, μ with an arbitrarily small and quantifiable margin of error. For non GDP algorithms, we provide a post-processing procedure that can amplify existing privacy guarantees to meet the GDP condition. As applications, we compare two single-parameter families of privacy notions, ε-DP, and μ-GDP, and show that all ε-DP algorithms are intrinsically also GDP. Lastly, we show that the combination of our measurement process and the composition theorem of GDP is a powerful and convenient tool to handle compositions compared to the traditional standard and advanced composition theorems.

## 12. Unsupervised learning in data integration studies using JIVE with Gaussian mixtures

Benjamin Risk

**Abstract:** A common goal in data integration studies is to identify subgroups. JIVE (joint and individual variation explained) has been proposed as a method to extract shared (joint) and unique (individual) information from each dataset, and cluster analysis is applied after extraction of joint and individual scores. We present a probabilistic JIVE model with mixture of Gaussians (JIVE-mix) that enables joint probabilistic clustering of subjects with multiple data sources. Our simulations demonstrate improvement over existing approaches. We apply our method to MRI brain imaging and CSF biomarker measurements in the Alzheimer's Disease Neuroimaging Initiative, which reveals interesting clusters that suggest distinct pathologies.

## 13. S-GMAS: Shape based Genome-wide Mediation Analysis

Chao Huang

**Abstract:** Causal mediation analysis is widely utilized in neuroscience to investigate the role of brain image phenotypes in the neurological pathways from genetic exposures to clinical outcomes. However, it is still difficult to conduct a genome-wide mediation analysis with the shapes of brain regions as mediators due to several challenges including (i) large-scale genetic exposures, i.e., millions of single-nucleotide polymorphisms (SNPs); (ii) nonlinear Hilbert space for shape mediators; and (iii) statistical inference on the direct and indirect effects. To tackle these challenges, this paper proposes a mediation analysis framework with high dimensional genetic exposures and shape mediators. First, the square-root velocity function representations are extracted from the shapes, which fall in an unconstrained linear Hilbert subspace. Second, to address the issue caused by the high dimensionality in genetic exposures, the global sure independence screening procedure is conducted to discover candidate SNPs influencing the shape mediators. To identify the underlying causal pathways from the detected SNPs to the clinical outcome implicitly through the shape mediators, we proposed a framework consisting of a function-on-scalar model and a scalar-on-function model. Furthermore, the bootstrap resampling approach is adopted to investigate both global and local significant mediation effects. Finally, our framework is applied to the ADNI multiple subregion shape data and we successfully identify the mediation effect of a subset of candidate SNPs on Alzheimer's Disease through different brain subregions.

## 14. Surface-based Brain Connectivity Analysis

Zhengwu Zhang

**Abstract:** Brain structural networks are often represented as discrete adjacency matrices, where each element in the matrix provides a summary of the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined a-priori using a brain atlas; a parcellation of the cortical surface constructed from anatomical considerations. Unfortunately, the choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This talk introduces an atlas-independent framework that overcomes these issues by modeling brain connectivity using smooth functions. In particular, our framework assumes that the pattern of observed white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain, referred to as the continuous connectivity. As a result, our framework is inherently both atlas and resolution independent, and so prevents information loss caused by large ROIs. Under this framework, we studied several important problems in brain connectivity analysis, e.g., how to define a network node, how to align brain connectivity across subjects and how to statistical analyze the the continuous connectivity.

## 15. Identification and Estimation of Treatment Effects in the Limited Overlap Region

Xiaohong Chen

**Abstract:** Strong ignorability is a commonly used assumption to identify average treament effects based on observational data. It is often argued that the conditional independence assumption can be made more plausible by using more covariates. However using more covariates makes the overlapping assumption less likely to hold. In most empirical applications, the supports of distributions of the covariate vector for different groups do not fully overlap or have limited overlap. Without imposing additional assumptions on the limited or no overlap region, average treatment effects for either the limited overlap region or for the whole population are not point identified.In this paper, we make a natural domain shift assumption for the limited overlap region based on optimal transport theory. We study identification of average treatment effects for the limited overlap region and propose three-step estimators of the average treatment effect and quantile treatment effect for the treated in the limited overlap region. We establish consistency and asymptotic normality of the proposed estimators under high level assumptions on the estimator of the optimal transport map. Three examples of the estimator of the optimal transport map are studied in detail and are shown to satisfy the high level assumptions under primitive conditions. We investigate the finite sample performance of our estimator and Wald inference via simulation.

## 16. On Robustness of Individualized Decision Rules

Yufeng Liu

**Abstract:** With the emergence of precision medicine, estimation of optimal individualized decision rules (IDRs) has attracted tremendous attentions in many scientific areas. Most existing literature has focused on finding optimal IDRs that can maximize the expected outcome for each individual. Motivated by complex individualized decision making procedures and the popular conditional value at risk, in this talk, I will introduce two new robust criteria to evaluate IDRs: one is focused on the average lower tail of the subjects' outcomes and the other is on the individualized lower tail of each subject's outcome. The proposed criteria take tail behaviors of the outcome into consideration, and thus the resulting optimal IDRs are robust in controlling adverse events. The optimal IDRs under our criteria can be interpreted as the distributionally robust decision rules that maximize the "worst-case" scenario of the outcome within a probability constrained set. Simulation studies and a real data application are used to demonstrate the robust performance of our methods.

## 17. Testing the effects of high-dimensional covariates via aggregating cumulative covariances

Runze Li

**Abstract:** In this paper, we test for the 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 new 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.

## 18. Reweighted Anderson-Darling Tests of Goodness-of-Fit

Chuanhai Liu

**Abstract:** Assessing goodness-of-fit is a fundamental statistical problem. This paper considers the class of reweighted Anderson-Darling tests, sheding new light on a geometric understanding of the problem via establishing an explicit one-to-one correspondence between the weights and their focal directions of distributional deviations under alternative hypothesis. It is shown that the weights that produce the test statistic with minimal variance put equal focuses on all the standardized deviations and, thereby, can serve as a general-purpose test. This practically optimal weights-based test is found to be practically equivalent to the Zhang test, which has been commonly perceived powerful. New large-sample results are established for this test. It is shown that these results can be useful for both understanding the large-sample behvior of the test statistic and large sample-based approximations.

## 19. Stability Approach to Regularization Selection for Reduced-Rank Regression

Yuan Jiang

**Abstract:** The reduced-rank regression model is a popular model to deal with multivariate response and multiple predictors, and is widely used in biology, chemometrics, econometrics, engineering, and other fields. In the reduced-rank regression modelling, a central objective is to estimate the rank of the coefficient matrix that represents the number of effective latent factors in predicting the multivariate response. Although theoretical results such as rank estimation consistency have been established for various methods, in practice rank determination still relies on information criterion based methods such as AIC and BIC or subsampling based methods such as cross validation. Unfortunately, the theoretical properties of these practical methods are largely unknown. In this paper, we present a novel method called StARS-RRR that selects the tuning parameter and then estimates the rank of the coefficient matrix for reduced-rank regression based on the stability approach. We prove that StARS-RRR achieves rank estimation consistency, i.e., the rank estimated with the tuning parameter selected by StARS-RRR is consistent to the true rank. Through a simulation study, we show that StARS-RRR outperforms other tuning parameter selection methods including AIC, BIC, and cross validation as it provides the most accurate estimated rank. In addition, when applied to a breast cancer dataset, StARS-RRR discovers a reasonable number of genetic pathways that affect the DNA copy number variations and results in a smaller prediction error than the other methods with a random-splitting process.

## 20. Transfer Learning: Optimality and Adaptive Algorithms

Tony Cai

**Abstract:** Human learners have the natural ability to use knowledge gained in one setting for learning in a different but related setting. This ability to transfer knowledge from one task to another is essential for effective learning. However, in statistical learning, most procedures are designed to solve one single task, or to learn one single distribution, based on observations from the same setting. In this talk, we discuss statistical transfer learning in various settings under the posterior drift model, which is a general framework and arises in many practical problems. The results show that significant benefit of incorporating data from the source distributions for learning under the target distribution.

## 21. Covariate-adjusted Expected Shortfall

Xuming He

**Abstract:** Expected shortfall, measuring the average outcome (e.g., portfolio loss) above a given quantile of its probability distribution, is a common financial risk measure. The same measure can be used to characterize treatment effects in the tail of an outcome distribution, with applications ranging from policy evaluation in economics and public health to biomedical investigations. Expected shortfall regression is a natural approach of modeling covariate-adjusted expected shortfalls. Because the expected shortfall cannot be written as a solution of a convex loss function at the population level, computational as well as statistical challenges around expected shortfall regression have led to stimulating research. We discuss some recent developments in this area, with a focus on a new optimization-based semiparametric approach to estimation of conditional expected shortfall that adapts well to data heterogeneity with minimal model assumptions.

## 22. Kidney Paired Donation Programs

Peter Song

**Abstract:** This talk aims to introduce Kidney Paired Donation Programs. Strategies of matching patient-donor pairs through volunteering exchanges of kidney organs from living donors are applied routinely as part of clinical decisions for life saving. Kidney paired donation programs (KPD) in the USA has contributed to approximately 25 percent (~6,600/25,000) of kidney transplants in 2021, the second largest donation next to the deceased donation. KPD may be formulated as an irregular market in Economics or attentively, in Data Science, a dynamic patient network with well-defined clinical utilities. Optimization of matji strategies, reinforced learning decision-making over time, and fairness of organ allocation policies give rise to a number of challenging but important problems that call for solutions.

## 23. Semi-parametric Learning for feature selection

Jiayang Sun

**Abstract:** The prolific accumulation of data from multiple domains provides a beautiful landscape of many interacting factors to target outcomes. These data challenge existing model and feature selection procedures used in statistics and data science. For example, features selected often depend on the model assumption that may be unrealistic. Determining variable transformations to make the model more realistic in a multivariate fashion is not trivial. This talk presents our preliminary work on a semi-parametric learning pipeline to study feature, transformation, and model selection in a "triathlon." We discuss the challenges and some guarantees and open up dialogues for paradigm changes.

## 24. iProMix: A mixture model for studying the function of ACE2 based on bulk proteogenomic data

Pei Wang

**Abstract:** Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused over six million deaths in the ongoing COVID-19 pandemic. SARS-CoV-2 uses ACE2 protein to enter human cells, raising a pressing need to characterize proteins/pathways interacted with ACE2. Large-scale proteomic profiling technology is not mature at single-cell resolution to examine the protein activities in disease-relevant cell types. We propose iProMix, a novel statistical framework to identify epithelial-cell specific associations between ACE2 and other proteins/pathways with bulk proteomic data. iProMix decomposes the data and models cell-type-specific conditional joint distribution of proteins through a mixture model. It improves cell-type composition estimation from prior input, and utilizes a non-parametric inference framework to account for uncertainty of cell-type proportion estimates in hypothesis test. Simulations demonstrate iProMix has well-controlled false discovery rates and favorable powers in non-asymptotic settings. We apply iProMix to the proteomic data of 110 (tumoradjacent) normal lung tissue samples from the Clinical Proteomic Tumor Analysis Consortium lung adenocarcinoma study, and identify interferon response pathways as the most significant pathways associated with ACE2 protein abundances in epithelial cells. Strikingly, the association direction is sex-specific. This result casts light on the sex difference of COVID-19 incidences and outcomes, and motivates sex-specific evaluation for interferon therapies.

## 25. Copula-based Multiple Indicator Kriging for non-Gaussian Random Fields

Huixia Wang

**Abstract:** In spatial statistics, the kriging predictor is the best linear predictor at unsampled locations, but not the optimal predictor for non-Gaussian processes. In this talk, I will introduce a copula-based multiple indicator kriging model for the analysis of non-Gaussian spatial data by thresholding the spatial observations at a given set of quantile values. The proposed copula model allows for flexible marginal distributions while modeling the spatial dependence via copulas. We show that the covariances required by kriging have a direct link to the chosen copula function. We then develop a semiparametric estimation procedure. The developed method provides the entire predictive distribution function at a new location, and thus allows for both point and interval predictions. The method demonstrates better predictive performance than the commonly used variogram approach and Gaussian kriging in the simulation studies. This is a joint work with Gaurav Agarwal and Ying Sun.

## 26. Efficient Multimodal Sampling via Tempered Distribution Flow

Xiao Wang

**Abstract:** Sampling from high-dimensional distributions is a fundamental problem in statistical research and practice, and has become a central task in Bayesian computing, Monte Carlo simulation, and energy-based models. However, great challenges emerge when the target density function is unnormalized and contains multiple modes that are isolated with each other. We tackle this difficulty by fitting an invertible transformation mapping applied to the target distribution, such that

the original distribution is warped into a new one that is much easier to sample from. The transformation mapping is constructed based on the normalizing flow model in deep learning. To address the multi-modality issue, our method adaptively learns a sequence of tempered distributions, which we term as a tempered distribution flow, to progressively approach the original distribution. Various experiments demonstrate the superior performance of this novel sampler compared to traditional methods. This is a joint work with Yixuan Qiu.

## 27. Conditional Survival Function Estimation Using Neural Networks for Censored Data with Time-Varying Covariates

Bin Nan

**Abstract:** We consider estimating the conditional distribution function using neural networks for censored survival data. The algorithm is built upon the data structure particularly constructed for the Cox regression with time-dependent covariates. Without imposing any model assumption, we consider a loss function that is based on the full likelihood where the conditional hazard function is the only unknown nonparametric parameter, for which unconstrained optimization methods can be applied. Through simulation studies, we show the proposed method possesses desirable performance, whereas the partial likelihood method yields biased estimates when model assumptions are violated. This is a joint work with Bingqing Hu.

## 28. Quantile Regression for Nonignorable Missing Data with Its Application of Analyzing Electronic Medical Records

Ying Wei

**Abstract:** Over the past decade, there has been growing enthusiasm for using electronic medical records (EMRs) for biomedical research. Quantile regression estimates distributional associations, providing unique insights into the intricacies and heterogeneity of the EMR data. However, the widespread nonignorable missing observations in EMR often obscure the true associations and challenges its potential for robust biomedical discoveries. We propose a novel method to estimate the covariate effects in the presence of nonignorable missing responses under quantile regression. This method imposes no parametric specifications on response distributions, which subtly uses implicit distributions induced by the corresponding quantile regression models. We show that the proposed estimator is consistent and asymptotically normal. We also provide an efficient algorithm to obtain the proposed estimate and a randomly weighted bootstrap approach for statistical inferences. Numerical studies, including an empirical analysis of real-world EMR data, are used to assess the proposed method's finite-sample performance compared to existing literature.

## 29. Distribution-invariant differential privacy

Xuan Bi

**Abstract:** Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in social science, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of the original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. We mitigate this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy in a wide range of simulations and real-world benchmarks.

## 30. Interpretability of Deep Neural Networks for Structured Input Data

Rui Feng

**Abstract:** Neural networks have demonstrated strong capabilities of discovering potential complex patterns for predicting outcomes. However, it is challenging to interpret the contribution of each input variable due to multi-layer over-parameterized black-box model. Layer-wise relevance propagation offers a computationally efficient solution to explain neural networks but is limited to independent input variables, which do not fit to many biomedical studies with correlated variables. In this paper, we developed a new inheritance relevance (IR) measure, to account for each variable's independent contribution to the overall prediction, at the presence of known correlated structure. IR are based on a previously proposed Peel Learning (PL) model, back-propagate nodes' relevance scores through structured layers, and allow for the gradients of various activation transformations. We demonstrated the utility of IR through simulations and an application to a TCGA breast cancer study.

## 31. Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses

Jiwei Zhao

**Abstract:** In this paper we consider the estimation of a high-dimensional parameter in an individualized linear threshold with binary responses. While the problem can be formulated into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to discontinuity of the sign function. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter, we propose an empirical risk minimization approach based on a regularized smoothed non-convex loss function. The Fisher consistency of the proposed method is guaranteed as the bandwidth of the smoothed loss is shrunk to zero. Statistically, we show that the convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. Computationally, an efficient path-following algorithm is proposed to compute the solution path. We show that this algorithm achieves geometric rate of convergence for computing the whole path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis from the ChAMP (Chondral Lesions And Meniscus Procedures) Trial. This is based on a joint work with Huijie Feng and Yang Ning.

## 32. ANNORE: genetic fine-mapping with functional annotation

Ching-Ti Liu

Genome-wide association studies (GWASs) have successfully identified loci of the human genome implicated in numerous complex traits. However, the limitations of this study design make it difficult to identify specific causal variants or biological mechanisms of association. We propose a method, AnnoRE, which uses GWAS summary statistics, local correlation structure among genotypes and functional annotation from external databases to prioritize the most plausible causal single-nucleotide polymorphisms (SNPs) in each trait-associated locus. Our proposed method improves upon previous fine-mapping approaches by estimating the effects of functional annotation from genome-wide summary statistics, allowing for the inclusion of many annotation categories. By implementing a multiple regression model with differential shrinkage via random effects, we avoid reductive assumptions on the number of causal SNPs per locus. Application of this method to a large GWAS meta-analysis of body mass index identified six loci with significant evidence in favor of one or more variants. In an additional 24 loci, one or two variants were strongly prioritized over others in the region. The use of functional annotation in genetic fine-mapping studies helps to distinguish between variants in high LD and to identify promising targets for follow-up studies.

## 33. Model diagnostics of discrete data regression: a unifying framework using functional residuals

Dungang Liu

**Abstract:** Model diagnostics is an indispensable component of regression analysis, yet it is not well addressed in standard textbooks on generalized linear models. The lack of exposition is attributed to the fact that when outcome data are discrete, classical methods (e.g., Pearson/deviance residual analysis and goodness-of-fit tests) have limited utility in model diagnostics and treatment. This paper establishes a novel framework for model diagnostics of discrete data regression. Unlike the literature defining a singlevalued quantity as the residual, we propose to use a function as a vehicle to retain the residual information. In the presence of discreteness, we show that such a functional residual is appropriate for summarizing the residual randomness that cannot be captured by the structural part of the model. We establish its theoretical properties, which leads to the innovation of new diagnostic tools including the functional-residual-vscovariate plot and Function-to-Function (Fn-Fn) plot. Our numerical studies demonstrate that the use of these tools can reveal a variety of model misspecifications, such as not properly including a higher-order term, an explanatory variable, an interaction effect, a dispersion parameter, or a zero-inflation component. The functional residual yields, as a byproduct, Liu-Zhang's surrogate residual mainly developed for cumulative link models for ordinal data (Liu and Zhang, 2018, JASA). As a general notion, it considerably broadens the diagnostic scope as it applies to virtually all parametric models for binary, ordinal and count data, all in a unified diagnostic scheme.

## 34. Pre-Clinical Drug Development: Computational Design, Screening, Formulation and Pharmacokinetics/Pharmacodynamics

Huan Xie

**Abstract:** Preclinical drug development bridges the gap between basic drug discovery and product translation into the clinic. It is a very important step towards to commercialization of novel drugs for all diseases. In recent years, there have been tremendously increased demands for computation drug design and screening, dosage formulation, and pharmacokinetic (PK) and pharmacodynamic (PD) characterizations of lead compounds from researchers in academic institutions and small companies throughout the nation. Traditionally, such drug developmental resources/services are only available through expensive commercial contract research organizations (CROs) with very limited resources available in laboratories scattered throughout Texas. We have filled this critical gap through the expansion of our previous drug development resources and the establishment of the GCC Center for Comprehensive PK/PD & Formulation (CCPF) at Texas Southern University (TSU) with $5.3M support from Cancer Prevention and Research Institute of Texas (CPRIT) in 2018. We have collaborated with over 40 investigators now for more than 100 projects since then. In 2020, we received another $8.63M grant from National Institute on Minority Health and Health Disparity (NIMHD) and built a Center for Biomedical and Minority Health Research (CBMHR), which conducts basic and translation research on diseases impacting minority health as well as substantial community engagement activities in great Houston area. In 2022, we were awarded another contract from National Cancer Institute (NCI) Experimental Therapeutics (NExT) Program to conduct drug discovery and development studies for researchers throughout the United States. As the PI/Director of those centers, I will introduce the capability of our centers, and discuss several projects that we conducted and are now in the commercialization stage, with a focus on computation drug design, formulation and PK/PD development.

## 35. Understanding the functional contributions of epigenetic deregulations in pediatric solid tumors

Xiang Chen

**Abstract:** Pediatric cancers seldom have strong environmental image and are essentially a disease of deregulated development. Pediatric cancer genome project and other cancer genomic studies consistently revealed that pediatric cancer harbors few genetic drivers. On the other side, epigenetic regulations play a critical role in both physiological and pathological development. However, limited availability of high quality tissues for ChIP-seq profiling and lack of functional interpretation of DNA methylome changes remain as major roadblocks to decipher the epigenetic drivers in pediatric tumors. We have developed a deep-learning based approach to interpret the functional consequence of DNA methylation changes. I will further demonstrate the functionality in a real biological application of rhabdomyosarcomas.

## 36. Analyzing Continuous Glucose Monitoring Data for Diabetes

Xiaohua Zhang

**Abstract:** Diabetes mellitus is a chronic metabolic disease associated with long-term damage to various organ systems. The rapid development of continuous glucose monitoring (CGM) brings new insights into diagnosis and treatment of diabetes. CGM is generating a huge amount of data. Because the glucose dynamics in our body is controlled by our complex system, the analysis of the CGM data requires methods from advanced complex systems. These methods include recently developed complexity analytic methods such as multiscale entropy analysis (MSE). Recently, we have not only developed a R package CGManalyzer implementing MSE and other methods to analyze CGM data but also apply these methods to analyze CGM data for diagnosis and treatment of diabetes in clinical studies. We are the first to apply MSE to assess the treatment effect in a pregnant woman with preexisting type 2 diabetes. We have developed algorithms to handling missing values in CGM data. In this presentation, I will introduce MSE methods, explore the CGM metrics most applicable to clinical practice (recommended by the 2019 International Consensus Group), describe our R package CGManalyzer, give examples of our research in continuous monitoring of blood glucose for diabetes.

## 37. Optimal One-pass Nonparametric Estimation Under Memory Constraint

Zhenhua Lin

**Abstract:** For nonparametric regression in the streaming setting, where data constantly flow in and require real-time analysis, a main challenge is that data are cleared from the computer system once processed due to limited computer memory and storage. We tackle the challenge by proposing a novel one-pass estimator based on penalized orthogonal basis expansions and developing a general framework to study the interplay between statistical efficiency and memory consumption of estimators. We show that, the proposed estimator is statistically optimal under memory constraint, and has asymptotically minimal memory footprints among all one-pass estimators of the same estimation quality. Numerical studies demonstrate that the proposed one-pass estimator is nearly as efficient as its non-streaming counterpart that has access to all historical data.

## 38. Orthogonal Common-Source and Distinctive-Source Decomposition Between High-Dimensional Data Views

Hai Shu

**Abstract:** Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of two high-dimensional data views/sets is to decompose each data matrix into three parts: a low-rank common-source matrix that captures the shared information across data views, a low-rank distinctive-source matrix that characterizes the individual information within each single data view, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common-source and distinctive-source matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive-source matrices. The latter guarantees that no more shared information is extractable from the distinctive-source matrices. We propose a novel decomposition method that defines the common-source and distinctive-source matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with a careful construction of the orthogonal relationship between distinctive-source matrices. The proposed estimators of common-source and distinctive-

source matrices are shown to be asymptotically consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and the real data analysis.

## 39. Bayesian Model Assessment for Jointly Modeling Multidimensional Response Data with Application to Computerized Testing

Minghui Chen

**Abstract:** Computerized assessment provides rich multidimensional data including trial-by-trial accuracy and response time (RT) measures. A key question in modeling this type of data is how to incorporate RT data, for example, in aid of ability estimation in item response theory (IRT) models. To address this, we propose a joint model consisting of a two-parameter IRT model for the dichotomous item response data, a log-normal model for the continuous RT data, and a normal model for corresponding pencil-and-paper scores. Then, we reformulate and reparameterize the model to capture the relationship between the model parameters, to facilitate the prior specification, and to make the Bayesian computation more efficient. Further, we propose several new model assessment criteria based on the decomposition of deviance information criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML). The proposed criteria can quantify the improvement in the fit of one part of the multidimensional data given the other parts. Finally, we have conducted several simulation studies to examine the empirical performance of the proposed model assessment criteria, and have illustrated the application of these criteria using a real dataset from a computerized educational assessment program.