R X-AGI IFoDS

第 17 届中国 R 会议
The 17th China-R Conference

2024 X 智能大会
The 2024 X-AGI Conference

2024 数据科学国际论坛
The 2024 International Forum on Data Science

时间
2024.7.20 - 2024.7.22

地点
中国·北京
中国人民大学

欢迎辞

第 17 届中国 R 会议 & 2024 X 智能大会 & 2024 数据科学国际论坛联合会议将于 2024 年 7 月 20-22 日在中国人民大学召开，本次会议由中国人民大学应用统计科学研究中心、中国人民大学统计学院、统计之都和中国商业统计学会人工智能分会主办，由《Journal of Data Science》编辑部和中国人民大学数据科学与大数据统计系承办，得到宽德投资、明汯投资、和鲸科技、子博设计赞助支持。

从 2008 年开始，中国 R 会议在过去的 16 年里一直致力于探讨数据科学在各学科、各行业的探索和实践，先后联合 20 多个院校，在 15 个城市共举办约 50 多场 R 会议，近 2000 个演讲，线上线下累计参会 10 万多人次。话题覆盖机器学习、人工智能、统计理论、文本、语音、图像、天文、地理、心理、社会、医疗、公共卫生、生物信息、互联网、可视化、车联网、自动驾驶、城市规划、量化金融、工业工程、智能制造等领域。

近年来，算法、算力、数据相关技术的迅猛发展拉开了通用人工智能（AGI）的序幕，正在对科学探索、产业进步、社会治理产生深远的影响。

为了更好地介绍和推广先进生产力，促进交流，激发思考，启迪创新，统计之都携手中国商业统计学会人工智能分会发起 X 智能（英文名：XAGI）项目：涵盖开源软件、大会、沙龙、奖项、编程训练和科研支持等多个活动。X 代表探索和交叉，X 智能项目的格言是：“交叉智能、计算未来”——智能技术和各学科、产业深度交叉融合，用计算创未来。X 智能大会引入了更多智能技术相关话题，包括最新的大模型、Agent 技术，以及智能科技在自然科学、社会科学、数学、医疗、金融、互联网等领域的最新实践。

数据科学国际论坛（IFoDS）由中国人民大学统计学院主办，《Journal of Data Science》承办。IFoDS 的核心使命是为来自不同背景的数据科学家（包括研究人员、从业人员、专业人员和学生）提供一个平台，分享对数据科学方法、计算技术和现实应用不断发展的见解，旨在激发数据科学各个领域的协作和创新。

第 17 届中国 R 会议 & 2024 X 智能大会 & 2024 数据科学国际论坛联合会议将携手为我们带来一场数据科学和人工智能共舞的盛会！让我们相约中国人民大学，共赴这场汇聚创新与洞察的盛会！我们欢迎您的到来，一同感受数据科学与人工智能为这个时代带来的惊喜与挑战。

统计之都敬上

2024 年 7 月 20 日

会议日程

会议	嘉宾姓名	演讲题目	时间
开幕式 7 月 20 日上午 主席：王子翀 会场：逸夫楼第一报告厅	吴喜之	开幕致辞	9:00-9:30
	吴芃希		
Keynote 7 月 20 日上午 主席：常象宇&魏太云 会场：逸夫楼第一报告厅	董彬	AI for Mathematics	9:30-10:30
	刘红升	大模型时代下的 AI4Science 发展和设想	10:30-11:30
基座大模型专场 7 月 20 日下午 主席：马诺 会场：立德楼 801	阎栋	From Imitation to Emergence: The Journey of Alignment for LLMs	13:00-13:25
	张舸	高能力全透明开源双语大语言模型 MAP-Neo	13:25-13:50
	吉嘉铭	大模型对齐的机理和高效对齐微调技术	13:50-14:15
	余天予	MiniCPM-V	14:15-14:40
	张松阳	Open-Compass 的大模型评测实践	14:40-15:05
通用大模型应用实践专场 7 月 20 日下午 主席：莫欣 会场：立德楼 801	莫欣	让开发者文思泉涌：一个好的 AI 应用开发框架应该为开发者提供的基本体验	15:30-16:00
	黄志国	langchain-chatchat：开源、可离线部署的检索增强生成(RAG)大模型知识库项目	16:00-16:30
	唐飞虎	长文本应用如何推理加速	16:30-17:00
	冯伟健	如何制造合成数据进行模型 SFT	17:00-17:30
大模型应用行业实践专场 7 月 20 日下午 主席：张源源 会场：立德楼 802	孙一乔	AIxEdu：超级应用的最佳赛道，迈向 AGI 的数学推理练兵场	13:00-13:30
	胡修涵	Building For Fun Agents	13:30-14:00
	Connor Wang	Agent in Action	14:00-14:30
	祝海林	大模型时代下的 AI 数据库 Byzer 和 编程工具 Chat-Auto-Coder	14:30-15:00

大模型基础设施专场 7 月 20 日下午 主席：张先轶 会场：立德楼 802	吴智楷	Modelscope-Agent 开源框架:功能完备且生产可落地的 Agent 框架	15:30-16:00
	张先轶	针对国产硬件优化的 PerfXLM 推理引擎与 PerfXCloud 推理云系统	16:00-16:30
	袁进辉	大模型部署成本降低 10000 倍之路	16:30-17:00
	黄锦涛	SWIFT 工具箱：简化大模型应用之旅	17:00-17:30
数据科学实践专场 7 月 20 日下午 主席：刘思喆 会场：立德楼 803	Kim	GPU 在量化投资中的优势及应用	13:00-13:30
	许以言	面向有组织科研的的模型生命周期管理	13:30-14:00
	刘思喆	因果推断技术在工业界的实践应用	14:00-14:30
	张丹	数据分析落地的最佳实践	14:30-15:00
AI 产业实践专场 7 月 20 日下午 主席：王小宁 会场：立德楼 803	朱赛赛	统计数据大模型的应用场景和解决方案的探索与实践	15:30-16:10
	王小宁	智能教育革命：如何借助大语言模型改善统计和数据科学教学	16:10-16:50
	冯晟洋	人工智能+数字员工企业最佳实践	16:50-17:30
评估因果主效应、调节效应和中介效应的进展专场 7 月 20 日下午 主席：秦旭 会场：立德楼 807	秦旭	A Causal Investigation of Heterogeneity in Mediation Mechanisms in Multisite Randomized Trials	13:00-13:40
	王杰彪	Heterogeneous Causal Mediation Analysis with Bayesian Additive Regression Trees	13:40-14:20
	洪光磊	Organizational Effectiveness: A New Strategy to Leverage Multisite Randomized Trials for Valid Assessment	14:20-15:00
因果推断及其应用的最新发展专场 7 月 20 日下午 主席：李伟 会场：立德楼 807	解海天	Data-driven Policy Learning for a Continuous Treatment	15:30-16:10
	马慧娟	Quantile Regression Models for Compliers in Randomized Experiments with Noncompliance	16:10-16:50
	王林勃	The promises of multiple outcomes	16:50-17:30

IFoDS 会议	嘉宾姓名	演讲题目	时间
开幕式 7 月 21 日上午 主席：Xiaoling Lu 会场：逸夫楼第二会议室	王轶	开幕致辞	9:00-9:30
	王晓军		
	闫军		
Keynote 7 月 21 日上午 主席：Jun Yan &Xiaoling Lu 会场：逸夫楼第二会议室	Songxi Chen	Digital Twin of Economic Systems	9:30-10:30
	Chuanhai Liu	First Principles of Advanced Data Analysis: the Prediction Principle	10:30-11:30
Complex Data Analysis 专场 7 月 21 日下午 主席：Zhibo Cai 会场：立德楼 810	Yue Liu	Quantifying Individual Risk for Binary Outcome:Bounds and Inference	13:00-13:25
	Wenxuan Zhong	MedReader: a query-based multisource AI learner of medical publications	13:25-13:50
	Liping Tong	Statistical Research Projects Using Electronic Health Records	13:50-14:15
	Zhezhen Jin	On detecting the effect of exposure mixture	14:15-14:40
	Ju-Young Park	Fitting an Accelerated Failure Time Model with Time-dependent Covariates via Nonparametric Mixture	14:40-15:05
Efficient Analysis in Statistics and Related Fields 专场 7 月 21 日下午 主席：Chunyan Wang 会场：立德楼 810	Danyang Huang	Subsampling Spectral Clustering for Stochastic Block Models in Large-Scale Networks	15:30-15:55
	Haonan Wang	Recent developments for multi-channel factor analysis	15:55-16:20
	Jie Yang	Statistical Models for Categorical Data Analysis	16:20-16:45
	Ping Ma	Statistical Computing Meets Quantum Computing	16:45-17:10
	Sangbum Choi	Interval-censored linear quantile regression	17:10-17:35

Modern Statistical Methods on Time Series and Functional Data 专场 7 月 21 日下午 主席：Hui Huang 会场：立德楼 811	Wu Wang	A Stock Price Trend Prediction Model Based on Supply Chain Matrix	13:00-13:25
	Jie Li	Testing conditional quantile independence with functional covariate	13:25-13:50
	Zerui Guo	Unified Principal Components Analysis of Irregularly Observed Functional Time Series	13:50-14:15
	Qin Shao	Forecasting Interval for Autoregressive Time Series with trend	14:15-14:40
	Mengyu Xu	Inference for Quantile Change Points in High-Dimensional Time Series	14:40-15:05
Machine Learning and Data Science 专场 7 月 21 日下午 主席：Jie Li 会场：立德楼 811	Feng Zhou	Accelerating Convergence in Bayesian Few-Shot Classification	15:30-15:55
	Zhibo Cai	A Variable Selection Tree and Its Random Forest	15:55-16:20
	Xinyue Wang	U.S.-U.K. PETs Prize Challenge: Anomaly Detection via Privacy-Enhanced Federated Learning	16:20-16:45
	Jiancheng Jiang	Partition-Insensitive Parallel ADMM Algorithm for High-dimensional Linear Models	16:45-17:10
	Sangwook Kang	Deep Neural Network-based Accelerated Failure Time Models Using Rank Loss	17:10-17:35

会议	嘉宾姓名	演讲题目	时间
New Statistical Methods I 专场 7 月 21 日下午 主席：成慧敏 会场：立德楼 812	成慧敏	Network Tight Community Detection	13:00-13:40
	赵博娟	Two variable screening procedures with restrictions on the positive or negative effects	13:40-14:20
	沈梓梁	分布式高维分位数回归：估计效率和支撑恢复	14:20-15:00
New Statistical Methods II专场 7 月 21 日下午 主席：李雪瞳 会场：立德楼 812	师佳鑫	Mixture Conditional Regression with Ultrahigh Dimensional Text Data for Estimating Extralegal Factor Effects	15:30-16:10
	余柏辰	A Gaussian Mixture Model for Multiple Instance Learning with Partially Subsampled Instances	16:10-16:50
	李雪瞳	Gaussian Mixture Model with Rare Event	16:50-17:30
生物统计前沿方法专场 7 月 21 日下午 主席：周静 会场：立德楼 807	李忻月	Functional Adaptive Double-Sparsity Estimator for High-Dimensional Sensor Data Analysis	13:00-13:30
	罗翔宇	Bayesian Integrative Region Segmentation in Spatially Resolved Transcriptomic Studies	13:30-14:00
	孙韬	Enhancing Treatment Strategies and Risk Assessment in Hip Fracture Elderly Patients: A Copula-Based Approach for Semi-Competing Risks Analysis	14:00-14:30
	梅好	Network and Covariate Adjusted Response-Adaptive Design	14:30-15:00

线上会议	嘉宾姓名	演讲题目	时间
医药行业的软件开发（与 DIA 中国统计社区的联动） 7 月 22 日晚上 主席：李扬&魏志军 会场：学说直播平台	段晓丽	What happens when your validated ecosystem is a Graph?	19:00-19:30
	程鼎	Integrating LLM Coding Capabilities in End-to-End Data Science: Challenges and Reflections	19:30-20:00
	曹心怡	Patient Narrative Generation in R	20:00-20:30
	王杰，刘晓畅	双剑合璧: R 和 Python 协同构建数据应用	20:30-21:00
Advancements in Statistical Testing, Estimation, and Design of Experiments 7 月 22 日晚上 主席：王春燕 会场：学说直播平台	张春明	Simultaneous jump detection for multiple sequences via screening and multiple testing	19:00-19:30
	马长兴	Common Odds Ratio Test and Interval Estimation for Stratified Bilateral and Unilateral Data	19:30-20:00
	刘笑	Assessing heterogeneous causal effects across clusters in partially nested designs	20:00-20:30
	王春燕	Construction of strong orthogonal Latin hypercubes	20:30-21:00

中国人民大学地图

中区周边

① 清真餐厅
② 汇贤府
③ 中区西饼屋

东区周边

④ 徽湘阁
⑤ New Pattem餐厅

西区周边

⑥ 学子居
⑦ 小渝府
⑧ 教授餐厅
⑨ 美食厨房

南区周边

⑩ 集天餐厅
⑪ 留二餐厅（营业至7月19日）
⑫ 是饭餐厅（营业至7月19日）

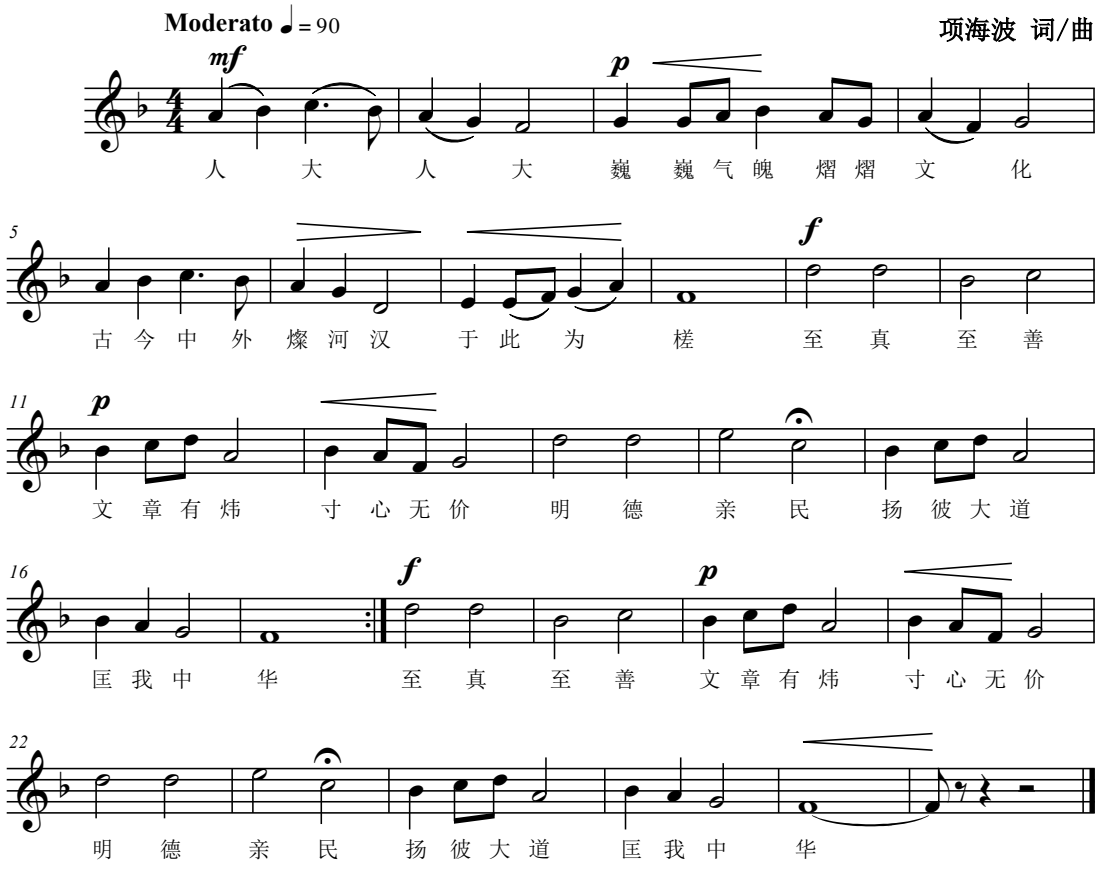
明德商学楼地下一层

⑬ 醉面
⑭ 肯德基

注：上图标记的餐厅均可以微信、支付宝消费。

人 之 大 者

为中国人民大学而作



目录

主办方

中国人民大学应用统计科学研究中心

中国人民大学应用统计科学研究中心是中华人民共和国教育部所属百所人文社会科学重点研究基地之一，成立于 2000 年 9 月，其前身是 1988 年成立的中国人民大学统计科学研究所。中心始终将建立和发展应用统计学科基地作为战略定位，着重从制定应用统计研究的科学规划、密切联系实际选准科研攻关方向、注重研究工作的长期积累、加强重点研究平台建设等方面开展工作。中心着力培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析、风险管理与精算、生物卫生统计、数据科学与大数据统计等五个各具特色的研究方向，围绕各个方向的统计理论创新与应用建设重点研究平台，获得丰硕的研究成果。“十四五”期间，中心将围绕经济社会的数字化转型展开科研攻关，继续为统计学科的发展提供支撑平台。

中国人民大学统计学院

中国人民大学统计学科始建于 1950 年，两年后成立统计学系，是新中国经济学科中最早设立的统计学系，2003 年 7 月，成立中国人民大学统计学院。多年来，本学科一直强调统计理论和统计应用的结合，不断拓宽统计教学和研究领域，成为统计学全国重点学科，在 2012 年、2017 年教育部全国统计学一级学科评估中排名第一。学院拥有统计学一级学科博士点和博士后流动站，拥有经济统计学和风险管理与精算学两个二级学科博士点，拥有预防医学与公共卫生一级学科硕士授权点，统计学、概率论与数理统计、风险管理与精算学、流行病与卫生统计学四个学术型硕士点，应用统计学专业学位硕士点，统计学、经济统计学、应用统计学（风险管理与精算）、数据科学与大数据技术四个本科专业，是全国拥有理学、经济学、医学三大门类统计学专业最齐全的统计学院。

统计之都

统计之都（Capital of Statistics，简称 COS，网址 https://cosx.org/），成立于 2006 年 5 月，是一家旨在推广与应用统计学知识的网站和社区，其口号是“中国统计学门户网站，免费统计学服务平台”。统计之都发源于中国人民大学统计学院，由谢益辉创建，现由世界各地的众多志愿者共同管理维护，理事会现任主席为常象宇。统计之都致力于搭建一个开放的平台，使得科研人员、数据分析人员和统计学爱好者能互相交流合作，一方面促进彼此专业知识技能的增长，另一方面为国内统计学和数据科学的发展贡献自己的力量。

中国商业统计学会人工智能分会

中国商业统计学会人工智能分会
COMMERCE STATISTICAL SOCIETY OF CHINA SECTION ON AI

中国商业统计学会成立于 1987 年，由原商业部、国家粮食局、国家烟草专卖局、中华全国供销合作总社、中石化销售总公司等九大行业部门和单位共同发起，在民政部正式注册的全国性、学术性、非营利性社会组织。学会现隶属于国务院国有资产监督管理委员会，由国家统计局及商务部对其进行业务指导。中国商业统计学会人工智能分会是 2023 年由统计之都、中国人民大学统计学院等数十家机构共同发起并筹备的组织。分会为推动统计和人工智能在学界与业界的交叉互通，为统计和人工智能的从业者提供交流互动平台，为实现未来通用人工智能提供技术与理论支撑。

承办方

Journal of Data Science

Journal of Data Science（JDS）现由中国人民大学统计学院和教育部人文社科重点研究基地应用统计科学研究中心主办。美国康涅狄格大学统计学系教授闫军博士担任主编。为了更好的支持数据科学的发展，JDS 重点关注数据科学的三大支柱方向，计算机科学、统计学、和应用领域的最新前沿问题。期刊常设栏目：Philosophy of Data Science、Statistical Data Science、Computing in Data Science、Data Science in Action、Data Science Review、Education in Data Science 以及 Data Science Conversation。JDS 采取单盲同行评审，每篇文章由 3 位专家把关，我们努力做到投稿文章在 3 个月内收到第一次评审结果，并且所有被接收文章都要求提供数据和代码文档，由期刊代码复现团队对文章的方法进行复现，以确保结果的可重复性。

期刊网址：https://jds-online.org/journal/JDS

投稿地址：https://www.e-publications.org/ruc/sbs/JDS/login

中国人民大学数据科学与大数据统计系

中国人民大学数据科学与大数据统计系

中国人民大学数据科学与大数据统计系成立于 2020 年，使命是培养未来的数据科学家。本系致力于为不同专业背景（比如商业分析、金融科技、健康信息学、工程、数学以及计算机）的学生提供扎实的数据科学知识，努力把握当前数据科学时代的机遇和挑战，发展自身优势，用统计方法解决机器学习、人工智能、大数据分析等领域中的重要问题，发展成为具有持久的区域和全球社会影响的世界一流的数据科学中心。

赞助商

明汯投资

明汯投资于 2014 年在上海虹口对冲基金产业园成立，借助强大的数据挖掘、统计分析和技术开发能力，构建了覆盖全周期、多品种、多策略的资产管理平台。自成立以来，明汯一直致力于成为国际一流量化投资机构，始终秉承“专业谦逊、务实高效、敬畏市场、感恩客户”的经营理念。作为国内最早一批将人工智能技术成功应用到金融市场的私募机构，公司管理规模位居行业前列，并成为国内首批管理规模突破 500 亿的量化私募管理人。经过多年的持续投入和研发，在基础设施硬件及投研框架、交易系统等方面均已构建起行业领先综合优势。

宽德投资

宽德投资是一家国内领先、业务全面的量化对冲基金。基于先进的高频交易构架，以及完善的资产管理系统，宽德投资在国内期货、股票、期权等主流市场具有出色的盈利能力。

和鲸科技

和鲸科技成立于 2015 年，是国内领先的数据智能科技企业，以“Connect People with Data 人与数据的价值连接”为使命，志在与开拓者同行，以“协同平台 + 实践社区 + 竞赛”三位一体的数据科学与人工智能基础设施建设体系，助力各行各业打通数据的价值闭环，实现 AI 赋能应用落地。客户覆盖气象、教育、医疗、航空航天、金融、通信、能源、零售等领域，与众多高校、科研机构、企业等单位展开了深度合作。

子博设计

子博设计，赋能企业增长，成就商业之美。我们致力于成为中国最受信赖的品牌设计公司，用设计的力量为企业赋能，提升企业的核心竞争力，帮助我们的客户成为最杰出的公司。自成立以来，团队服务客户超过 200 家，包括微软、阿里巴巴、腾讯、字节跳动、宁德时代、四季沐歌等客户。我们时刻关注行业发展，积极把握业务与技术结合的机遇，积极尝试 AIGC 为业务场景赋能，从而提升客户服务体验，驱动业务增长。

第 17 届中国 R 会议 & 2024 X 智能大会 & 2024 数据科学国际论坛联合会议筹备委员会

主　席：吴芃希
秘书长：王子翀
秘书团：李怡萱、徐瀚臣、顾正煊、钟轶伦、刘致远

2024 数据科学国际论坛（IFoDS）筹备委员会

Program Committee:
Xiaoling Lu (co-chair), Professor, School of Statistics, Renmin University of China
Jun Yan (co-chair), Professor, Department of Statistics, University of Connecticut
Hui Huang, Professor, School of Statistics, Renmin University of China
Jing Zhou, Associate Professor, School of Statistics, Renmin University of China

Local Organizing Committee:
Feifei Wang (chair), Associate Professor, School of Statistics, Renmin University of China
Hao Mei, Assistant Professor, School of Statistics, Renmin University of China
Chunyan Wang, Assistant Professor, School of Statistics, Renmin University of China
Feng Zhou, Assistant Professor, School of Statistics, Renmin University of China

Web Support Committee:
Jinyu Sun, Graduate Student, School of Statistics, Renmin University of China

Student Volunteers:
Wenxuan Song, Undergraduate Student
Jiaheng Wang, Undergraduate Student
Pengxi Wu, Undergraduate Student

统计之都简介及活动回顾

“统计之都”(Capital of Statistics，简称 COS) 网站成立于 2006 年 5 月 19 日，其主旨为传播统计学知识并将其应用于实际领域。纵观现今国内统计学理论和应用的发展，一方面我们不难发现统计学在应用领域的巨大潜力——现代管理、咨询、商业、经济、金融、医药、生物等等，无不需要数据的力量，而另一方面我们也不得不承认，国内统计学的应用很大程度上受理论的制约——无论是应用界的人们对统计学基础理论知识的欠缺，还是学术界所研究的理论对应用领域问题的轻视。“统计之都”网站便是基于这样的认识而创建的。我们希望，统计理论研究者能充分关注应用问题，而统计应用者也能正确把握统计学基本知识，将统计学这门应用学科真正的潜力开发出来。“统计之都”为非赢利性质网站，但大力欢迎所有商界和研究领域的朋友与我们在实际应用问题上合作。我们的口号是：

中国统计学门户网站，免费统计学服务平台

我们怀着“十年磨一剑”的决心，要将“统计之都”创建成中国的统计学“正直、人本、专业”的社区；我们抱着“己欲立而立人、己欲达而达人”的信条，要将“统计之都”以免费统计学服务平台的形式坚持办下去。我们希望“统计之都”在专业知识体系上有真正的王者风范，在面对用户需求时却又以谦恭的态度为大家服务。统计之都（下文简称 COS）目前由线上与线下两部分构成。其中，线上内容主要包括主站（http://cosx.org/）以及微信公众号（CapStat）；随着越来越多喜爱数据科学的朋友们加入，大家对于线下活动和书稿撰写翻译等等的需求也越来越旺。COS 线下活动总结：

1. 中国 R 会：目前已开展到第十七届，分别在北京、上海、广州、杭州、西安、武汉、成都、贵阳、南昌、厦门、合肥、太原、哈尔滨等地举办。历届会议纪要和幻灯片共享都可以在 COS 主站上找到：http://china-r.org/
2. 线下沙龙：目前我们在北京、上海和广州深圳开展线下沙龙活动。不同于规模庞大的 R 语言会议，沙龙形式更为轻巧，注重讨论交流。目前已经举办过 50 期，目前主要在北京、上海每月举办，详情参见统计之都主站及微信公众号。
3. 海外在线视频沙龙：我们在 Google Hangouts 举办在线沙龙，主要由海外嘉宾来分享学术、生活中的点点滴滴。目前已经举办 23 期：http://meetup.cos.name/.
4. 书籍出版，包括写作和翻译。如《Dynamic Documents with R and knitr》(2ndedition) 谢益辉著，《Implementing Reproducible Research》谢益辉等著，《bookdown: Authoring Books and Technical Documents with R Markdown》谢益辉著,《数据科学中的 R 语言》李舰、肖凯著,《R 语言实战》高涛、肖楠、陈钢翻译,《ggplot2：数据分析与图形艺术》统计之都翻译,《R 语言核心技术手册》刘思喆、李舰、陈钢、邓一硕翻译,《R 语言编程艺术》陈堰平、邱怡轩、潘岚锋等翻译,《R 数据可视化手册》肖楠、邓一硕、魏太云翻译,《R 语言统计入门》邓一硕、郝智恒、何通翻译,《数据科学实战》冯凌秉、王群锋翻译,《R 语言实战》(第 2 版) 王小宁、刘撷芯、黄俊文翻译,《Rcpp: R 与 C++ 的无缝结合》寇强、张哗翻译,《R 绘图系统》呼思乐、张哗、蔡俊翻译,《R 语言编程实战》冯凌秉翻译,《量化投资与 R》（待出版）邓一硕、冯凌秉、杨环翻译,《金融风险建模与投资组合优化》（待出版）邓一硕、郑志勇等翻译、《ggplot2: 数据分析与图形艺术（第 2 版)》黄俊文、王小宁、于嘉傲、冯璟烁,《统计之美：人工智能时代的科学思维》李舰，海恩著等等。

AI for Mathematics

董彬（北京大学）

时间：7.20 9:30-10:30

简介：董彬，北京大学，北京国际数学研究中心教授、国际机器学习研究中心副主任。主要研究领域为机器学习、科学计算和计算成像。2014 年获得求是杰出青年学者奖，2022 年受邀在世界数学家大会（ICM）做 45 分钟报告，2023 年获得新基石研究员项目，同年获得王选杰出青年学者奖。

摘要：本报告将重点关注近年来人工智能在辅助数学探索中的一些进展。首先，我们将回顾人工智能为数学研究赋能的背景和一些发展现状，包括机器学习在激发数学家进行前沿探索中的应用。其次，我们将介绍目前正在进行的一些工作的初步成果。最后，我们将展望人工智能与数学交叉研究领域的未来机遇与挑战。

大模型时代下的 AI4Science 发展和设想

刘红升（华为）

时间：7.20 10:30-11:30

简介：刘红升，中国科学技术大学少年班学院本科，北卡罗莱纳大学教堂山分校统计学博士。现任华为 2012 实验室昇思 MindSpore 架构师/AI4Sci Lab 负责人。基于昇腾 AI 基础软硬件及昇思 MindSpore AI 框架构建了面向 AI4Sci 领域的 MindScience 开源框架，覆盖生物、化学、流体、气象、电磁等多个领域。

摘要：本次报告回顾 AI for Science 在各领域的最新业界进展，并介绍华为 AI4Sci Lab 基于昇腾 AI 基础软硬件及昇思 MindSpore AI 框架在大模型赋能各方向的最新研究与未来展望。

From Imitation to Emergence: The Journey of Alignment for LLMs

阎栋（百川智能）

时间：7.20 13:00-13:25

简介： 阎栋，百川智能强化学习负责人。博士毕业于清华大学计算机系。主要从事决策算法/系统和大语言模型对齐方面的研究。在算法方面，提出了通过奖励分配机制连接无模型和基于模型的强化学习算法的求解框架。在系统方面，作为架构师设计的强化学习编程框架“天授”，在 Github 获得超过 7.4k 星标/1.1k 二次开发。在 ICLR、ICML、IJCAI、AAAI、JMLR、Pattern Recognition 等会议/期刊发表论文十余篇。带领团队基于 RLHF 增强的大语言模型 Baichuan3，在 4 月份的 Superclue 评测中荣获国内第一。

摘要： 大语言模型的对齐技术在过去两年中迅速发展。除了 InstructGPT 所采用的 Supervised Fine Tuning 和 Reinforcement Learning with Human Feedback 方式之外，Rejection Sampling、Direct Preference Optimization、Identity-Preference Optimization 等方法纷纷涌现，为各种目标和条件下的行业落地提供了丰富的工具。但想要用好这些对齐工具，不仅需要对了解各种方法的底层数学原理，而且需要辅以坚实的工程支持。本次分享从对齐技术的理论图景开始，深入对齐技术的工程实践进行讨论，以展望对齐技术的未来收尾。通过对对齐技术全景式的回顾和讨论，帮助听众了解对齐技术的挑战并在业务场景落地。1. Theoretical Landscape of Alignment；2. Practical Data-Centric Process；3. Scaleable Oversight and Beyond。

高能力全透明开源双语大语言模型 MAP-Neo

张舸（零一万物）

时间：7.20 13:25-13:50

简介： 张舸，加拿大滑铁卢大学博士生，M-A-P 社区发起人，COIG 系列工作的发起人。

摘要： 第一个工业级的透明中英文双语大模型-Neo 的开源，我们提供了全部的 4.7T 预训练数据，训练 pipeline，基于 spark 的预训练数据 pipeline，OCR pipeline，以及复现 deepseek-math 提出的迭代地从预训练数据中召回高质量数据的直接可用的 pipeline。我们的模型在 7B 大小，与 OLMo 和 Amber 相比，Neo 作为基座的性能基本达到了可比工业级 SOTA 的水准。

大模型对齐的机理和高效对齐微调技术

吉嘉铭（北京大学）

时间：7.20 13:50-14:15

简介：北京大学人工智能研究院博士生，导师是杨耀东助理教授，主要从事大模型的安全与价值对齐方面的研究，获得首批国家自然科学基金青年学生基础研究项目 (博士研究生) 资助，北京大学校长奖学金获得者，详见个人主页：jijiaming.com。

摘要：围绕大模型的对齐机理和高效对齐微调技术展开汇报。从机理上探究：模型是否抗拒对齐，拒绝被改变，模型在预训练塑造的意图与价值是否能够在对齐阶段被修改。相比预训练阶段的数据量和参数更新次数，对齐所需的数据量和参数更新显著更少。即使是经过精细化对齐的模型也很容易被故意或无意地规避。本报告探讨了大模型参数中是否存在弹性，以及对齐是否真正改变了模型的内在特性，还是仅仅只是表面对齐，以及如何不通过 RLHF 实现高效的大模型对齐技术。

MiniCPM-V

余天予（清华大学）

时间：7.20 14:15-14:40

简介：清华大学自然语言处理实验室博士生，主要从事多模态大模型相关工作。

摘要：迈向实用多模态大模型的路上存在许多阻碍，MiniCPM-V 系列模型通过 MiniCPM 高效 scaling law，VisCPM 跨语言泛化、RLHF-V 和 RLAIF-V 可信行为学习、LLaVA-UHD 高清图编码等技术在端侧实现了接近 GPT-4V 级别的效果。

Open-Compass 的大模型评测实践

张松阳（上海人工智能实验室）

时间：7.20 14:40-15:05

简介：上海人工智能实验室青年研究员，OpenCompass 技术负责人

摘要：评测是大模型研发的指南针，如何全面、科学、客观的评测大模型的能力是产学研各界都关心的重点问题。OpenCompass 旨在从能力体系、工具链、评测数据和模型榜单多个维度对大模型评测进行体系建设和技术研发。本演进将介绍 OpenCompass 在大模型评测上的具体实践和相关思考。

让开发者文思泉涌：一个好的 AI 应用开发框架应该为开发者提供的基本体验

莫欣（智体纪元（*Agently.cn*））

时间：7.20 15:30-16:00

简介：北京智体纪元科技有限公司创始人，Agently AI 应用开发框架项目负责人，前光年之外开发者生态产品经理。

摘要：从利用 Agently 开发框架开发的开源项目说起，带领开发者逐层拆解 Agently AI 应用开发框架在帮助开发者完成项目开发落地过程中，从模型单次请求的能力放大脚手架，到代码级工作流编排管理能力，这些不同的特性都将如何帮助开发者顺畅、高效将思路转化为高可用、高质量的业务代码。

langchain-chatchat：开源、可离线部署的检索增强生成 (RAG) 大模型知识库项目

黄志国（*LangchainChatChat*）

时间：7.20 16:00-16:30

简介：南开大学精算学博士，浙江大学科学技术研究院博士后，langchain-chatchat 核心开发组成员。

摘要：一种利用 langchain 思想实现的基于本地知识库的问答应用，目标期望建立一套对中文场景与开源模型支持友好、可离线运行的知识库问答解决方案。依托于本项目支持的开源 LLM 与 Embedding 模型，本项目可实现全部使用开源模型离线私有部署。与此同时，本项目也支持 OpenAI GPT API 的调用，并将在后续持续扩充对各类模型及模型 API 的接入。 本项目实现原理如下图所示，过程包括加载文件 -> 读取文本 -> 文本分割 -> 文本向量化 -> 问句向量化 -> 在文本向量中匹配出与问句向量最相似的 top k 个 -> 匹配出的文本作为上下文和问题一起添加到 prompt 中 -> 提交给 LLM 生成回答。

长文本应用如何推理加速

唐飞虎（月之暗面）

时间：7.20 16:30-17:00

简介：月之暗面高级研发工程师、开发者关系负责人。

摘要：推理性能是目前长文本大模型的瓶颈之一，本文介绍各种可利用在长文本模型中的推理加速技术，重点介绍目前在实践中使用将多的方法。在典型的 AI 工作流中，您可能会将相同的输入令牌反复传递给模型。使用上下文缓存功能，您可以将一些内容传递给模型一次，缓存输入令牌，然后引用缓存的令牌以用于后续请求。在某些数量下，使用缓存的令牌比重复传入同一语料库的令牌更低的费用（并且延迟更低）。

如何制造合成数据进行模型 SFT

冯伟健（明日知己）

时间：7.20 17:00-17:30

简介：为明教育集团 AI 部门负责人、明日知己教育科技 CTO、香港中文大学大一休学。

摘要：模型 SFT 是一个非常重要的工作，但通常受制于数据量的大小和质量会影响到 SFT 效果。本分享将重点介绍 SFT 数据的不同分类、如何合成可以使用的 SFT 数据、以及合成数据与真实数据的配比

AIxEdu：超级应用的最佳赛道，迈向 AGI 的数学推理练兵场

孙一乔（悉之智能）

时间：7.20 13:00-13:30

简介： 悉之智能创始人，致力于用 AI 革新教育行业，7 年 AI 解题和讲解经验，AI 教师大模型多次刷新行业 SOTA，广泛应用于国内外，获得启明、经纬、真格、新东方等一线 VC 3000 万美元融资。旗下北美产品用户超过百万，ARR 超百万美金，积累千万题目数据，App Store 评分 4.8+，独有行业最高的解题率和多模态互动 AI 讲解功能。国内赋能教育公司 AI 升级，与新东方优编程合作开发 U-shannon 大模型，与紫光合作推出 AI 答疑普惠平台，致力于用 AI 实现真正的个性化、普惠教育。

摘要： AGI 的超级应用时代即将到来，中国的超级应用将引领全球。教育行业因其刚需性，是最有可能率先跑出超级应用的赛道之一。数学推理能力是通向通用人工智能（AGI）的关键，而教育行业恰恰是 LLM 提升数理推理能力的最佳练兵场。最重要的是，AI 对教育的变革，非常契合并支持国家政策和发展方向，能够让每个孩子都拥有普惠的个性化 AI 教师。

Building For Fun Agents

胡修涵（捏 *Ta*）

时间：7.20 13:30-14:00

简介： 北京大学智能科学本科，哥伦比亚大学硕士，机器人实验室研究生。曾任 Facebook 视频产品 Tech Lead，阿里巴巴研发团队负责人，特赞（上海）信息科技有限公司技术副总裁，系统性发布特赞内容数字资产管理系统（DAM）并带领团队完成产品收入过亿。2022 年创办看见概念（上海）智能科技有限公司，打造 AI 驱动的幻想创作平台“捏 Ta”。

摘要： AGI 时代的娱乐应用以 For Fun Agents 作为基础服务单元，而其中 Agents 的打造和建设可能需要大众广泛的参与。如何通过系统性的建设角色内容框架和事件，构造出未来最值得与之互动和可以高效产出优质内容的 Agents，是捏 Ta 这个平台的核心价值。

Agent in Action

Connor Wang（*Six*）

时间：7.20 14:00-14:30

简介： Founder & CEO @ Six AI.

摘要： Action Agent 的实现与应用. 如何将基于 AI 应用与移动互联网的产品和经济生态整合, 让 LLM 的“动脑”能力赋能 classic engineering 的“动手”能力. 让 AI 帮你做每天无聊又必须做的事情。

大模型时代下的 AI 数据库 Byzer 和编程工具 Chat-Auto-Coder

祝海林（*Kyligence*）

时间：7.20 14:30-15:00

简介： Byzer 社区 PMC/资深数据架构师/Kyligence 技术合伙人，拥有 15+ 年研发经验。一直专注在 Data + AI 融合方向上，致力于帮助企业更好的落地 Data+AI。个人热衷于开源产品的设计和研发，Byzer/MLSQL 为其主要开源作品，最新产品 auto-coder 超越自动代码补全的编程工具，旨在帮助企业获得倍数级别研发效率提升。Byzer AI 数据库获得 22 年中国开源创新大赛二等奖，23 年浦东新区人工智能创新大赛一等奖，个人入选中国 22 年开源先锋 33 人，荣获 23 年全球人工智能开发者先锋大会「开发者先锋」称号。

摘要： 大模型是 AI 发展的一个里程碑，它正在改变社会的方方面面。Byzer AI 数据库，使用 SQL 作为交互语言，创新性可以将主流大模型注册成 UDF 来使用，同时具备预训练，微调，部署大模型能力，帮助企业快速的在诸如 ETL, 数据分析，流式计算（风控）以及 APP 应用等各种场景中使用大模型。Chat-Auto-Coder 可以帮助用户实现对已有项目（包括 SQL 类项目）的阅读和迭代，用户甚至可以不打开编辑器即可完成对代码的修改和测试。

Modelscope-Agent 开源框架: 功能完备且生产可落地的 Agent 框架

吴智楷（阿里巴巴通义实验室）

时间：7.20 15:30-16:00

简介： 阿里巴巴魔搭社区 Modelscope-Agent 框架开发者。

摘要： 在大模型时代，Agent 作为未来可能的落地场景，受到了广泛的关注。为了满足复杂多变的用户需求和生产场景，Modelscope-Agent 框架积极与 Modelscope 等开源社区生态结合，提供了包括可定制 Agent，开源大语言模型支持，API 服务集成，分布式多智能体任务等完备的功能。用户仅需若干行代码，即可定制自己的 Agent，并在相应场景中落地使用。

针对国产硬件优化的 PerfXLM 推理引擎与 PerfXCloud 推理云系统

张先轶（澎峰科技）

时间：7.20 16:00-16:30

简介： 张先轶, 本科和硕士毕业于北京理工大学, 博士毕业于中国科学院大学, 曾于中科院软件所工作，之后分别在 UT Austin 和 MIT 进行博士后研究工作。国际知名开源矩阵计算项目 OpenBLAS 发起人和主要维护者。中国计算机学会高性能计算专业委员会委员, ACM SIGHPC China 执行委员。2016 年, 创办 PerfXLab 澎峰科技, 提供异构计算软件栈与解决方案。获得 2016 年中国计算机学会科学技术二等奖，2017 年中国科学院杰出科技成就奖，2020 年美国 SIAM Activity Group on Supercomputing 最佳论文奖，2023 年北京市自然科学二等奖，2023 Bench Council 世界开源贡献奖。

摘要： 当前智算中心 N 卡与非 N 卡的混合架构下，如何充分发挥国产卡计算资源，达到可用、好用、高效利用是亟需解决的问题。我们提出了 PerfXLM 大模型推理引擎与 PerfXCloud 推理云系统，主要针对多种国产 GPU 和 NPU 硬件进行适配与优化，已经完成语言模型，embedding 模型等多种主流模型的迁移与适配。

大模型部署成本降低 10000 倍之路

袁进辉（硅基流动）

时间：7.20 16:30-17:00

简介：袁进辉, 2003 年于西安电子科技大学 (Xidian University) 计算机专业获得学士学位, 2008 年于清华大学计算机系获得工学博士学位,清华大学优秀博士学位论文奖获得者,2008 2011 年在清华博士后期间开展计算神经科学方面的研究,2013 2016 年他任微软亚洲研究院主管研究员 (Lead Researcher), 负责研发大规模机器学习系统 LightLDA 并服务于微软产品。2016 年 2023 年, 他发起和主导研发了开源深度学习框架 OneFlow，在分布式深度学习系统编程易用性和高效性方向设计了一系列新方法，并为工业界广泛采用。目前他的研究领域为 AI Infrastructure，致力于通过算法、系统、硬件协同设计研发大模型推理加速引擎，降低大模型应用成本和开发门槛。

摘要：ChatGPT 背后的大模型技术最为新一轮的技术变革已经成为共识，各种 AI 原生应用呼之欲出，预期在不久的未来，AI 将在我们的工作和生活中无处不在。目前限制 AI 应用发展的一个主要因素是大模型部署的成本，这次分享将和大家探讨如何解决目前 AI 应用快速爆发与算力资源短缺推理成本昂贵的矛盾，是否有机会将大模型推理成本降低 10000 倍，加速 AGI 时代的到来。

SWIFT 工具箱：简化大模型应用之旅

黄锦涛（阿里巴巴通义实验室）

时间：7.20 17:00-17:30

简介：魔搭社区 swift 框架开发者

摘要：当前，大语言模型和多模态大模型正逐步成为推动技术创新和应用的关键力量。然而，如何有效整合这些多元的模型，特别是在多模态领域，以提供简洁且统一的使用接口，对许多从业者而言是一项棘手的挑战。为此，我们提供了 SWIFT：一个旨在简化大模型使用的工具箱。SWIFT 支持 250+ 大语言模型和 35+ 多模态大模型的微调、人类反馈对齐、推理、评估、量化和部署，包括：Qwen、Llama、GLM、Internlm、Yi、百川、DeepSeek、Llava 等系列模型。除此之外，我们构建了丰富的 Adapter 库，汇集了包括 LoRA+、GaLore、Llama-Pro 在内的最新训练技术，作为 PEFT 轻量级训练方案的补充。同时，SWIFT 提供了基于 Gradio 的 Web-UI 界面和众多最佳实践，帮助研究者和开发者轻松上手大模型的微调与应用。

GPU 在量化投资中的优势及应用

Kim（头部量化私募）

时间：7.20 13:00-13:30

简介： Kim 就职于头部量化私募，负责量化交易低延时，高性能计算系统的有关开发工作。

摘要： 2007 年英伟达发布 CUDA 编程范式以来，经过 17 年的发展，GPU 在算力和显存都已经逐步远超通用 CPU 的能力。量化投资领域一直走在技术的最前沿，原有用 CPU 来进行的高性能计算的程序，也逐步在切换到使用 GPU 来加速的模式。这里将介绍日常工作中 GPU 的应用场景，实际开发中遇到的问题，以及分享 GPU 提升对应业务效率的具体案例。

面向有组织科研的的模型生命周期管理

许以言（和鲸科技）

时间：7.20 13:30-14:00

简介： 许以言，和鲸科技产品专家，专注于数据驱动研究与 AI for Science 场景的数据科学平台产品设计与方法创新，参与了 ModelWhale 数据科学协同平台在气象、地质、遥感、空间科学、临床等众多科研智能领域的落地，对数据智能场景的多角色协同研究流程有独到的见解与丰富的经验积累。

摘要： 随着有组织科研的快速发展，数据信息与数据价值正以更高维的形式体现在模型中，数据分析的过程也需要由多领域专家参与其中，本报告将围绕空间数据智能分析场景的模型生命周期管理流程，介绍 ModelOps 方法，并探讨面对交叉领域研究场景，如何通过平台化的工具与社区化的方法支撑有组织科研。

因果推断技术在工业界的实践应用

刘思喆（统计之都）

时间：7.20 14:00-14:30

简介： 刘思喆，统计之都理事会成员。先后在彩票、电信、电商、教培、交通、餐饮行业从事算法、数据科学、营销赋能等相关工作。曾任 51Talk 数智中心助理副总裁、首席数据科学家，也曾任京东推荐平台部高级经理，京东技术名人堂成员之一。中国人民大学大数据分析实验班、首经贸信息学院校外硕士生导师。国内 R 语言的布道者，21 年的使用经验，《153 分钟学会 R》的作者，《R 语言核心技术手册》的译者。

摘要： 本报告围绕工业界中因果推断的核心价值展开，探讨其在产品优化、市场策略调整、供应链管理等业务中的重要性。本报告也将尝试系统梳理常见的因果推断技术，包括随机实验、倾向得分匹配、断点回归分析、合成控制等方法，并探讨它们之间的内在联系、适用场景及其潜在局限。通过剖析企业中的真实业务案例，我们将生动展示，如何利用这些方法提炼出精准的因果洞见，持续赋能企业的高质量决策的完整过程。

数据分析落地的最佳实践

张丹（北京青萌数海科技有限公司）

时间：7.20 14:30-15:00

简介： 张丹，R 语言实践者，北京青萌数海科技有限公司 CTO，微软 MVP。10 年以上互联网应用架构经验，在 R、大数据、数据分析等方面有深厚的积累。精通量化投资交易策略，熟悉中国金融二级市场、交易规则和投研体系。熟悉数据学科方法论，在海关、药监、外汇等监管科技领域均有落地项目。著有《R 的极客理想：量化投资篇》、《R 的极客理想：工具篇》、《R 的极客理想：高级开发篇》，图书英文版被 CRC 出版集团引进，在美国发行。个人博客：http://fens.me 。

摘要： 现在我们正处于大数据时代，处处都产生数据，大部分数据已经不在稀缺，分析方法和算法模型都也写在了教课书中。如何挖掘出数据的价值，让数据分析落地，把数据价值转换为自身价值，是数据分析师核心要考虑的。数据分析要解决实际业务场景问题，伪需求、不清晰的目标，都会造成项目失败。数据分析不只是指标体系、更不是指标堆积，市场在变，数据也在变，我们的知识结构也要跟着变化。数据分析是跨学科的工作，对人的要求也越来越高，调包侠的时代已过。要以新的视角，看数据、看业务、看技术发展、看我们自己，适应变化，才能把项目做好、落地。

统计数据大模型的应用场景和解决方案的探索与实践

朱赛赛（同方知网）

时间：7.20 15:30-16:10

简介：朱赛赛，同方知网图书工具书与志鉴产品总监。2014 年加入中国知网，2019 年至今负责经济社会大数据产品的运营、市场推广及项目支持等工作，为国内外千余家高校、科研及企事业单位提供服务和支持。基于多年在农业部、统计局等系统的项目合作实践，积累了丰富的统计数据采集、治理、管理及业务应用经验。

摘要：本次报告介绍了以应用和服务为导向如何构建统计数据治理体系，并基于多维度、高质量海量统计数据与华知大模型，进行数据问答、数据解读、专业数据分析模型、数据分析报告的规划设计，用以解决数值型数据的深度应用。

智能教育革命：如何借助大语言模型改善统计和数据科学教学

王小宁（中国传媒大学）

时间：7.20 16:10-16:50

简介：王小宁，现为中国传媒大学数据科学与智能媒体学院副教授，大语言模型智能体书卷侠负责人，硕士生导师，中国商业统计学会理事，中国人民大学中国调查与数据中心研究员，中国商业统计学会人工智能分会秘书长，统计之都秘书长，中国人民大学统计学博士，研究方向为大语言模型、抽样设计、统计机器学习和文本挖掘。

摘要：本次分享将探讨如何利用大语言模型来革新统计和数据科学的课程教学，将从传统统计教学方法的挑战和局限开始讨论，引入数字化教育的重要性。重点介绍中国传媒大学数据科学教学团队围绕大语言模型 +Agent 在这领域的一些探索，同时介绍基于大语言模型的 AI 助教-书卷侠（https://scholarhero.cn/），展示其如何通过智能化解答和个性化教学资料来提升教学效果和学习体验。我们还会探讨这种技术在《数据科学导论》等课程中的具体应用，并展望未来教育技术的发展趋势，讨论这些新技术在教育实践中的潜在应用，以及它们对未来数据科学教育格局的深远影响。

人工智能 + 数字员工企业最佳实践

冯晟洋（上海蓝衫科技有限公司）

时间：7.20 16:50-17:30

简介：上海蓝衫科技有限公司 Blueshirt Technology 的联合创始人，GPT 元宇宙创始人、渗透智能–ShirtAI 创始人

摘要：

- 市场痛点：企业用工成本不断增加，员工难以处理海量数据和繁琐任务
- 解决方案：构建数字机器人扮演数字员工帮助企业降本增效，扮演数字助理提高个人生产力
- 业务介绍：提供 AI 应用开发和服务，以 NLP 作为大脑，RPA 作为双手，为组织和个人构建数字机器人
- 商业模式：本地化 + 远程部署
- 竞争优势：为广泛客户提供更具实操性、更具性价比的数字化转型方案和应用

官网：https://www.bluelsqkj.com/robot-development

A Causal Investigation of Heterogeneity in Mediation Mechanisms in Multisite Randomized Trials

秦旭（*University of Pittsburgh*）

时间：7.20 13:00-13:40

简介： Dr. Xu Qin is an Assistant Professor of Research Methodology at the School of Education (primary) and an Assistant Professor of Biostatistics at the School of Public Health (secondary). She holds a Ph.D. from the Department of Comparative Human Development at the University of Chicago and a B.S. and an M.S. in Statistics from the Renmin University of China. Her research focuses on solving cutting-edge methodological problems in causal mediation analysis and multilevel modeling. She is also interested in using rigorous and innovative quantitative methods to evaluate the impacts of interventions and the underlying mechanisms. Methodologically, she has developed statistical methods and software for investigating the heterogeneity in causal mediation mechanisms in both multilevel and single-level settings, as well as sensitivity analysis and power analysis methods for causal mediation analysis. Substantively, she is interested in applying advanced statistical methods in developmental, educational, and health research. Dr. Qin has served as the Principal Investigator or Co-Principal Investigator for grants funded by the Spencer Foundation, the National Science Foundation, and the U.S. Department of Education's Institute of Education Sciences. She is a recipient of the 2024 NSF CAREER award and the 2022 National Academy of Education/Spencer Postdoctoral Fellowship.

摘要： Multisite randomized trials have been pervasive in the past three decades. The importance of investigating the variation in the total impact of an intervention has become increasingly valued. An intervention may generate heterogeneous impacts due to natural variations in participant characteristics, context, and local implementation. Important research questions include whether the intervention impact is generalizable across individuals and contexts, for whom and under what contexts the intervention is effective, and why. To advance this line of research, this study develops a method to assess the mediation mechanism underlying the total impact of the intervention in multisite randomized trials and how it varies by individual and contextual factors. The findings may help practitioners improve and tailor intervention designs and implementations for different individuals and contexts. The method is evaluated through comprehensive Monte Carlo simulations. It is also applied to the National Study of Learning Mindsets for evaluating the mediation mechanism underlying the impact of a growth mindset intervention on math performance and its heterogeneity.

Heterogeneous Causal Mediation Analysis with Bayesian Additive Regression Trees

王杰彪（*University of Pittsburgh*）
时间：7.20 13:40-14:20

简介： Assistant Professor of Biostatistics and Clinical and Translational Science at the University of Pittsburgh

摘要： Causal mediation analysis can help explain the mechanism of how an exposure affects an outcome. The mediation effects are often heterogeneous based on individual characteristics, but most existing methods ignore this heterogeneity and estimate the population average effects. To address this gap, we develop a heterogeneous causal mediation analysis method using Bayesian Regression Tree Ensembles. Distinct from traditional methods, our approach captures complex non-linear interactions and heterogeneous effects in mediation processes more flexibly, offering a refined understanding of the heterogeneity of causal mechanisms. By sampling from the posterior trees of mediator and outcome models, we are able to obtain rigorous credible intervals for causal mediation effects. We also use partial dependent plots to illustrate which moderators play more important roles and how each effect changes with a moderator. Utilizing simulated datasets, we demonstrate the superiority of our approach in accurate estimation and inference of heterogeneous mediation effects, especially in scenarios characterized by non-linear relationships and interaction effects.

Organizational Effectiveness: A New Strategy to Leverage Multisite Randomized Trials for Valid

洪光磊（*University of Chicago*）
时间：7.20 14:20-15:00

简介： Guanglei Hong is Professor in the Department of Comparative Human Development (https://humdev.uchicago.edu/) at the University of Chicago. She was the Inaugural Chair of the University-wide Committee on Quantitative Methods in Social, Behavioral, and Health Sciences (https://voices.uchicago.edu/qrmeth/) and is a member of the Committee on Education (https://voices.uchicago.edu/coed/). She attained a master's degree in Applied Statistics in 2002 and a Ph.D. in Education in 2004 from the University of Michigan. Before joining the University of Chicago faculty in July 2009, she had been an Assistant Professor in the Human Development and Applied Psychology Department in the Ontario Institute for Studies in Education of the University of Toronto (OISE/UT). Prof. Hong has focused her research on developing causal inference theories and methods for understanding the impacts of large-scale societal changes and the effects of social

and educational policies and programs on child and youth development. She has contributed original concepts and developed multiple methods for drawing valid inferences about causal relationships, for investigating heterogeneity in responses to external interventions across individuals and contexts, and for rigorously testing theories about the mechanisms through which such exposures generate impacts. Her book "Causality in a social world: Moderation, mediation, and spill-over" was published by Wiley in 2015. She guest edited the Journal of Research on Educational Effectiveness special issue on the statistical approaches to studying mediator effects in education research in 2012. Additionally, through publishing in first-tier statistics, education, psychology, sociology, and public policy journals and disseminating new methods through workshops and training institutes, her research has generated a broad impact among quantitative methodologists as well as applied researchers. She has received research and training grants from the National Science Foundation, the U.S. Department of Education, the William T. Grant Foundation, the Spencer Foundation, and the Social Sciences and Humanities Research Council of Canada among other funding agencies. She was awarded a prestigious John Simon Guggenheim Memorial Foundation Fellowship in 2021. For more information, please visit her website: https://humdev.uchicago.edu/directory/guanglei-hong.

摘要： In education, health, and human services, an intervention program is usually implemented by many local organizations. Determining which organizations are more effective is essential for theoretically characterizing effective practices and for intervening to enhance the capacity of ineffective organizations. In multisite randomized trials, site-specific intention-to-treat (ITT) effects are likely invalid indicators for organizational effectiveness and may lead to inequitable decisions. This is because sites differ in their local ecological conditions including client composition, alternative programs, and community context. Applying the potential outcomes framework, this study proposes a mathematical definition for the relative effectiveness of an organization. The estimand contrasts the performance of a focal organization with those that share the features of its local ecological conditions. The identification relies on relatively weak assumptions by leveraging observed control group outcomes that capture the confounding impacts of alternative programs and community context. Simulations demonstrate significant improvements when comparing with site-specific ITT analyses or analyses that adjust for between-site differences in the observed baseline participant composition only. We illustrate its use through an evaluation of the relative effectiveness of individual Job Corps centers by reanalyzing data from the National Job Corps Study, a multisite randomized trial that included 100 Job Corps centers nationwide serving disadvantaged youths. The new strategy promises to alleviate severe misclassifications of some of the most effective Job Corps centers as least effective and vice versa.

Data-driven Policy Learning for a Continuous Treatment

解海天（北京大学）

时间：7.20 15:30-16:10

简介： 解海天 2023 年毕业于美国加州大学圣地亚哥分校。主要研究方向为因果推断理论，包括工具变量、断点回归等因果推断方法的非参数/半参数识别与估计，以及基于因果模型的政策分析评估、策略学习与统计决策等。研究成果发表于 Journal of Business and Economic Statistics, Oxford Bulletin of Economics and Statistics 等国际期刊。

摘要： This paper studies policy learning under the condition of unconfoundedness with a continuous treatment variable. Our research begins by employing kernel-based inverse propensity-weighted (IPW) methods to estimate policy welfare. We aim to approximate the optimal policy within a global policy class characterized by infinite Vapnik-Chervonenkis (VC) dimension. This is achieved through the utilization of a sequence of sieve policy classes, each with finite VC dimension. Preliminary analysis reveals that welfare regret comprises of three components: global welfare deficiency, variance, and bias. This leads to the necessity of simultaneously selecting the optimal bandwidth for estimation and the optimal policy class for welfare approximation. To tackle this challenge, we introduce a semi-data-driven strategy that employs penalization techniques. This approach yields oracle inequalities that adeptly balance the three components of welfare regret without prior knowledge of the welfare deficiency. By utilizing precise maximal and concentration inequalities, we derive sharper regret bounds than those currently available in the literature. In instances where the propensity score is unknown, we adopt the doubly robust (DR) moment condition tailored to the continuous treatment setting. In alignment with the binary-treatment case, the DR welfare regret closely parallels the IPW welfare regret, given the fast convergence of nuisance estimators.

Quantile Regression Models for Compliers in Randomized Experiments with Noncompliance

马慧娟（华东师范大学）

时间：7.20 16:10-16:50

简介： 华东师范大学统计学院与统计交叉科学研究院副教授。中国科学技术大学统计学博士，美国埃默里大学博士后。主要研究方向包括生存分析,分位数回归,因果推断等。在统计学期刊《Biometrika》,《Biometrics》,《Journal of Business & Economic Statistics》和《Statistica Sinica》等期刊发表论文二十余篇。曾主持国家自然科学基金青年项目一项和上海市浦江人才项目一项。现主持国家自然科学基金重点项目子项目一项，参与国家自然科学基金重点项目及科技部重点研发项目等。担任中国现场统计研究会生存分析分会理事。

摘要： Understanding the causal effect of a treatment in randomized experiments with noncompliance is of fundamental interest in many domains. Utilizing the instrumental variable (IV) framework, compliers are the only subpopulation that closely relevant to the assessment of causal treatment effect. In this pape we study flexible quantile regression models for compliers with and without treatment. We establish unbiased estimating equations by investigating the relationship between observed data and latent subgroup indicators. A novel iterated algorithm is proposed to solve the discontinuous equations that involve unknown parameters in a complicated way. The complier average treatment effect and quantile treatment effects can be estimated. The consistency and asymptotic normality of the proposed estimators are established. Numerical results, including extensive simulation studies and real data analysis of the Oregon health insurance experiment, are presented to show the practical utility.

The promises of multiple outcomes

王林勃（*University of Toronto*）

时间：7.20 16:50-17:30

简介： Linbo Wang is an associate professor in the Department of Statistical Sciences and the Department of Computer and Mathematical Sciences, University of Toronto. He is also a faculty affiliate at the Vector Institute, a CANSSI Ontario STAGE program mento and affiliated with the Department of Statistics, University of Washington, and Department of Computer Science, University of Toronto. Prior to these roles, he was a postdoc at Harvard T.H. Chan School of Public Health. He obtained his Ph.D. from the University of Washington. His research interest is centered around causality and its interaction with statistics and machine learning.

摘要： A key challenge in causal inference from observational studies is the identification and estimation of causal effects in the presence of unmeasured confounding. In this pape we introduce a novel approach for causal inference that leverages information in multiple outcomes to deal with unmeasured confounding. The key assumption in our approach is conditional independence among multiple outcomes. In contrast to existing proposals in the literature, the roles of multiple outcomes in our key identification assumption are symmetric, hence the name parallel outcomes. We show nonparametric identifiability with at least three parallel outcomes and provide parametric estimation tools under a set of linear structural equation models. Our proposal is evaluated through a set of synthetic and real data analyses.

Digital Twin of Economic Systems

Songxi Chen (Tsinghua University)

时间： 7.21 9:30-10:30

简介: Dr. Songxi Chen is an Academician of the Chinese Academy of Sciences. He is currently serving as the President of the Chinese Society for Probability and Statistics for the term 2023-2026. Dr. Chen earned his Ph.D. in Statistics from the Australian National University in 1993. Prior to his full-time return to China, he held faculty positions at the National University of Singapore and Iowa State University. From 2010 to 2019, Dr. Chen served as the Founding Director of the Center for Statistical Science at Peking University. His research interests are diverse and include high-dimensional data inference, environmental modeling and assessment, empirical likelihood, statistical and machine learning, and stochastic process inference. Dr. Chen is a Fellow of the Institute of Mathematical Statistics (IMS), the American Statistical Association, and the American Association for the Advancement of Science. He is also an elected member of the International Statistical Institute (ISI).

摘要: A digital twin of a system is a high-precision numerical simulation based on the integration of system models and observational data, representing the pinnacle of understanding of that system. I will discuss the importance and feasibility of establishing a digital twin for the Chinese economic system, as well as the requirements for high spatiotemporal resolution economic datasets and the development of large-scale econometric models.

First Principles of Advanced Data Analysis: the Prediction Principle

Liu Chuanhai (Purdue University)

时间： 7.21 10:30-11:30

简介: Chuanhai Liu earned his correspondence diploma from Central China Normal University in 1985, master's degree in Probability and Statistics from Wuhan University in 1987, and PhD in Statistics from Harvard University in 1994. He worked at Bell Laboratories for ten years starting in 1995 and at Texas A&M as an Associate Professor in Spring 2024. Since 2005, he has been a Professor of Statistics at Purdue University. His research interests include the foundations of statistical inference, statistical computing, and applied statistics. Much of his work on iterative algorithms, such as Quasi-Newton, EM, and MCMC methods, is discussed in his book titled "Advanced Markov Chain Monte Carlo Methods" (2010), co-authored with F. Liang and R. J. Carroll. His work on the foundations of statistical inference, developing a new inferential framework for prior-free probabilistic inference,

is included in his book titled ”Inferential Models: Reasoning with Uncertainty” (2015), co-authored with R. Martin. For his research on statistical computing, he spent several years experimenting with a multi-threaded and distributed R software system called SupR for big data analysis. Currently, he is working on topics for a potential new book titled ”Scientific Modeling: Principles, Methods, and Examples.”

摘要: This era of big data is fascinating for data analysis in particular and statistics in general. It has also clearly revealed more than ever different scientific attitudes toward data analysis and statistical research from different perspectives. As statisticians, we see both challenges and responsibility for foundational developments in both statistical inference and scientific modeling. This talk introduces a new principle, called the prediction principle. We argue that this principle can serve as a first principle for valid and efficient inference by exploring its implications in three key research directions: (a) how the prediction principle can be used to refine both the principle of maximum likelihood and the likelihood principle, (b) how statistical inference should be formalized, as the required reasoning is deductive, and (c) how a general theory of scientific modeling might be achievable, despite the inherent challenges of inductive reasoning. These discussions are illustrated using seemingly simple but unsolved problems in high-dimensional statistics and deep learning models. To prompt deeper reflections, the talk concludes with a few challenging problems.

Quantifying Individual Risk for Binary Outcome: Bounds and Inference

Liu Yue (Renmin University of China)

时间: 7.21 13:00-13:25

简介: 刘越, 中国人民大学讲师, 2019 年博士毕业于北京大学。多篇文章发表于 Journal of Machine Learning Research（JMLR）, Artificial Intelligence（AIJ）, IEEE Transactions on Knowledge and Data Engineering（TKDE）, IEEE Transactions on Neural Networks and Learning Systems（TNNLS）, International Conference on Machine Learning（ICML）, Knowledge Discovery and Data Mining（KDD）, The Conference on Uncertainty in Artificial Intelligence（UAI）等机器学习与统计学期刊及会议。研究兴趣主要包括因果推断，贝叶斯网络以及基于因果推断的机器学习算法等。

摘要: Understanding treatment heterogeneity is crucial for reliable decision-making in treatment evaluation and selection. While the conditional average treatment effect (CATE) is commonly used to capture treatment heterogeneity induced by covariates and design individualized treatment policies, it remains an averaging metric within subpopulations. This limitation prevents it from unveiling individual-level risks, potentially leading to misleading results. This article addresses this gap by examining individual risk for binary outcomes, specifically focusing on the fraction negatively affected (FNA) –a metric assessing the percentage of individuals experiencing worse outcomes with treatment compared to control. Under the strong ignorability assumption,FNA is unidentifiable, and we find that previous Fréchet-Hoeffding bounds are usually wide and unattainable in practice. By introducing a plausible positive correlation assumption for the potential outcomes, we obtain significantly improved bounds compared to previous studies. We show that even with a positive and statistically significant CATE, the lower bound on FNA can be positive, i.e., in the best-case scenario many units will be harmed if receiving treatment. Additionally, we establish a nonparametric sensitivity analysis framework for FNA using the Pearson correlation coefficient as the sensitivity parameter thereby exploring the relationships among the correlation coefficient, FNA, and CATE. We also present a practical and tractable method for selecting the range of correlation coefficients. Furthermore, we propose flexible estimators for the refined FNA bounds and prove their consistency and asymptotic normality. Extensive simulations are conducted to evaluate the effectiveness of the proposed estimators. We apply our method to the right heart catheterization (RHC) data to explore the percentage of patients harmed by RHC.

MedReader: a query-based multisource AI learner of medical publications

Wenxuan Zhong (University of Georgia)

时间： 7.21 13:25-13:50

简介： Dr. Zhong is an Athletic Association Professor in the Department of Statistics at the University of Georgia. She holds a B.S. in Statistics from Nankai University, China, and a Ph.D. in Statistics from Purdue University. After completing her Ph.D., Dr. Zhong pursued a postdoctoral fellowship in Statistics and Computational Biology at Harvard University. She served as an Assistant Professor in the Department of Statistics at the University of Illinois at Urbana-Champaign from 2007 to 2013, before joining the University of Georgia in 2013. Dr. Zhong is an ASA Fellow and an elected Fellow of the International Statistical Institute. She is the co-Director of the big data analytics lab.

摘要： As the volume and velocity of medical publications have increased at an unprecedented pace, a computational-based learning system is essential to avoid expensive and time-consuming human annotations which in general hinders the deployment of novel therapeutic methods in clinical practice. To achieve this goal, we develop MedReader, a novel multi-channel learning system that can summarize (topic learning), understand (knowledge-graph constructing), and generalize (hypothesis generating) knowledge simultaneously from query-related publications. As with human learners, MedReader can assess how faithfully a discovered concept is by using data beyond publications and conducting a novel enrichment analysis. We applied MedReader to a COVID-19 related publication set, which includes 4,117 abstracts that are deposited into the MEDLINE database from 1/1/2020 to 4/30/2020. The hypotheses generated from the 4,117 publications significantly overlapped with the hypotheses that appeared in subsequent publications. For example, 71% of the predicted gene-gene interactions and 100% of the predicted disease-disease interactions are enriched in subsequent articles. Moreover, the whole learning process only takes 3 minutes—a negligible time-frame for clinical practice. Our analysis shows that this system can help us to learn from publications at an unprecedented speed and scale. Such a learning system not only helps us promptly summarize but also affords opportunities for discovery.

Statistical Research Projects Using Electronic Health Records

Liping Tong (Advocate Aurora Healthcare)

时间： 7.21 13:50-14:15

简介： Liping Tong is currently a senior statistician in Advocate Aurora Health, leading a team

of research and analysis. Liping got her B.A. in 1997 from the Department of Mathematics, Nankai University. She had two years of graduate school in Nankai before going to the Department of Statistics, University of Chicago in 1999. Liping got her PhD in statistics in 2004 and started to work as a research associate in the Department of Statistics, University of Washington. Starting from 2007, she became an Assistant Professor Department of Mathematics, Loyola University Chicago. In 2010, she switched to the Department of Public Health Sciences, Loyola University, Stritch School of Medicine. In 2015, she started her career in Advocate Aurora Healthcare, as a senior statistician. The main responsibilities are:

1. Lead the development of prediction models based on millions of patients' electronic medical records for questions such as readmission risk or chronic disease management. Statistical and computational methods, such as logistic models, hierarchical models, survival analysis, support vector machine, random forest and boosting methods, are used to optimize predictions.
2. Lead the analysis on the evaluation of interventions to reduce adverse events such as emergency department visits and 30-day readmissions after hospitalization. Cox Proportional Hazard models with time dependent covariates are applied in the analysis.
3. Mentor interns, junior statisticians, and data analysts on multiple projects, including evaluation of the program of Palliative Care, application of deep learning and big data strategy in medical science, and so on.
4. Involve in other team members' projects as a reliable source of expert support.

In addition, Liping has an active collaboration with the professors from the Department of Psychiatry, University of Illinois at Chicago since 2020. The main interest is in the data collected for the Chicago Follow-up Study (CFS) that was designed as a naturalistic prospective longitudinal, multi-follow-up research study to investigate the course, outcome, symptomatology, effects of medication, and recovery in participants with serious mental illness disorders. Statistical methods, such as logistic generalized estimating equation (GEE) models, the latent class analysis (LCA), network analysis and clustering methods, have been applied for a wide range of hypotheses of interest.

摘要: With the advent of electronic medical records (EMR), hospitals find themselves overwhelmed with vast quantities of patient data with diverse applications. Given the critical nature of medical data storage and utilization, numerous specialized companies such as Epic, Oracle, and Cerner have emerged. Moreove hospitals typically employ their own cadre of experts including statisticians, data analysts, and data scientists. Data analysis in hospitals spans a spectrum, ranging from fundamental tasks like data summarization and demonstration using tables and plots to more intricate efforts involving the refinement and creation of statistical methods and models. In this presentation, I will illustrate the necessity of connecting time-dependent survival models with logistic models through a compelling example. Additionally, I will underscore the significance of selecting the most suitable analytical tool to maximize insights from data, drawing from a concrete case study.

On detecting the effect of exposure mixture

Zhezhen Jin (Columbia University)

时间: 7.21 14:15-14:40

简介: Zhezhen Jin is Professor of Biostatistics in the Department of Biostatistics in Mailman School of Public Health at Columbia University. He received his BS and MS in probability and statistics from Nankai University in 1989 and in 1992 respectively, MA in applied mathematics from the University of Southern California in 1994 and Ph.D. degree in Statistics from Columbia University in 1998. After 1998-2000 two years of postdoctoral studies at Harvard School of Public Health, he returned to Columbia as a faculty member in the Department of Biostatistics in 2000. He has been conducting statistical and biostatistical methodological research on resampling methods, survival analysis, nonparametric and semiparametric methods, smoothing methods, and statistical computing. He has also been collaborating with clinical investigators to address statistical issues in neurology, cardiology, oncology, transplantation, psychiatry, pathology and alternative medicine. He was a co-founding editor of the Contemporary Clinical Trials Communication. He is Statistical Editor for the Journal of American Cardiology College—Cardiovascular Imaging. He has served as an associate editor for several statistical journals including Journal of American Statistical Association, Statistica Sinica, Lifetime Data Analysis, Communications for Statistical Applications and Methods, Journal of Statistical Theory and Practice, and is on the editorial board for Kidney International, the Journal of the International Society for Nephrology. He received Career Award from the National Science Foundation in 2002. He is a Fellow of the American Statistical Association, a Fellow of the Institute of Mathematical Statistics, and an elected member of International Statistical Institute. He served as the President of the International Chinese Statistical Association (ICSA) in 2022.

摘要: To study the effect of exposure mixture on the continuous health outcomes, one can use the linear model with a weighted sum of multiple standardized exposure variables as an index predictor and its coefficient for the overall effect. The unknown weights typically range between zero and one, indicating contributions of individual exposures to the overall effect. Because the weight parameters present only when the parameter for overall effect is non-zero, testing hypotheses on the overall effect can be challenging, especially when the number of exposure variables is above two. This paper presents a working model based approach to estimate the parameter for overall effect and to test specific hypotheses, including two tests for detecting the overall effect and one test for detecting unequal weights when the overall effect is evident. The statistics are computationally easy and one can apply existing statistical software to perform the analysis. A simulation study shows that the proposed estimators for the parameters of interest may have better finite sample performance than some other estimators.

Fitting an Accelerated Failure Time Model with Time-dependent Covariates via Nonparametric Mixture

Ju-Young Park (Yonsei University)

时间: 7.21 14:40-15:05

简介： BS in Statistics, Seoul National University, South Korea, 2001 PhD in Biostatistics, University of North Carolina at Chapel Hill, 2007 Assistant Professor, University of Georgia, US (2007-2010), University of Connecticut, US (2010 - 2013) Assistant, Associate, Full Professor, Yonsei University, South Korea (2013 -)

摘要: An accelerated failure time (AFT) model is a popular regression model in survival analysis. It models the relationship between the failure time and a set of covariates via a log link with an addition of a random error. The model can be either parametric or semiparametric depending on the degree of specification of the error distribution. The covariates are usually assumed to be fixed—'time independent'. In many biomedical studies, however, 'time-dependent' covariates are frequently observed. In this work, we consider a semiparametric time-dependent AFT model. We assume that the distribution of the baseline failure time is an infinite scale mixture of Gaussian densities. Thus, this model is highly flexible compared to those that assume a one-component parametric density. We consider a maximum likelihood estimation and propose an algorithm based on the constrained Newton method for estimating model parameters and mixing distributions. The proposed methods are investigated via simulation studies to assess the finite sample properties. The proposed methods are illustrated with a real data set.

Subsampling Spectral Clustering for Stochastic Block Models in Large-Scale Networks

Danyang Huang (Renmin University of China)

时间： 7.21 15:30-15:55

简介： 黄丹阳：中国人民大学统计学院教授，博士生导师。主持国家自然科学基金面上项目，北京市社会科学基金重点项目等省部级及以上科研课题，入选北京市科协青年人才托举工程，曾获北京市优秀人才培养资助。长期从事复杂网络建模、超高维数据分析、分布式计算等方向的理论研究，以及统计理论研究在中小微企业信用风险评估，企业数字化发展中的应用研究。在 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *Journal of Econometrics*, *Journal of Business & Economic Statistics* 等国内外权威期刊发表论文 30 余篇。

摘要： The rapid development of science and technology has generated large amounts of network data, leading to significant computational challenges for network community detection. A novel subsampling spectral clustering algorithm is proposed to address this issue, which aims to identify community structures in large-scale networks with limited computing resources. The algorithm constructs a subnetwork by simple random subsampling from the entire network, and then extends the existing spectral clustering to the subnetwork to estimate the community labels for entire network nodes. As a result, for large-scale datasets, the method can be realized even using a personal computer. Moreover, the proposed method can be generalized in a parallel way. Theoretically, under the stochastic block model and its extension, the degree-corrected stochastic block model, the theoretical properties of the subsampling spectral clustering method are correspondingly established. Finally, to illustrate and evaluate the proposed method, a number of simulation studies and two real data analyses are conducted.

Recent Developments for Multi-channel Factor Analysis

Haonan Wang (Colorado State University)

时间： 7.21 15:55-16:20

简介： Haonan Wang received his Ph.D. degree in statistics from the University of North Carolina at Chapel Hill in 2003. Currently, he is a Professor of Statistics at Colorado State University. His research interests are in object-oriented data analysis, functional dynamic modeling of neuron activities, spatial and spatio-temporal modeling, and statistical learning.

摘要： As modern data collection techniques evolve, complex and inhomogeneous data are frequently collected from multiple sources with unobserved interference and idiosyncratic noise. Multi-channel factor analysis, introduced by Ramírez et al. (2020), allows for the extraction of low-dimensional latent factors that highlight the commonalities across various channels as well as identify

unique structures within each channel. In this talk, we discuss some of the important properties of the MFA, including identifiability and the asymptotic behavior of the quasi-Gaussian maximum likelihood estimators. Furthermore, we extend this framework to model time series data, incorporating both temporal and spatial dependencies.

Statistical Models for Categorical Data Analysis

Jie Yang (University of Illinois at Chicago)

时间： 7.21 16:20-16:45

简介： Jie Yang is a professor in the Department of Mathematics, Statistics, and Computer Science at the University of Illinois at Chicago. He received his Ph.D. in Financial Mathematics from Nankai University in 2001 and his Ph.D. in Statistics from the University of Chicago in 2006. His research focuses on statistical methods, financial mathematics, bioinformatics, and big data statistical analysis. His research achievements include rapid classification methods for biological macromolecules, high-dimensional data statistical classification methods, optimal experimental design theory and applications, real-time pricing methods for financial derivatives, and big data sampling analysis methods.

摘要: Categorical responses, whose measurement scale consists of a set of categories, arise naturally in many different scientific disciplines. Multinomial logistic models have been widely used in the literature, which cover four kinds of logit models: baseline-category (also known as multiclass logistic regression model), cumulative, adjacent-categories, and continuation-ratio logit models. We propose a unified multinomial link model for analyzing categorical responses. It not only covers the existing multinomial logistic models and their extensions as special classes, but also allows the observations with NA or Unknown responses to be incorporated as a special category in the data analysis. We provide explicit formulas for computing the likelihood gradient and Fisher information matrix, as well as detailed algorithms for finding the maximum likelihood estimates of the model parameters. Our algorithms solve the infeasibility issue of existing statistics software on estimating parameters of cumulative link models. The applications to real datasets show that the proposed multinomial link models can fit the data significantly better and the corresponding data analysis may correct the misleading conclusions due to missing data.

Statistical Computing Meets Quantum Computing

Ping Ma (University of Georgia)

时间： 7.21 16:45-17:10

简介： Professor Ma is a Distinguished Research Professor in the Department of Statistics at the University of Georgia and co-director of the big data analytics lab. He was a Beckman Fellow at the Center for Advanced Study at the University of Illinois at Urbana-Champaign, a Faculty Fellow at the US National Center for Supercomputing Applications, and a recipient of the National Science Foundation CAREER Award. His paper won the best paper award from the Canadian Journal of Statistics in 2011. He delivered the 2021 National Science Foundation Distinguished Lecture. Professor Ma serves on multiple editorial boards. He is a Fellow of the American Association for the Advancement of Science and the American Statistical Association.

摘要: The recent breakthroughs in quantum computers have shown quantum advantage (aka quantum supremacy), i.e., quantum computers outperform classic computers for solving a specific problem. These problems are highly physics-oriented. A more relevant fact is that there are already general-purpose programmable quantum computing devices available to the public. A natural question for statisticians is whether these computers will benefit statisticians in solving some statistics or data science problems. If the answer is yes, what kind of statistics problems should statisticians resort to quantum computers? Unfortunately, the general answer to this question remains elusive.

In this talk, I will present challenges and opportunities for developing quantum algorithms. I will introduce a novel quantum algorithm for a statistical problem and demonstrate that the intersection of statistical computing and quantum computing is an exciting and promising research area. The development of quantum algorithms for statistical problems will not only advance the field of quantum computing but also provide new tools and insights for solving challenging statistical problems.

Interval-censored linear quantile regression

Sangbum Choi (Korea University)

时间： 7.21 17:10-17:35

简介： Dr. Sangbum Choi received his Ph.D degree in Statistics in 2010 from the University of Wisconsin at Madison. He was an assistant professor in Biostatistics at The University of Texas Health Science Center at Houston and now he is a full professor in Statistics at Korea University. His research interest covers semiparametric methods in survival analysis, joint modeling, longitudinal data analysis, and actuarial data science.

摘要: Censored quantile regression has emerged as a prominent alternative to classical Cox' s proportional hazards model or accelerated failure time model in both theoretical and applied statistics. While quantile regression has been extensively studied for right-censored survival data, methodologies for analyzing interval-censored data remain limited in the survival analysis literature. This paper introduces a novel local weighting approach for estimating linear censored quantile regression, specifically tailored to handle diverse forms of interval-censored survival data. The estimation equation and the corresponding convex objective function for the regression parameter can be constructed as a weighted average of quantile loss contributions at two interval endpoints. The weighting components are nonparametrically estimated using local kernel smoothing or ensemble machine learning techniques. To estimate the nonparametric distribution mass for interval-censored data, a modified EM algorithm for nonparametric maximum likelihood estimation is employed by introducing subject-specific latent Poisson variables. The proposed method' s empirical performance is demonstrated through extensive simulation studies and real data analyses of two HIV/AIDS datasets.

A Stock Price Trend Prediction Model Based on Supply Chain Matrix

Wu Wang (Renmin University of China)

时间：7.21 13:00-13:25

简介： Wu Wang is a lecturer in the Department of Mathematical Statistics at Renmin University of China, holding a Ph.D. in Mathematical Statistics from Fudan University. His primary research interests include functional data analysis, spatial data analysis, machine learning, and deep learning methods applied in energy and industrial domains. His work has been published in journals such as *Biometrics* and the *Scandinavian Journal of Statistics*.

摘要： This work explores the integration of industry chain network matrices into graph neural network models to enhance the predictive ability of deep learning factors for future stock returns. Historically, subjective investors have predominantly utilized industry chain analysis methods but have been constrained by data limitations, preventing their full utilization in quantitative investment. With natural language processing technology's maturation, data providers can extract relationships between companies and products from annual reports, combining expert knowledge to construct industry chain upstream and downstream relationships. Based on this foundation, we compute a matrix of interrelatedness between listed companies derived from the industry chain. Subsequently, this matrix is introduced into the graph neural network model as prior information. Experimental results demonstrate that our proposed model outperforms the baseline GRU model in terms of predictive performance on the test set, with significantly increased IC mean values and decreased IC standard deviations. This finding is consistent with existing research, while the differences in the stock pool and graph structure information selected in this study contribute as a supplement to the field. Additionally, this research extensively explores and explains the model structure, lookback periods, training labels, and other factors through numerous experiments.

Testing conditional quantile independence with functional covariate

Jie Li (Renmin University of China)

时间：7.21 13:25-13:50

简介： Jie Li is a lecturer at the School of Statistics, Renmin University of China. He obtained his Ph.D. in Statistics from Tsinghua University in 2022. His research interests include functional data analysis and time series analysis. He is currently leading a National Natural Science Foundation

of China (NSFC) Young Scientist Project and a China Postdoctoral Science Foundation Project. His work has been published in journals such as *Biometrics* and *Statistica Sinica*.

摘要: We propose a new nonparametric conditional independence test for a scalar response and a functional covariate over a continuum of quantile levels. We build a Cramer–von Mises-type test statistic based on an empirical process indexed by random projections of the functional covariate, effectively avoiding the "curse of dimensionality" under the projected hypothesis which is almost surely equivalent to the null hypothesis. The asymptotic null distribution of the proposed test statistic is obtained under some mild assumptions. The asymptotic global and local power properties of our test statistic are then investigated. We specifically demonstrate that the statistic is able to detect a broad class of local alternatives converging to the null at the parametric rate. Additionally, we recommend a simple multiplier bootstrap approach for estimating the critical values. The finite-sample performance of our statistic is examined through a number of Monte Carlo simulation experiments. Finally, an analysis of an EEG data set is used to show the utility and versatility of our proposed test statistic.

Unified Principal Components Analysis of Irregularly Observed Functional Time Series

Zerui Guo (Sun Yat-sen University)

时间: 7.21 13:50-14:15

简介： Zerui Guo is a Ph.D. student at the School of Mathematics, Sun Yat-sen University, focusing on functional data analysis and epidemiological modeling. His work has been published in journals such as *European Journal of Epidemiology* and *Chinese Journal of Preventive Medicine*.

摘要: Irregularly observed functional time series (FTS) are increasingly available in many real-world applications. To analyze FTS, it's crucial to account for both serial dependencies and the irregularly observed nature of functional data. However, existing methods for FTS often rely on specific model assumptions in capturing serial dependencies or cannot handle the irregular observational scheme of functional data. To address these issues, we propose a novel dimension reduction method for FTS based on the framework of dynamic functional principal component analysis (FPCA). Through a new concept called optimal functional filters, we unify the theories of FPCA and dynamic FPCA, providing a parsimonious and optimal representation for FTS that adapts to its serial dependence structure. This framework, referred to as principal analysis via dependency-adaptivity (PADA), is established under a hierarchical Bayesian model for dimension reduction via FPCA. Our method accommodates both sparsely and densely observed FTS and is capable of predicting future functional data. We investigate the theoretical properties of PADA and demonstrate its effectiveness through extensive simulation studies. Finally, we illustrate our method through dimension reduction and prediction of daily PM2.5 data.

Forecasting Interval for Autoregressive Time Series with Trend

Qin Shao (University of Toledo)

时间： 7.21 14:15-14:40

简介： Dr. Qin Shao received her bachelor's and master's degrees from Nankai University in 1990 and 1993, respectively. She completed her doctoral studies in Statistics at the University of Georgia in 2002 and joined the University of Toledo as an Assistant Professor of Statistics the same year. She was promoted to the rank of Professor in 2013. Her research interests encompass both the methodology and applications of statistics, with a focus on semi-parametric time series modeling and using statistics to address societal issues.

摘要: We propose a kernel distribution estimator (KDE) for the cumulative distribution function of autoregressive time series with trend. We demonstrate that under certain assumptions, this estimator performs as efficiently as an infeasible KDE that assumes the trend is known. The oracular KDE is utilized to estimate quantiles for constructing a forecasting interval. Simulation studies confirm the asymptotic properties of the KDE estimator. To illustrate the method's application, we apply it to monthly average hourly wages data.

Inference for Quantile Change Points in High-Dimensional Time Series

Mengyu Xu (University of Central Florida)

时间： 7.21 14:40-15:05

简介： Mengyu Xu received her Bachelor' s Degree in Statistics from Renmin University of China, Beijing, China in 2010. She received the M.S. and Ph.D. degrees from the Department of Statistics at the University of Chicago, Chicago, USA in 2012 and 2016, respectively. Her research interests include covariance matrix estimation, time-varying network recovery from high-dimensional time series, and the distribution theory of quadratic forms and high-dimensional hypotheses testing.

摘要: Change-point detection methods based on quantiles are effective for identifying changes in extreme values. In this study, we propose a novel change-point detection scheme using fixed quantiles of moving sums from high-dimensional time series data. Our approach utilizes a moving sum (MOSUM) test statistic that aggregates the component series using the ℓ^∞ norm. We analyze the asymptotic properties of our test statistic under weak temporal dependence in high-dimensional time series, allowing for both strong and weak cross-sectional dependence. Our analysis employs a uni-

form Bahadur representation result extended to the high-dimensional setting for dependent data. We demonstrate the effectiveness of our approach through a simulation study.

Accelerating Convergence in Bayesian Few-Shot Classification

Feng Zhou (Renmin University of China)

时间： 7.21 15:30-15:55

简介： Feng Zhou is a lecturer at the School of Statistics, Renmin University of China, and an Outstanding Young Scholar at Renmin University of China. He has led projects supported by the National Natural Science Foundation of China, the China Postdoctoral Science Foundation (special fund, general fund), and has been selected for the Postdoctoral International Exchange Program introduction project. His primary research interests include statistical machine learning, Bayesian methods, stochastic processes, spatiotemporal data analysis, among others. He has published over 20 papers in journals and conferences such as JMLR, MLJ, STCO, NeurIPS, ICLR, AAAI, AISTATS.

摘要: Bayesian few-shot classification has become a focal point in few-shot learning research. This paper integrates mirror descent-based variational inference into Gaussian process-based few-shot classification to address the challenge of non-conjugate inference. Leveraging non-Euclidean geometry, mirror descent accelerates convergence by providing the steepest descent direction along the corresponding manifold. It exhibits parameterization invariance concerning the variational distribution. Experimental results demonstrate competitive classification accuracy, improved uncertainty quantification, and faster convergence compared to baseline models. The study also investigates the impact of hyperparameters and model components.

A Variable Selection Tree and Its Random Forest

Zhibo Cai (Renmin University of China)

时间： 7.21 15:55-16:20

简介： Zhibo Cai is currently a lecturer at the School of Statistics, Renmin University of China, specializing in data science, big data statistics, sufficient dimension reduction, variable selection, and their applications in machine learning. His research has been published in academic journals and conferences such as JASA, NeurIPS, ICLR.

摘要: A novel screening approach is proposed by partitioning the sample into subsets sequentially and creating a tree-like structure of sub-samples called the SIS-tree. SIS-tree is straightforward to implement and can be integrated with various measures of dependence. Theoretical results are established to support this approach, including its "sure screening property". Additionally, SIS-tree is extended to a forest with improved performance. Through simulations, the proposed methods are demonstrated to have great improvement comparing with existing SIS methods. The selection of a cutoff for the screening is also investigated through theoretical justification and experimental study.

As a direct application of the screening, the classification of high-dimensional data is considered, and it is found that the ranking and cutoff can substantially improve the performance of existing classifiers.

U.S.-U.K. PETs Prize Challenge: Anomaly Detection via Privacy-Enhanced Federated Learning

Xinyue Wang (Rutgers University)

时间： 7.21 16:20-16:45

简介： Xinyue Wang received her Ph.D. from Rutgers University in Newark, NJ, USA. Her research interests lie in the interdisciplinary areas of data privacy and security, deep learning, and their applications in various fields such as bioinformatics and finance.

摘要: Privacy Enhancing Technologies (PETs) have the potential to enable collaborative analytics without compromising privacy. The U.S. and U.K. governments partnered in 2021 for the PETs prize challenge to seek privacy-enhancing solutions for financial crime prevention and pandemic response. This article presents Rutgers ScarletPets, a privacy-preserving federated learning approach for identifying anomalous financial transactions in a payment network system (PNS). The approach uses a two-step anomaly detection methodology: mining features based on account-level data and labels, and augmenting these features using a privacy-preserving encoding scheme within the PNS. In the second step, the PNS learns a highly accurate classifier from the augmented data. The approach ensures minimal accuracy loss between federated and centralized settings, and allows the PNS to continually improve its model without additional computational or privacy burdens on banks. ScarletPets won the first prize in the U.S. for its privacy, utility, efficiency, and flexibility.

Partition-Insensitive Parallel ADMM Algorithm for High-dimensional Linear Models

Jiancheng Jiang (University of North Carolina)

时间： 7.21 16:45-17:10

简介： Dr. Jiancheng Jiang is Professor of statistics at the Department of Mathematics and Statistics & School of Data Science, University of North Carolina at Charlotte. His research interests include Financial Econometrics, Theoretical and Applied Statistics, Biostatistics, and Data Science.

摘要: The parallel alternating direction method of multipliers (ADMM) algorithms have gained popularity in statistics and machine learning due to their efficient handling of large sample data

problems. However, the parallel structure of these algorithms, based on the consensus problem, can lead to an excessive number of auxiliary variables when applied to high-dimensional data, resulting in a large computational burden. In this paper, we propose a partition-insensitive parallel framework based on the linearized ADMM (LADMM) algorithm to solve nonconvex penalized high-dimensional regression problems. Our algorithm does not rely on the consensus problem, significantly reducing the number of variables updated at each iteration. It remains robust regardless of how the total sample is divided, demonstrating partition-insensitivity. We establish convergence under mild assumptions and validate the algorithm through numerical experiments on synthetic and real datasets. We also provide an R software package for easy implementation of our proposed algorithm.

Deep Neural Network-based Accelerated Failure Time Models Using Rank Loss

Sangwook Kang (Yonsei University)

时间: 7.21 17:10-17:35

简介： Sangwook Kang received his BS in Statistics from Seoul National University, South Korea, in 2001, and his PhD in Biostatistics from the University of North Carolina at Chapel Hill in 2007. He served as Assistant Professor at the University of Georgia, US (2007-2010), and the University of Connecticut, US (2010-2013), before joining Yonsei University, South Korea, where he is currently an Associate Professor.

摘要: An accelerated failure time (AFT) model assumes a log-linear relationship between failure times and a set of covariates. Unlike other survival models that focus on hazard functions, AFT models directly interpret the effects of covariates on failure times, which is intuitive. The semiparametric nature of AFT models, which do not assume a specific error distribution, offers flexibility and robustness against distributional assumptions. Despite these advantages, existing AFT models typically assume linear predictors for the mean, neglecting non-linear relationships. Deep neural networks (DNNs) have shown remarkable success in capturing non-linear relationships across various domains. In this work, we propose DeepR-AFT, a DNN-based AFT model using Gehan-type loss and sub-sampling techniques to handle non-linear predictors. We evaluate the finite sample properties of DeepR-AFT through extensive simulations and demonstrate its superiority over parametric and semiparametric AFT models in scenarios involving non-linear predictors and large covariate dimensions. We validate the performance of DeepR-AFT using three real datasets.

Network Tight Community Detection

成慧敏（*Boston University*）
时间：7.21 13:00-13:40

简介： I am an Assistant Professor in the Department of Biostatistics at Boston University. I am affiliated with the Rafik B. Hariri Institute for Computing and Computational Science Engineering and Nanotechnology Innovation Center at Boston University. I received my Ph.D. in statistics from the University of Georgia in 2023.

摘要： Conventional community detection methods often categorize all nodes into clusters. However, the presumed community structure of interest may only be valid for a subset of nodes (named as "tight nodes"), while the rest of the network may consist of noninformative "scattered nodes". For example, a protein-protein network often contains proteins that do not belong to specific biological functional modules but are involved in more general processes, or act as bridges between different functional modules. Forcing each of these proteins into a single cluster introduces unwanted biases and obscures the underlying biological implication. To address this issue, we propose a tight community detection (TCD) method to identify tight communities excluding scattered nodes. The algorithm enjoys a strong theoretical guarantee of tight node identification accuracy and is scalable for large networks. The superiority of the proposed method is demonstrated by various synthetic and real experiments.

Two Variable Screening Procedures with Restrictions on the Positive or Negative Effects

赵博娟（天津财经大学）
时间：7.21 13:40-14:20

简介： 参会申请人毕业于南开大学数学系 (数理统计专业, 博士)，曾在美国美国南卫理公会大学（Southern Methodist University）和美国哈佛大学 (Harvard School of Public Health) 做过博士后研究，在美国 Meharry Medical College 工作，现在天津财经大学工作（教授、博导)。

摘要： In this paper, two variable screening procedures, the local significant forward and backward procedure with restrictions on the positive or negative effects (FBRPN) and the backward procedure with restrictions on the positive or negative effects (BRPN), are proposed to obtain meaningful protective and risk factors in fast and sequential ways in models with a linear component such as the GLMs to avoid multicollinearity. The two fitted models from the procedures are compared to obtain the most efficient model and the representative variables of the original predictors. The new procedures are compared with stepwise and best subsets regression in three illustration examples. Simulation

studies are carried out to get some insights on how different covariance structures can affect the final fitted models obtained from the procedures. Cross-validation comparisons with stepwise, LASSO and LAR methods are made based on the Efron diabetes data. Finally, practical issues are discussed, and applications of the new procedures in big data analysis are envisioned.

分布式高维分位数回归：估计效率和支撑恢复

沈梓梁（上海财经大学）

时间：7.21 14:20-15:00

简介：我目前在上海财经大学统计与管理学院攻读统计学博士学位。我的导师是王绍立副教授。我对统计学和机器学习理论充满热情，特别分布式计算领域。此前，我曾在南昌大学获得学士学位。

摘要：本报告深入探讨了高维线性分位数回归问题中的分布式估计和支持恢复技术。分位数回归作为一种对异常值和数据异质性具有较强鲁棒性的最小二乘回归替代方法，已获得广泛应用。然而，其检查损失函数的非平滑特性，在分布式计算和理论分析中带来了重大挑战。为了克服这些难题，我们提出了一种创新的转换策略，将分位数回归问题转化为最小二乘优化问题。本报告中，我们采用了双平滑技术，对先前牛顿型分布式方法进行了扩展，消除了对误差项与协变量之间独立性的严格假设。我们开发了一种高效的算法，该算法在计算和通信效率方面表现出色。从理论上讲，我们提出的分布式估计器在经过一定数量的迭代后，能够达到接近最优的收敛速度，并实现高准确度的支持恢复。此外，本报告还通过在合成数据和真实数据集上的广泛实验，进一步验证了所提出方法的有效性。实验结果表明，我们的方法在处理高维数据时，不仅能够提供准确的估计，还能有效地恢复数据中的关键支持结构。总体而言，本报告为高维分位数回归的分布式估计和支持恢复提供了一种新的视角和解决方案，具有重要的理论和实际应用价值。

Mixture Conditional Regression with Ultrahigh Dimensional Text Data for Estimating Extralegal Factor Effects

师佳鑫（北京大学）

时间：7.21 15:30-16:10

简介： 师佳鑫，北京大学光华管理学院商务统计与经济计量系在读博士生。主要研究方向为高维数据中的潜在结构分析，因子模型，计算法学，复杂网络数据分析等。研究论文被 Annals of Applied Statistics 期刊接收。

摘要： Testing judicial impartiality is a problem of fundamental importance in empirical legal studies, for which standard regression methods have been popularly used to estimate the extralegal factor effects. However, those methods cannot handle control variables with ultrahigh dimensionality, such as those found in judgment documents recorded in text format. To solve this problem, we develop a novel mixture conditional regression (MCR) approach, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a naïve Bayes type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, we are able to estimate the key parameters of interest as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offenses is analyzed for illustration purposes.

A Gaussian Mixture Model for Multiple Instance Learning with Partially Subsampled Instances

余柏辰（北京大学）

时间：7.21 16:10-16:50

简介： 余柏辰，北京大学光华管理学院商务统计与经济计量系在读博士生，师从王汉生教授。本科毕业于华东师范大学统计学院。主要研究方向为图像数据分析、高维数据分析等。

摘要： Multiple instance learning is a powerful machine learning technique, which is found useful when numerous instances can be naturally grouped into different bags. Accordingly, a bag-level label can be created for each bag according to whether the instances contained in the bag are all negative or

not. Thereafter how to train a statistical model with bag-level labels with/without partially labeled instances becomes the problem of great interest. To this end, we develop a Gaussian mixture model (GMM) framework to describe the stochastic behavior of the instance-level feature vectors. Both the instance-based maximum likelihood estimator (IMLE) and the bag-based maximum likelihood estimator (BMLE) are theoretically investigated. We found that the statistical efficiency of the IMLE could be much better than that of the BMLE, if the instance-level labels are relatively hard to be predicted. To fix the problem, we develop here a subsampling-based maximum likelihood estimation (SMLE) approach, where the instance-level labels are partially provided through carefully subsampling. This leads to a significantly reduced labeling cost with little sacrifice in terms of statistical efficiency. To demonstrate the finite sample performance, extensive simulation studies are presented. A real data example using whole-slide images (WSIs) to diagnose metastatic breast cancer is illustrated.

Gaussian Mixture Model with Rare Events

李雪曈（北京大学）

时间：7.21 16:50-17:30

简介： 李雪曈，北京大学光华管理学院商务统计与经济计量系在读博士生，师从王汉生教授。主要研究方向包括非均衡数据分析，网络结构数据分析，分布式计算等。研究论文发表在 Statistica Sinica, Electronic Journal of Statistics 期刊上。

摘要： We study here a Gaussian Mixture Model (GMM) with rare events data. In this case, the commonly used Expectation-Maximization (EM) algorithm exhibits an extremely slow numerical convergence rate. To theoretically understand this phenomenon, we formulate the numerical convergence problem of the EM algorithm with rare events data as a problem about a contraction operator. Theoretical analysis reveals that the spectral radius of the contraction operator in this case could be arbitrarily close to 1 asymptotically. This theoretical finding explains the empirical slow numerical convergence of the EM algorithm with rare events data. To overcome this challenge, a Mixed EM (MEM) algorithm is developed, which utilizes the information provided by partially labeled data. As compared with the standard EM algorithm, the key feature of the MEM algorithm is that it requires additionally labeled data. We find that the MEM algorithm significantly improves the numerical convergence rate as compared with the standard EM algorithm. The finite sample performance of the proposed method is illustrated by both simulation studies and a real-world dataset of Swedish traffic signs.

Functional Adaptive Double-Sparsity Estimator for High-Dimensional Sensor Data Analysis

李忻月（香港城市大学）

时间：7.21 13:00-13:30

简介： Prof. Li received her PhD in Biostatistics from Yale University. Prior to Yale University, she spent one year at Peking University and three years at the University of Chicago, receiving her B.A. and M.S. in Statistics from the University of Chicago. Prof. Li's research focuses on statistical methods for wearable device data, medical imaging data, large population studies, and precision medicine. Her research papers were published in high-impact journals, such as The Lancet, JAMA Network Open, Advanced Science, IEEE Internet of Things Journal, NPJ Digital Medicine, and Statistica Sinica. Prof. Li has established collaboration with China, Europe, and US to join international efforts in developing statistical methods for analyzing wearable sensor data in large population health studies.

摘要： Wearable sensors have been increasingly used in health monitoring and early anomaly detection. Wearable devices can collect objective and continuous information on physical activity and vital signs and have great potential in studying the association with health outcomes. However, effectively analyzing high-frequency multi-dimensional sensor data is challenging. In this talk, we propose a new Functional Adaptive Double-Sparsity Estimator (FadDoS) based on functional regularization of sparse group lasso with multiple functional predictors, which can achieve global sparsity via functional variable selection and local sparsity via zero-subinterval identification within coefficient functions. We prove that the FadDoS estimator converges at a bounded rate and satisfies the oracle property under mild conditions. Extensive simulation studies confirm the theoretical properties and exhibit excellent performances compared to existing approaches. We applied FadDoS to a Kinect sensor study that utilized an advanced motion sensing device tracking human multiple joint movements and conducted among community-dwelling elderly, and we demonstrated how FadDoS can effectively characterize the detailed association between joint movements and physical health assessments. The proposed method is not only effective in Kinect sensor analysis but also applicable to broader fields where multi-dimensional sensor signals are collected simultaneously. The R code for FadDoS is available at https://github.com/Cheng-0621/FadDoS.

Bayesian Integrative Region Segmentation in Spatially Resolved Transcriptomic Studies

罗翔宇（中国人民大学）

时间：7.21 13:30-14:00

简介：罗翔宇 2018 年 9 月起任职于中国人民大学统计与大数据研究院，现为准聘副教授。他 2018 年博士毕业于香港中文大学统计系。罗翔宇的研究兴趣包括贝叶斯统计、非参数贝叶斯、生物信息学、统计计算等。他热衷于开发新的统计模型来解决实际中的生物问题。其具体研究方向包括利用统计图模型构建基因调控或共表达网络、纠正高通量数据中的批次效应、对于批量层次的基因表达或 DNA 甲基化数据进行去卷积化、发现单细胞分辨率上的个体异质性、空间转录组及多组学数据融合分析等。

摘要：The spatially resolved transcriptomic study is a recently developed biological experiment that can measure gene expressions and retain spatial information simultaneously, opening a new avenue to characterize fine-grained tissue structures. In this article, we propose a nonparametric Bayesian method named BINRES to carry out the region segmentation for a tissue section by integrating all the three types of data generated during the study—gene expressions, spatial coordinates, and the histology image. BINRES is able to capture more subtle regions than existing statistical partitioning models that only partially make use of the three data modes and is more interpretable than neural-network-based region segmentation approaches. Specifically, due to a nonparametric spatial prior, BINRES does not require a prespecified region number and can learn it automatically. BINRES also combines the image and the gene expressions in the Bayesian consensus clustering framework and thus flexibly adjusts their label alignment contribution weights in a data-adaptive manner. A computationally scalable extension is developed for large-scale studies. Both simulation studies and the real application to three mouse spatial transcriptomic datasets demonstrate that BINRES outperforms the competing methods and easily achieves the uncertainty quantification of the integrative partition.

Enhancing Treatment Strategies and Risk Assessment in Hip Fracture Elderly Patients: A Copula-Based Approach for Semi-Competing Risks Analysis

孙韬（中国人民大学）

时间：7.21 14:00-14:30

简介：孙韬，中国人民大学统计学院副教授，博士毕业于匹兹堡大学生物统计系，主要研究方向为复杂生存数据模型，老年失能风险管理。

摘要：Hip fracture is a severe complication in the elderly. The affected people are at a higher risk of second fracture and death occurrence, and the best treatment for hip fractures is still being

debated. Aside from the treatment, many factors, such as comorbidity conditions, may be associated with second fracture and death occurrence. This study aims to identify effective treatments and important covariates and estimate their effects on the progression of second fracture and death occurrence in hip fracture elderly patients using the semi-competing risks framework, because death dependently censors a second fracture but not vice versa. Due to the complex semi-competing risks data, performing variable selection simultaneously for second fracture and death occurrence is difficult. We propose a penalised semi-parametric copula method for semi-competing risks data. Specifically, we use separate Cox semi-parametric models for both margins and employ a copula to model the two margins' dependence. We apply the proposed method to a population-based cohort study of hip fracture elderly patients, providing new insights into their treatment and clinical management.

Network and Covariate Adjusted Response-Adaptive Design

梅好（中国人民大学）

时间：7.21 14:30-15:00

简介：梅好，中国人民大学统计学院讲师，中国人民大学杰出青年学者，2021 年博士毕业于耶鲁大学，曾就职于耶鲁纽黑文医院临床实效研究中心，腾讯医疗健康事业部。主要研究方向为网络数据分析、生存分析、复杂数据建模等统计学方法及其在在医疗健康、决策预测等领域的应用。在 Biometrics, Statistics in Medicine, Annals of Emergency Medicine, BMC health services research 等期刊发表论文十余篇，总引用量 300 次以上。

摘要：Randomization is a distinguishing feature of clinical trials for unbiased assessment of treatment efficacy. With a growing demand for more flexible and efficient randomization schemes and motivated by the idea of adaptive design, in this article we propose the network and covariate adjusted response-adaptive (NCARA) design that can concurrently manage three challenges: 1) maximizing benefits of a trial by assigning more patients to the superior treatment group randomly; 2) balancing social network ties across treatment arms to eliminate potential network interference; and 3) ensuring balance of important covariates, such as age, gender, and other potential confounders. We conduct simulation with different network structures and a variety of parameter settings. It is observed that the NCARA design outperforms four alternative randomization designs in solving the above-mentioned problems and has comparable power and type I error for detecting true difference between treatment groups. In addition, we conduct real data analysis to implement the new design in two clinical trials. Compared to equal randomization (the original design utilized in the trials), the NCARA design slightly increases power, largely increases the percentage of patients assigned to the better-performing group, and significantly improves network and covariate balances. It is also noted that the advantages of the NCARA design are augmented when the sample size is small and the

level of network interference is high. In summary, the proposed NCARA design assists researchers in conducting clinical trials with high-quality and high-efficiency.

What happens when your validated ecosystem is a Graph?

段晓丽（罗氏制药）

时间：7.22 19:00-19:30

简介： Xiaoli Duan has been a Data Scientist in Roche PD Data Sciences since she received her Ph.D. degree in Industrial Engineering in 2022, with a research focus on statistical machine learning in healthcare. She is an R developer of the NEST project (chevron family) and a Python developer of automatic tumor segmentation algorithms. She is a product owner of the R interface to Roche's distributed ecosystem across multiple semantic platforms.

摘要： A validated environment to use R to develop clinical trials reporting tools and deliver reproducible data analytic results (i.e. table, listing, and figure outputs) for regulatory submission is a must. The Comprehensive R Archive Network (CRAN) which sets up the highest standard of validating a new/upgraded package assesses the cohort of package reverse dependencies upon submission and evaluates if the package continues to serve as expected as a dependency in the current validated ecosystem. Indeed, the evaluation of the heaviness of package dependencies and the risk of inter-dependency impacting reproducibility is a complex process, given that active package up-versioning and data standards publications make our Auto-validation R Submission Portal a dynamic system/network on a daily basis.

Our goal is to effortlessly touch the comprehensive review of a validated ecosystem's all available package dependencies via a directed Graph - a non-linear data structure in graph theory - and simplify the validation task workflow in terms of computational complexity. We will

(1) linearly traverse/search and visualize package dependencies within a user-defined scope,

(2) linearly order/schedule pending packages to be validated in the queue and automatically trigger which is the next to be performed in the validation pipelines to minimize any newly broken package behaviors due to package upgrades, and

(3) automatically make package owners/maintainers notified if their package dependencies get upgraded up to certain versions by any other package requests, which will make the package re-submitted for validation again (but thinking about this is a heads-up of potential test failures due to package up-versioning).

Our demos will cover three CRAN-released clinical trial analysis tools: tern (Roche), tidytlg (J&J), and forestly (Merck). Note that our proposed framework can be generalized for any complex dataflow system, regardless of performing tasks, programming languages, package managers, etc. The dynamic QC (for results) process (and data dependencies) can also be supported if we provide an end-to-end R solution to clinical reporting in a centralized platform.

Integrating LLM Coding Capabilities in End-to-End Data Science: Challenges and Reflections

程鼎（艾伯维）

时间：7.22 19:30-20:00

简介： Ding Cheng is currently working at AbbVie - Allergan Aesthetics, where he is responsible for commercial and business-related data analysis and modeling. With extensive experience in clinical research development, IT and innovation, and business intelligence, Ding is passionate about integrating advanced digital technologies with medical practices to drive improvements in the healthcare industry.

摘要： This presentation will explore the integration of large language model (LLM) coding capabilities within the end-to-end data science workflow. Using a case study of constructing a Chat Dashboard, we will delve into the challenges, insights, and reflections encountered throughout the process. The focus will be on development within the R programming environment, highlighting the application of statistical models to enhance data analysis and decision-making. The presentation will cover technical implementation details and share experiences in project management and interdisciplinary collaboration, providing practical guidance for professionals looking to leverage LLM advantages in the data science field.

Patient Narrative Generation in R

曹心怡（先声再明医药）

时间：7.22 20:00-20:30

简介： Zoe Cao
Statistical Programmer Employed at Simcere Zaiming Pharmaceutical Company
Graduated from the University of British Columbia with majors in Statistics and Economics
曹心怡
统计程序员，就职于先声再明医药有限公司
毕业于英属哥伦比亚大学，主修统计学与经济学

摘要： The Patient Narrative, or Adverse Event narrative, is critical in clinical trials for providing detailed safety data. Its distinctive features include patient-generated content and presentation in chronological order. However its creation involves tedious tasks like data retrieval and event timeline linking. The use of R for the automated generation of patient narrative reports significantly saves the resources in data collection and repetitive writing tasks, offering a notable improvement in accuracy compared to manual methods. This presentation will primarily focus on how to generate Narratives

using R, along with the usage of the current popular R packages. Moreover it will explore the potential for further automating Narrative generation in R.

双剑合璧: R 和 Python 协同构建数据应用

王杰，刘晓畅（强生）

时间：7.22 20:30-21:00

简介： 王杰，是强生创新制药中国研发临床统计编程部门技术解决方案的数据工程师。他是一位技术娴熟的统计程序员，专注于发现机会，推动优化和创新，应用传统和前沿的方法进行临床相关的数据分析。他拥有 12 年生物制药数据分析经验，在加入强生之前，曾在辉瑞研发工作过 5 年多从事临床数据分析相关工作。刘晓畅目前任职于 Johnson & Johnson 的临床与统计编程部门，担任的是 Data Engineer 的工作。他擅长使用 R，Python 以及其他编程语言和工具，在处理大规模临床数据、数据挖掘、机器学习，数据可视化和生成式人工智能应用方面具有丰富的经验和技能。他的工作目标是利用数据驱动的解决方案来支持临床研究和决策。他于 2018 年获得山东大学药学学士学位，2019 年获得英国爱丁堡大学药物发现与转化生物学硕士学位。

摘要： R 与 Python 是构建数据科学应用的过程中必不可少的重要工具，诚然，它们有着各自独特的优势：R 以其强大的统计分析能力和数据可视化功能闻名，而 Python 则以其易读性和广泛的库的支持在数据处理和机器学习领域中占据一席之地。对于一个完整的数据科学项目，R 与 Python 并非互斥的关系。我们可以通过结合它们各自的优势，在开发过程中实现协同效应。从需求出发，灵活选择工具，这样我们将极大地提高开发数据应用的速度，以及赋予应用一定程度的鲁棒性。我们将详细介绍这种协同工作的实践过程，以及如何最大限度地利用 R 和 Python 的优势，为数据科学家提供一个全新的构建数据应用的视角。对于大部分的数据清洗、可视化以及前端交互的需求，我们选择使用 R 作为我们的工具。其中，我们选择了 R Shiny 作为前端交互的框架并使用了 R golem 作为开发 R Shiny 应用的框架。对于一些特定的功能，我们利用 Python 的优势，通过 FastAPI 搭建接口为前端应用提供功能的实现。此外，我们也借助了微软提供的 Graph API，以此来丰富应用的功能。在具体的实践过程中，需要根据项目的需求和团队的技术能力来选择合适的工具和框架。通过合理地利用 R 和 Python 的协同构建，可以开发出高效、灵活和功能强大的数据应用，为数据科学工作提供更多可能性和创新空间。

Simultaneous jump detection for multiple sequences via screening and multiple testing

张春明（威斯康星大学）

时间：7.22 19:00-19:30

简介： Chunming Zhang is a Professor in the Department of Statistics at the University of Wisconsin-Madison. She earned her Ph.D. in Statistics from the University of North Carolina at Chapel Hill under the guidance of Jianqing Fan. She completed her B.S. in mathematical statistics at Nankai University, Tianjin, China, and an M.S. in Computational Mathematics from Academia Sinica, Beijing, China. Her research interests range from statistical learning and data mining, statistical methods with applications to imaging data, neuroinformatics, and bioinformatics, multiple testing, large-scale simultaneous inference and applications, statistical methods in financial econometrics, non- and semi-parametric estimation and inference, to functional and longitudinal data analysis. Her current research topics include new developments in the area of large-scale structure learning tasks and statistical inference procedures, with applications in neuroscience, biology, machine learning, and causal inference. She is an elected Fellow (2016) of the American Statistical Association (ASA) and an elected Fellow (2011) of the Institute of Mathematical Statistics (IMS) and is honored by a Medallion Award and Lecturer (2024) of the IMS.

摘要： The estimation of nonparametric discontinuous regression function is fundamental in many applied fields, but challenges arise when the number of jumps (or discontinuities) is large and unknown. We propose a new jump detection method, via the consecutive screening and multiple testing (SaMT) algorithm for estimating the unknown jump points in the flexible non-parametric regression model, guaranteeing the desired accuracy. The initial jump candidates are obtained in the consecutive screening procedure combined with locally-linear smoothing method. To further assess the significance of an individual jump candidate, we develop a novel test based on the profile likelihood inference. The ultimate selection of relevant jump points is conducted in multiple testing procedure, which rules out irrelevant jump points with large variations, due to heteroscedastic errors, from jump candidates. Moreover, we generalize the proposed SaMT algorithm to detect the common jump points shared across multiple aligned sequences. The proposed method is easy to implement, enjoys flexibility in choices of bandwidth parameter and threshold quantity in screening, and is illustrated through simulations and real data examples, as compared with existing methods.

Common Odds Ratio Test and Interval Estimation for Stratified Bilateral and Unilateral Data

马长兴（*University at Buffalo*）
时间：7.22 19:30-20:00

简介： Changxing Ma, PhD is Associate Professor, Co-Director for Master of Public Health (MPH) Biostatistics in the Department of Biostatistics at the University at Buffalo. He graduated from Nankai University in 1997. Before joining in Biostatics University at Buffalo, he worked at Nankai Dept of Statistics from 1992 to 2002, worked with longitudinal and birth cohort's data in University of Florida for 5 years from 2000 to 2005. He published more than 130 peer-reviewed publications in a wide range of statistical and biomedical journals. His Google scholar h-index is 46, i10-index 95.

摘要： In clinical research, data are commonly collected bilaterally from paired organs or bodily parts within individual subjects. However, unilateral data arise when constraints or limiting factors impede the collection of complete bilateral data. In this paper, we propose three large-sample tests and five confidence interval methods for making inferences on the common treatment effect, measured by the odds ratio, in a stratified design under integrated bilateral and unilateral data. Our simulation results show that the likelihood ratio-based and score-based tests, along with their associated confidence interval methods, demonstrate robust control of type I error and close-to-nominal coverage probabilities. We apply the proposed methods to real-world datasets of acute otitis media and myopic eyes to showcase their validity and applicability in clinical practice.

Assessing heterogeneous causal effects across clusters in partially nested designs

刘笑 *(The University of Texas at Austin)*
时间：7.22 20:00-20:30

简介： Xiao Liu is an assistant professor in the quantitative methods program of the Department of Educational Psychology at UT Austin. She is interested in causal inference methods, quasi-experimental methods (e.g., propensity score), causal mediation analysis, and longitudinal data analysis.

摘要： Partially nested designs are common in studies of psychological or behavioral interventions. In this type of design, after participants are assigned to study arms, participants in a treatment arm are subsequently assigned to clusters (e.g., teachers, therapy groups) to receive treatment, whereas participants in a control arm are unclustered (e.g., a wait-list control). As participants in the treatment arm receive treatment in clusters, it is often of interest to examine heterogeneity of treatment effects

across the clusters; but this is challenging in the partially nested design. Particularly, in defining a causal effect of treatment for a specific cluster (e.g., a specific therapist), it is unclear how the treatment and control outcomes should be compared, as the control arm has no clustering (e.g., no therapists). It may be tempting to compare outcomes of a specific cluster to outcomes of the entire control arm —however, this comparison may not represent a causal effect even when the treatment assignment is randomized, because the cluster assignment in the treatment arm may be nonrandomized (elaborated in this talk). In this talk, I will describe our study that extends the principal stratification framework and the principal score approach to assessing heterogeneous cluster-specific treatment effects in the partially nested design. Besides the effect definition and identification, our study obtains various estimators for the cluster-specific treatment effects, including a multiply-robust estimator that can provide more robustness to parametric model misspecification. In addition to simulation results, I will present an empirical example applying our methods to estimating the heterogeneous treatment effects across clusters in a partially nested design. I will end this talk with a discussion of the implications of our study and potential future directions.

Construction of strong orthogonal Latin hypercubes

王春燕（中国人民大学）

时间：7.22 20:30-21:00

简介：王春燕，中国人民大学统计学院讲师，南开大学博士，田纳西大学访问学者，普渡大学博士后助理研究员。研究方向包括统计试验设计、计算机试验、次序添加试验等。相关文章发表在《中国科学数学》、《Annals of Statistics》、《Statistica Sinica》等期刊上。

摘要：Column-orthogonality and space-filling property are perhaps two most desirable design properties for computer experiments. Column-orthogonality allows the estimates of the main effects in linear models to be uncorrelated with each other, while the space-filling property is appropriate for Gaussian process models. Orthogonal Latin hypercubes are widely used for computer experiments. They achieve both orthogonality and the maximum one-dimensional stratification property. When two-factor (and higher-order) interactions are active, two- and three-dimensional stratifications are also important. Unfortunately, little is known about orthogonal Latin hypercubes with good two (and higher) dimensional stratification properties. This paper proposes a method for constructing a new class of orthogonal Latin hypercubes whose columns can be partitioned into groups, such that the columns from different groups maintain two- and three-dimensional stratification properties. The proposed designs perform well under almost all popular criteria (e.g., the orthogonality, stratification, and maximin distance criterion). They are the ideal designs for computer experiments. The construction method can be straightforward to implement, and the relevant theoretical supports are well established. The proposed strong orthogonal Latin hypercubes are tabulated for practical needs.

R X-AGI IFoDS

第 17 届中国 R 会议
The 17th China-R Conference

2024 X 智能大会
The 2024 X-AGI Conference

2024 数据科学国际论坛
The 2024 International Forum on Data Science

主办方 中国人民大学应用统计科学研究中心 Center for Applied Statistics of Renmin University of China | 中国人民大学 RENMIN UNIVERSITY OF CHINA 统计学院 SCHOOL OF STATISTICS | CAPITAL OF STATISTICS PROFESSION, HUMANITY & INTEGRITY | 中国商业统计学会人工智能分会 COMMERCE STATISTICAL SOCIETY OF CHINA SECTION ON AI

承办方 JDS Journal of Data Science | 中国人民大学数据科学与大数据统计系

赞助方 明汯投资 MINGHONG INVESTMENT | 宽德投资 WIZARD QUANT | Heywhale 和鲸 | 子博设计