

Workshop on Integration of Transportation, Statistics and Big Data

July 21.- 22. 2017



**TECHNISCHE
UNIVERSITÄT
DRESDEN**



Yale University



ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



**PRINCETON
UNIVERSITY**



**Stanford
University**

SWUFE

Prof. Dr. Daimin Shi, Vice president of SWUFE in research and education

Prof. Dr. Jianjun Guo, the Dean of the School of Statistics

Prof. Dr. Hui Xiao, Research field: Management

Prof. Dr. Jinyuan Chang, Research field: Statistics, Econometrics

Prof. Dr. Feng Xiao, Research field: Transportation Economics

Prof. Dr. Shulin Zhang, Research field: Statistics, Econometrics

UMICH

Prof. Dr. Peter Song, Research field: Biostatistics, School of Public Health

YALE

Prof. Dr. Heping Zhang, Professor of Biostatistics, Statistics and Data Science, School of Public Health

UIUC

Prof. Dr. Annie Qu, Professor of Statistics, Department of Statistics

Princeton University

Jeffrey Zhang, Center for Statistics and Machine Learning

Stanford University

Leon Zhang, Institute for Economic Policy Research

TUD

Prof. Dr. Hans Georg Krauthäuser, Vice president in education and international affairs

Prof. Dr. Bernard Bäker, Dean of the Faculty of Transportation

Prof. Dr. Jörn Schönberger, Vice-dean of the Faculty of Transportation

Prof. Dr. Hartmut Fricke, Director of the Institute of Air Transportation and Logistics

Prof. Dr. Ulrike Stopka, Study dean at the Faculty of Transportation, Chair of Communications

Prof. Dr. Bernhard Schipp, Commissioner of foreign affairs of the School of Business and Economics,
Chair of Econometrics, School of Business and Economics

Prof. Dr. Ostap Okhrin, Commissioner of foreign affairs of the Faculty of Transportation, Chair of
Statistics and Econometrics, Faculty of Transportation

Danny Klotzsche, Consultant for Internationalization of the Field: Construction and Environment

Day	Time	Speaker	Title
Friday 21.07.17	12:00 - 13:20	Excursion to the Railway Laboratory Registration	
	Discussion on MOA Participants: SWUFE - Guo, Xiao, Chang, Xiao, Zhang TUD - Okhrin, Bäker, Fricke, Schipp, Schönberger, Stopka, Klotzsche		
	13:30 - 13:35	Prof. Dr. Ostap Okhrin	Welcome words
	13:35 - 13:45	Prof. Dr. Bernard Bäker	Opening Speech
	13:45- 13:55	Prof. Dr. Jianjun Guo	Speech
	13:55 - 14:05	Danny Klotzsche	Academic Exchange and international research
	14:05 - 14:45	Discussion Cooperation/ Masterprogram Joint Signature of the Memorandum of Agreement	
	14:45 - 15:00	Coffee Break	
	Scientific Part		
	15:00 - 15:25	Prof. Dr. Bernard Bäker	The flying car comes later - Mechatronical vehicles of the future 2030+
	15:25 - 15:50	Prof. Dr. Hui Xiao	Ranking and Selection: An Optimal Computing Budget Allocation Procedure
	15:50 - 16:15	Prof. Dr. Hartmut Fricke	Optimisation of ecological and safety aspects of the trajectory modeling
	16:15 - 16:25	Coffee Break	
	16:25 - 16:50	Prof. Dr. Jinyuan Chang	A new scope of penalized empirical likelihood with high-dimensional estimating equations
	16:50 - 17:15	Prof. Dr. Jörn Schönberger	Algorithmic Process Decision Making in Volatile Environments: Beyond Intelligent Search
	17:15 - 17:40	Jeffrey Zhang	Semidefinite Programming and Nash Equilibria in Bimatrix Games
	18:30	Excursion + Dinner in the city center	

Saturday 22.07.17	9:30 - 9:55	Prof. Dr. Ostap Okhrin	Using a Kernel-based regression to model the required space for inland vessels
	9:55 - 10:20	Prof. Dr. Peter Song	Identification of risk factors associated with injury in vehicle accidents using Big Data from the USA national automobile crash reporting system
	10:20 - 10:45	Prof. Dr. Heping Zhang	Integration and analysis of imaging and genetic data
	10:45 - 11:00	Coffee Break	
	11:00 - 11:25	Prof. Dr. Annie Qu	Individualized learning and integration for medical imaging data
	11:25 - 11:50	Prof. Dr. Bernhard Schipp	Risk management by statistics of extremes: Assessing the impact of the subprime financial crisis on stock market performance
	11:50 - 12:15	Prof. Dr. Feng Xiao	Exploration of Day-To-Day Route Choice Models By a Virtual Experiment
	12:15 - 12:25	Coffee Break	
	12:25 - 12:50	Prof. Dr. Shulin Zhang	Goodness-of-fit Test of Copula functions for semi-parametric Univariate Time Series Models
	12:50 - 13:15	Leon Zhang	On the empirical relevance of Myerson's Neutral Bargaining Solution
	13:45	Lunch in the city center	
Tuesday 25.07.17	13:00 - 13:30	Prof. Dr. H. G. Krauthäuser	Excursion to the Railway Laboratory
	14:00 - 15:00	Prof. Dr. Daimin Shi and others	Discussion at University level

Bernard Bäker, TU Dresden

Das fliegende Auto kommt später... - Mechatronische Fahrzeuge der Zukunft 2030+

Zu den aktuellen Herausforderungen der Automobilindustrie gehören die weitere Reduzierung der Emissionsgrenzwerte und damit in der Folge eine weiter gesteigerte Elektrifizierung der Antriebstechnologien. Dies geschieht zeitgleich vor dem Hintergrund neuer Stufen einer IT-Vernetzung und der Integration nächster automatisierter Fahrfunktionen.

Das Themenfeld der Mechatronik gewinnt hier zunehmend an Bedeutung, da nur durch integrative mechatronische Systemlösungen die Zielkonflikte bestehend aus Bauraum, Preis und Emissionsverhalten bewältigt werden können. Neuartige Elektrik-/Elektronik-Architekturen mit adaptiven Sensoren und einer hochdatenratigen Kommunikationsanbindung bieten Potenziale für personalisierte Backend-Fahrfunktionen im Kraftfahrzeug der Zukunft.

Jinyuan Chang, SWUFE

A new scope of penalized empirical likelihood with high-dimensional estimating equations

Statistical methods with empirical likelihood (EL) are appealing and effective especially in conjunction with estimating equations through which useful data information can be adaptively and flexibly incorporated. It is also known in the literature that EL approaches encounter difficulties when dealing with problems having high-dimensional model parameters and estimating equations. To overcome the challenges, we begin our study with a careful investigation on high-dimensional EL from a new scope targeting at estimating a high-dimensional sparse model parameters. We show that the new scope provides an opportunity for relaxing the stringent requirement on the dimensionality of the model parameter. Motivated by the new scope, we then propose a new penalized EL by applying two penalty functions respectively regularizing the model parameters and the associated Lagrange multipliers in the optimizations of EL. By penalizing the Lagrange multiplier to encourage its sparsity, we show that drastic dimension reduction in the number of estimating equations can be effectively achieved without compromising the validity and consistency of the resulting estimators. Most attractively, such a reduction in dimensionality of estimating equations is actually equivalent to a selection among those high-dimensional estimating equations, resulting in a highly parsimonious and effective device for high-dimensional sparse model parameters. Allowing both the dimensionalities of model parameters and estimating equations growing exponentially with the sample size, our theory demonstrates that the estimator from our new penalized EL is sparse and consistent with asymptotically normally distributed nonzero components. Numerical simulations and a real data analysis show that the proposed penalized EL works promisingly.

Ostap Okhrin, TU Dresden

Using a Kernel-based regression to model the required space for inland vessels

An important factor in the design of fairways for inland waterway traffic is the width a vessel requires for safe navigation. This is not only given by the constant dimensions of the vessel, but highly influenced by its surroundings. Furthermore it depends on the vessel speed which can be regulated by the steersman. We investigate a complex physical model describing these dependencies and looking for a simpler approach in terms of a multiple regression and a kernel supervised principal component regression. Thereby we reduce the dimensionality and identify the most important influencing variables.

Annie Qu, University of Illinois at Urbana-Champaign

Individualized learning and integration for medical imaging data

This work is motivated by breast cancer imaging data produced by a multimodal multiphoton optical imaging technique. One unique aspect of breast cancer imaging is that different individuals might have breast imaging at different locations, which also creates a technical difficulty in that the imaging background could vary for different individuals. We develop a multilayer tensor learning method to predict disease status effectively through utilizing subject-wise imaging information. In particular, we construct an individualized multilayer model which leverages an additional layer of individual structure of imaging in addition to employing a high-order tensor decomposition shared by populations. In addition, to incorporate multimodal imaging data for different profiling of tissue, cellular and molecular levels, we propose a higher order tensor representation to combine multiple sources of information at different modalities, so important features associated with disease status and clinical outcomes can be extracted effectively. One major advantage of our approach is that we are able to capture the spatial information of microvesicles observed in certain modalities of optical imaging through combining multimodal imaging data. This has medical and clinical significance since microvesicles are more frequently observed among cancer patients than healthy ones, and identification of microvesicles enables us to provide an effective diagnostic tool for early-stage cancer detection. This is joint work with Xiwei Tang and Xuan Bi.

Bernhard Schipp, TU Dresden

Risk management by statistics of extremes: Assessing the impact of the subprime financial crisis on stock market performance

Given the growing need for managing financial risk and the recent global crisis, accurate risk prediction is crucial in banking and finance. In this paper, we show how recent advances in the statistical analysis of extreme events can provide solid fundamentals for modeling these events. Our approach combines self-exciting marked point processes for estimating the tail of loss distributions.

The main result is that the time between extreme events typically plays an important role in the statistical analysis and could therefore be useful to forecast the size and intensity of future extreme events. We illustrate this point by measuring the impact of the subprime and global financial crisis on the various stock markets. Value at Risk (VaR) at different quantiles is used to assess the likeliness of different extreme movements on the DAX, S&P 500 and Nasdaq stock market indices. The results show that the proposed models provide accurate risk measures required by the Basel Committee.

Jörn Schönberger, TU Dresden

Algorithmic Process Decision Making in Volatile Environments: Beyond Intelligent Search

Process decision making in transportation and logistics is challenged by a quite high complexity, e.g. the number of possible decision alternatives is prohibitively high and a full enumeration of all possible decision alternatives is impossible. The concept of heuristic search exploits the idea to evaluate only a subset of these alternatives. This subset is called the search space. Artificial Intelligence (AI) has proposed a comprehensive bunch of ideas how the search space can be evaluated efficiently by incorporating several so-called search paradigms like neighbourhood search, genetic as well as evolutionary algorithms, and ant algorithms and so on. All these approaches perform sufficiently well after an intensive parameter tuning as long as the search space is given and invariable. However, changes in the underlying decision application imply modifications of the search space. Obviously, the so far appropriate algorithm parameters become inappropriate and the

formerly well-working search algorithm becomes inappropriate. This presentation discusses applications in which the aforementioned shortcomings of algorithmic decision making occur. Furthermore, ideas to overcome this shortcoming are outlined and their general applicability are shown. But we also address so far unsolved challenges in algorithmic decision making in dynamic environments that might benefit from setting up an adequate search space exploiting huge data sets. The integration of an appropriate search space definition using big data techniques with powerful search algorithms promises a significant progress in decision making in dynamic process environments beyond improving search algorithms.

Peter Song, UMICH

Identification of risk factors associated with injury in vehicle accidents using Big Data from the USA national automobile crash reporting system

The motor vehicle accident causes injury in that victims will subsequently suffer both physical and emotional hardships. Identifying risk factors associated with injury in vehicle accidents is important to develop preventive measures to minimize the chance of the occurrence of injury. Leveraging Big Data sources from the USA national automobile crash reporting system, we analyze the injury data using the method of divide-and-combine (MODAC) proposed by Tang et al. (2017). This MODAC allows us to perform Big Data analysis on a Hadoop platform via the MapReduce scheme, where the statistical inference is established by the means of confidence distribution. This analysis confirmed that seat belt and speed limit are the top risk factors among those associated with injury, and however, interestingly race is not an important risk factor for the chance of injury in vehicle accidents.

Feng Xiao, SWUFE

Exploration of Day-To-Day Route Choice Models By a Virtual Experiment

Hui Xiao, SWUFE

Ranking and Selection: An Optimal Computing Budget Allocation Procedure

Discrete event simulation is a widely used tool for evaluating the performance of stochastic systems since any level of detail of the system can be modeled via simulation. Examples of such successful applications can be found in inventory and supply chain systems, production systems and healthcare systems. However, due to the slow convergence of the mean performance, the efficiency of simulation remains a big concern, especially when the simulation is economically expensive or time consuming. This presentation gives an efficient ranking and selection procedure called Optimal Computing Budget Allocation that can greatly improve the efficiency of simulation. The key component of our methodology is introduced and the new development of this procedure will be also address in this presentation.

Heping Zhang, YALE

Integration and analysis of imaging and genetic data

Identification of genetic markers for neurological disorders has been challenging. It is now widely believed that genetic markers do not necessarily impact on the behaviors directly, and instead through brain functions which are commonly measured by imaging data. Therefore, it is important and useful to integrate imaging and genetic data, and analyze the integrated data in order to better

identify genetic markers for neurological disorders. I will present specific datasets and examples to illustrate how our methods work and the importance of our findings.

Jeffrey Zhang, Princeton University

Semidefinite Programming and Nash Equilibria in Bimatrix Games

We explore the power of semidefinite programming (SDP) for finding additive ϵ -approximate Nash equilibria in bimatrix games. We introduce an SDP relaxation for a quadratic programming formulation of the Nash equilibrium (NE) problem and provide a number of valid inequalities to improve the quality of the relaxation. If a rank-1 solution to this SDP is found, then an exact NE can be recovered. We show that for a (generalized) zero-sum game, our SDP is guaranteed to return a rank-1 solution. Furthermore, we prove that if a rank-2 solution to our SDP is found, then a $5/11$ -NE can be recovered for any game, or a $1/3$ -NE for a symmetric game. We propose two algorithms based on iterative linearization of smooth nonconvex objective functions which are designed to produce rank-1 solutions. Empirically, we show that these algorithms often recover solutions of rank at most two and ϵ close to zero. We then show how our SDP approach can address two (NP-hard) problems of economic interest: finding the maximum welfare achievable under any NE, and testing whether there exists a NE where a particular set of strategies is not played. Finally, we show that by using the Lasserre/sum of squares hierarchy, we can get an arbitrarily close spectrahedral outer approximation to the convex hull of Nash equilibria, and that the SDP proposed in this paper dominates the first level of the sum of squares hierarchy.

Leon Zhang, Stanford University

On the empirical relevance of Myerson's Neutral Bargaining Solution

This study compares a real-world bilateral bargaining game to a theoretically "fair" or "neutral" bargaining solution proposed by Myerson (1984). The neutral bargaining solution generalizes the Nash bargaining solution to incomplete information settings. We implement an algorithm for computing a neutral bargaining solution given estimates of buyer and seller valuation distributions. We then apply this procedure using estimates of buyer and seller valuation distribution from data on dealer-to-dealer bargaining over used cars. We find that the real-world bargaining outcomes differ starkly from those in the neutral bargaining solution, suggesting that real-world bargaining falls short of achieving "fairness" or "neutrality" in the Myerson sense.

Shulin Zhang, SWUFE

Goodness-of-fit Test of Copula functions for semi-parametric Univariate Time Series Models

In this paper, we propose a goodness-of-fit test, named "in-and-out-of-pseudo-likelihood" (PIOL) ratio test, to check for misspecification in semi-parametric copula models for univariate time series. The proposed test extends the idea of IOS test by Presnell and Boos (2004) and PIOS test by Zhang et al. (2016), which are problematic for univariate time series. The PIOL test provides a more general framework which can be applied to both independent data and correlated data. In addition, an approximation method is implemented to alleviate the computational burden, which also exists for IOS/PIOS tests. Asymptotic properties of the proposed PIOL statistic are discussed. The finite-sample performance is examined through simulation studies.