大数据时代
统计学的跨界融合与发展论坛

2019.5

主办单位：
国家自然科学基金委数学天元基金委员会
国家天元数学东北中心

承办单位：
东北师范大学数学与统计学院
应用统计教育部重点实验室
东北师范大学大数据研究院

中国·长春

会议指南

1. 会议时间及安排

5 月 18 日 8:00-12:00，会议报到及注册，东北师范大学数学与统计学院 104

5 月 18 日 -19 日，行业专家报告，东师数学与统计学院 104 报告厅

5 月 21 日，学术交流报告，东师数学与统计学院 104 报告厅

5 月 24 日，学术交流报告，东师就业指导中心三楼报告厅

2. 入住酒店

名称：长春海航名门饭店

地址：长春市朝阳区人民大街 4501 号

3. 会务组成员

朱文圣　齐春香　吴双

4. 联系方式

联 系 人：齐春香

邮件地址：qicx1224@nenu.edu.cn

办公电话：0431-85099763

手机号码：15044138879

通讯地址：吉林省长春市人民大街 5268 号东北师范大学

数学与统计学院 206 办公室

日程安排

时间：2019 年 5 月 18 日（星期六） 地点：东北师范大学数学与统计学院 104 报告厅			
时间	行业报告		主持人
13:50–14:00	郭建华 东北师范大学副校长　致辞		
14:00–14:30	于丹 （中国科学院数学与系统科学研究院）	统计分析与工业应用	高巍 （东北师范大学）
15:00–15:30	刘宏亮 （吉林省社会医疗保险管理局）	医保概述及问题分析	
16:00–16:15	茶歇		
16:15–16:45	尹俊平 （北京应用物理与计算数学研究所）	医疗数据中的典型数据科学问题	朱文圣 （东北师范大学）
17:15–17:45	刘秉辉 （东北师范大学）	东师大数据研究院产业化应用案例	
18:30	晚宴（名门饭店三楼贵宾厅）		

日程安排

时间：2019 年 5 月 19 日（星期日）
地点：东北师范大学数学与统计学院 104 报告厅

时间	行业报告		主持人
8:30–9:00	蔡　波 （长春市政府市长公开电话办公室）	应让数据说真话	郭建华 （东北师范大学）
9:30–10:00	王建立 （中国科学院长春光学精密机械与物理研究所）	智能光学成像系统及其应用	
10:30–10:45	茶歇		
10:45–11:15	陈博 （中国联通大数据有限公司）	运营商大数据助力教育行业共发展	刘秉辉 （东北师范大学）
12:00	午餐（林业宾馆 103 厅）		

时间：2019 年 5 月 21 日（星期二） 地点：东北师范大学数学与统计学院 415 会议室			
时间	学术交流		主持人
08:00–09:00	范剑青 （普林斯顿大学）	Statistics and AI	郭建华 （东北师范大学）
09:00–10:00	李润泽 （宾夕法尼亚州立大学）	Test of Significance for High-Dimensional Longitudinal Data	
10:00–10:15	茶歇		
10:15–10:45	陈敏 （中国科学院）	Semi-parametric inference for large-scale data with non-stationary non-Gaussian temporally dependent noises	高巍 （东北师范大学）
10:45–11:15	陈钊 （复旦大学）	Ultrahigh Dimensional Precision Matrix Estimation via Refitted Cross Validation	
11:15–11:45	黎德元 （复旦大学）	Extreme Quantile Estimation for Single Index Model	
11:45–12:15	冯兴东 （上海财经大学）	Lack-of-fit tests for quantile regression models	
12:15	午餐（林业宾馆）		

时间：2019 年 5 月 24 日（星期五） 地点：东北师范大学数就业指导中心三楼报告厅			
时间	学术交流		主持人
2:30–3:30	徐宗本 （西安交通大学）	TBD	徐海阳 （东北师范大学）

行业专家报告摘要

统计分析与工业应用

于丹（中国科学院数学与系统科学研究院）

个人简介：

中国科学院数学与系统科学研究院研究员。1984 年毕业于中国科学技术大学数学系（本科），1991 年毕业于北京大学概率统计系（硕士），1996 年毕业于中国科学院系统科学研究所（博士）。长期从事工业统计研究与应用。现担任中科院数学院“质量与数据科学研究中心”主任、“航天产品可靠性技术与质量科学联合实验室”主任、“纳米技术与统计科学联合实验室”副主任。曾获得国务院政府特殊津贴、全国优秀科技工作者、国防科学技术二等奖及国防科学技术进步二等奖各一项。

摘要：

本报告旨在与大家交流统计分析在工业领域应用研究的经验与体会，以及科研、教学与工业应用之间相互促进的案例。

医保概述及问题分析

刘宏亮（吉林省社会医疗保险管理局）

个人简介：

吉林省社会医疗保险管理局，副局长，东北师范大学数学与统计学院研究生兼职指导教师，吉林省社会医疗保险研究会常务理事。长期从事医疗生育保险经办管理工作，设计并推动吉林省医疗保险支付制度、异地就医直接结算管理、病种付费管理、特殊药品管理等多项医疗保险制度的建立及实施，积累了丰富的医疗生育保险经办管理实践经验及实际研究成果。

摘要：

本报告结合吉林省医保运行实际，简要介绍了医疗保障的组成、运行模式和管理机

制、医保数据分析面临的困难以及当前医保运行中重点问题与解决思路。内容主要包括当前医保的组织架构和运行机制、医保行业涉及的数据资源情况、重点问题的原因分析，提出了基于用户视角的问题解决构想。

医疗数据中的典型数据科学问题

尹俊平（北京应用物理与计算数学研究所）

个人简介：

北京应用物理与计算数学研究所，研究员，博士生导师、博士后合作导师。现任北京应用物理与计算数学研究所信号与数据处理技术研究联合实验室主任、大数据团队首席科学家。研究方向：数据科学。长期从事数据科学与人工智能在情报处理中的应用研究。

摘要：

本报告结合国内某知名医院临床数据，简要介绍了该知名医院几个科室的临床医疗数据科学中的典型实际问题，内容主要包括数据预处理（编码、缺失数据、数据平衡等等）、变量选择、分类、聚类、预测等数据分析的典型环节，还包括函数型数据、纵向数据、以及 RNN 的一些问题。

行业专家报告摘要

应让数据说真话

蔡波（长春市政府市长公开电话办公室）

个人简介：

长春市政府市长公开电话办公室处长，20 年一直从事市长公开电话电话工作，曾撰写的论文《政府热线整合》在 2008 年编入欧盟社会发展论坛一书中。

摘要：

数据挖掘离不开数据的真实，只有真实的数据才能挖掘正确的结论；只有使用全面的数据，才能保持结论导向的不偏、不变、不移。

智能光学成像系统及其应用

王建立（中国科学院长春光学精密机械与物理研究所）

个人简介：

工学博士，研究员，博士生导师，中国科学院长春光学精密机械与物理研究副所长。目前主要从事地基空间目标光电探测技术、地基大口径望远镜总体技术研究等工作。作为项目第一负责人先后完成国家重大工程项目 10 项，发表论文 120 余篇，获军队科技进步一等奖两项、二等奖一项。曾获第四届吉林省十大杰出青年，第十三届全国青年五四奖章等荣誉，科技部中青年创新领军人才，国家“万人计划”科技创新领军人才。

运营商大数据助力教育行业共发展

陈博（联通大数据有限公司）

个人简介：

陈博，博士 / 博士后，高级工程师，现任联通大数据有限公司数据科学总监，主要负责大数据分析挖掘、机器学习建模、人工智能方向的研发工作。在此之前，陈博

于 2008 年毕业于北京邮电大学，取得信号与信息处理专业博士学位，主要研究方向为机器学习、自然语言处理、信息检索；毕业后曾先后就职于 NEC 中国研究院、中国联通集团博士后科研工作站、中国联通集团总部技术部。

陈博在长期科研工作中，兼具数据挖掘、机器学习专业背景，以及电信运营商、移动互联网行业背景，具有复合型技术规划与研发能力。目前，作为主要作者已发表 SCI、EI 等论文 20 余篇、申请发明专利 6 项、出版著作 1 本，并作为核心人员主导、参与多个国家重大专项、核高基专项、运营商内部项目的研发实施工作。

摘要：

本报告首先将介绍联通大数据在海量运营商数据资源上，如何采用数据科学方法，将统计分析、机器学习、深度学习、图计算等算法技术，应用于构建大数据价值体系与大数据能力体系，以及在此过程中所形成的典型应用案例。之后，将介绍联通结合教育行业的关注点，利用大数据所作的应用探索，以及研发的数据科学教育云平台。

东师大数据研究院产业化应用案例

刘秉辉（东北师范大学）

个人简介：

东北师范大学副教授、统计系主任、统计学一流学科工作委员会委员兼秘书长；主要研究方向为应用统计、机器学习、网络数据分析；在 Artificial Intelligence、Journal of Machine Learning Research、The Annals of Applied Statistics 等国际知名期刊发表多篇学术论文；主持国家自然科学基金青年项目一项、面上项目一项、中央高校基本科研业务费青年拔尖人才项目一项；与中国联通吉林省公司合作，主持“互联网化运营咨询与培训”人力资源项目一项、“校园关怀”大数据开发项目一项。

摘要：

本报告将介绍东师大数据研究院在产业化应用方面的两个案例——竞标长春市社会治理非应急信息服务平台采购项目、联手中国联通吉林省分公司打造“校园关怀”主题系列。

学术报告摘要

Statistics and AI

范剑青（普林斯顿大学）

This talk first gives an overview on the genesis of machine learning and AI and how statistical and computational methods have evolved with growing dimensionality and sample sizes and become the foundation of modern machine learning and AI. It will also outline how ideas of trading modeling biases and variances have been developed into high-dimensional statistics and machine learning, with focus on deep learning models. We will outline the challenges of statistical sciences at this crossroad and offer some prospects. We will offer a general robustification principle and show how to use factor adjustments to deal with dependent measurements. In particular, Factor Adjusted Robust Multiple testing (FarmTest) and Model selection (FarmSelect) will be introduced for high-dimensional statistical inference and model selection. The effectiveness of these methods will be revealed with an application to predicting bond risk premia using macroeconomic time series. Further insights on the prospects of machine learning and AI will be offered.

Test of Significance for High–Dimensional Longitudinal Data

李润泽（宾夕法尼亚州立大学）

This paper concerns statistical inference for longitudinal data with ultrahigh dimensional covariates. We first study the problem of constructing confidence intervals and hypothesis tests for a low dimensional parameter of interest. The major challenge is how to construct a powerful test statistic in the presence of high-dimensional nuisance parameters and sophisticated within-subject

correlation of longitudinal data. To deal with the challenge, we propose a new quadratic decorrelated inference function approach, which simultaneously removes the impact of nuisance parameters and incorporates the correlation to enhance the efficiency of the estimation procedure. When the parameter of interest is of fixed dimension, we prove that the proposed estimator is asymptotically normal and attains the semiparametric information bound, based on which we can construct an optimal Wald test statistic. We further extend this result and establish the limiting distribution of the estimator under the setting with the dimension of the parameter of interest growing with the sample size at a polynomial rate. Finally, we study how to control the false discovery rate (FDR) when a vector of high-dimensional regression parameters is of interest. We prove that applying the Storey (2002)'s procedure to the proposed test statistics for each regression parameter controls FDR asymptotically in longitudinal data. We conduct simulation studies to assess the finite sample performance of the proposed procedures. Our simulation results imply that the newly proposed procedure can control both Type I error for testing a low dimensional parameter of interest and the FDR in the multiple testing problem. We also apply the proposed procedure to a real data example.

Ultrahigh Dimensional Precision Matrix Estimation via Refitted Cross Validation

陈钊（南开大学）

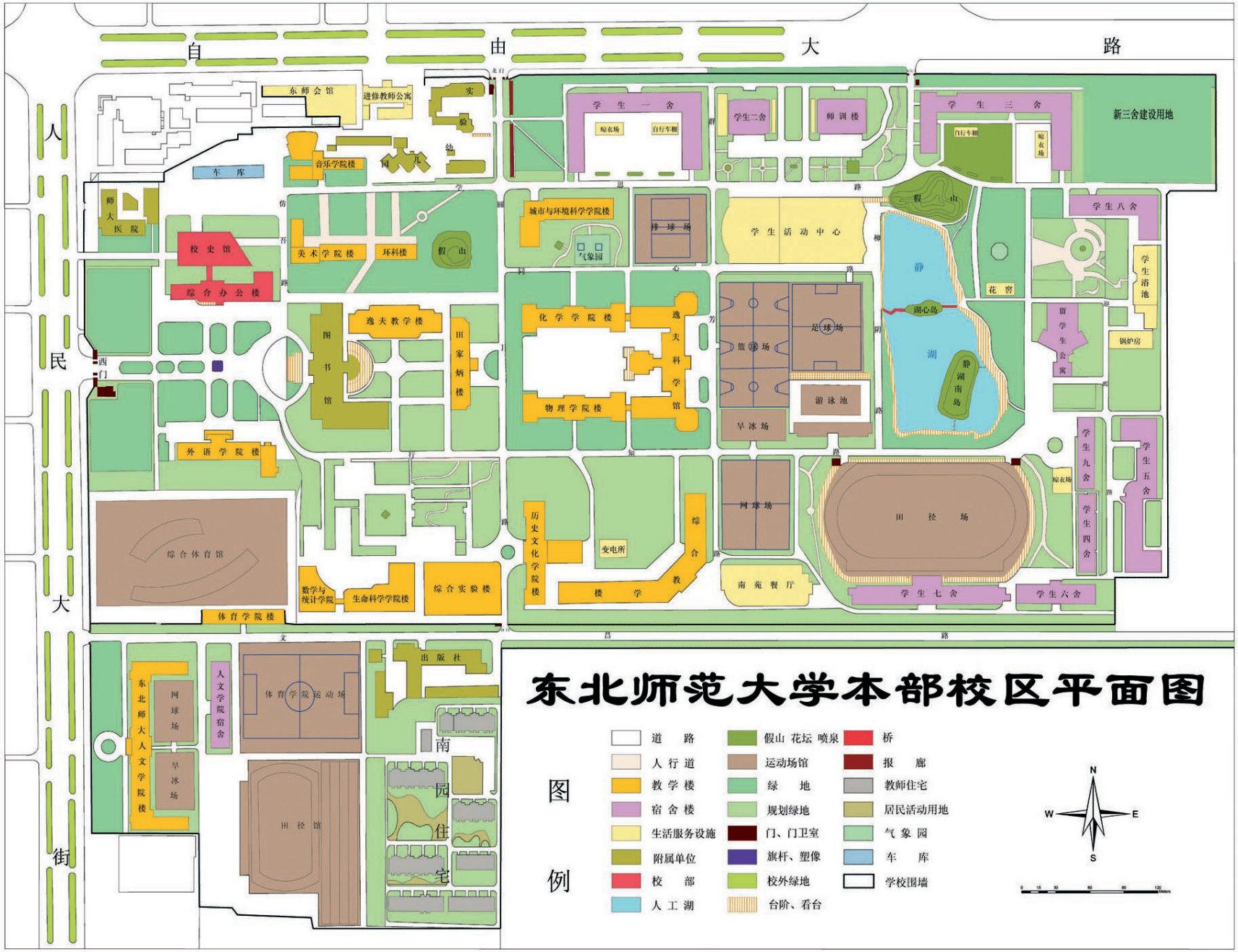
This paper develops a new estimation procedure for ultrahigh dimensional sparse precision matrix, the inverse of covariance matrix. Regularization methods have been proposed for sparse precision matrix estimation, but they may not perform well with ultrahigh dimensional data due to spurious correlation. We propose a refitted cross validation (RCV) method for sparse precision matrix estimation based on its Cholesky decomposition. The proposed RCV procedure can be easily implemented with existing software for ultrahigh dimensional linear regression.

We establish the consistency of the proposed RCV estimate and show that the rate of convergence of the RCV estimate without assuming banded structure is the same as those assuming the banded structure in Bickel and Levina (2008b). Monte Carlo studies were conducted to access the finite sample performance of the RCV estimate. Our numerical comparison shows that the RCV estimate can outperform existing ones in various sce.

Extreme Quantile Estimation for Single Index Model

黎德元（复旦大学）

Single Index model is a flexible semiparametric regression model and it reduces the dimension of the covariates. In this paper, we consider the estimation of the extreme conditional quantiles for the single index model and developed a so-called three-stepestimator. We first obtain a misspecified root-n estimator of the index parameter vector under the linear quantile regression. Secondly, we apply a local polynomial regression technique to estimate intermediate conditional quantiles. Finally, we extrapolate these estimates to tails by extreme value theory. We show the asymptotic properties of the provided estimator and study its performance for finite sample by simulation. A real application to the NMMAPS dataset of LA is also provided.


自由大路
人民大街
东师会馆
进修教师公寓
实验幼儿园
学生一舍
晾衣场
自行车棚
学生二舍
师训楼
学生三舍
新三舍建设用地
车库
音乐学院楼
师大医院
校史馆
综合办公楼
美术学院楼
环科楼
假山
城市与环境科学学院楼
气象园
排球场
学生活动中心
静湖心岛
静湖南岛
学生八舍
学生浴池
花窖
留学生公寓
锅炉房
图书馆
逸夫教学楼
田家炳楼
化学学院楼
物理学院楼
逸夫科学馆
篮球场
足球场
游泳池
旱冰场
外语学院楼
学生九舍
学生五舍
学生四舍
综合体育馆
历史文化学院楼
变电所
综合教学楼
网球场
田径场
南苑餐厅
学生七舍
学生六舍
数学与统计学院
生命科学学院楼
综合实验楼
体育学院楼
出版社
东北师大人文学院楼
网球场
人文学院宿舍
体育学院运动场
旱冰场
田径馆
南苑住宅
东北师范大学本部校区平面图
图例
道路
人行道
教学楼
宿舍楼
生活服务设施
附属单位
校部
人工湖
假山 花坛 喷泉
运动场馆
绿地
规划绿地
门、门卫室
旗杆、塑像
校外绿地
台阶、看台
桥
报廊
教师住宅
居民活动用地
气象园
车库
学校围墙
N
W
E
S

Record Pge