# Otsu Seminar

## Statistical Data Sciences

Date: March 01 — 03, 2020

Venue: Biwako Hotel

Organizer: Masanobu TANIGUCHI

(Research Institute for Science & Engineering, Waseda University)

Supported by:

"Introduction of General Causality to Various Data & Its Innovation of the Optimal Inference"

JSPS KAKENHI Kiban (S)
Grand-in-Aid No. 18H05290 (M. Taniguchi)

# Otsu Seminar

## “Statistical Data Sciences”

**Date: March 1 — 3, 2020**

**Venue:  Biwako Hotel**

(www.keihanhotels-resorts.co.jp/biwakohotel/english/)

**Supported by**

“Introduction of General Causality to Various Data
& Its Innovation of the Optimal Inference”

JSPS Kiban (S) No.18H05290

(M. Taniguchi, Waseda University)

# *Program*

## March 1

**Session (I)  16:30 — 18:00**

16:30—17:00: Goto, Yuichi

*Random Effect ANOVA model for Time Series Data*

17:00—17:30: Fujimori, Kou

*The variable selection by the Dantzig selector for Cox's proportional hazards model*

17:30—18:00: Kimura, Akitoshi

*The asymptotic variance estimators of the correlation estimator between latent processes and their asymptotic properties*

**Session (II) 20:00 — 22:00**

20:00—20:30: Akashi, Fumiya

*Self-weighted GEL method for heavy-tailed ARMA models and its application*

20:30—21:00: Liu, Yan

*Sphericity test for high-dimensional time series*

21:00—21:30: Dou, Xiaoling

*EM algorithms for estimating B-spline copulas*

# March 2

**Session (III) 9:30 — 10:30**

9:30—10:30: Master's theses corner (*Supervisor: Taniguchi, M.*)

Wulan, Kaneko, Kojima, Nakamura

**Session (IV) 10:45 — 12:15**

10:45—11:15: Shiohama, Takayuki

*Non-stationary Time Series Classification using Topological Data Analysis*

11:15—11:45: So, Mike K P (HKUST)

*Modeling of Coevolution in Financial Networks and its Implications for Market Risk Prediction*

11:45—12:15: Hirukawa, Junichi

*Weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation*

**Session (V) 13:30 — 15:00**

13:30—14:00: Chen, Ying (National University of Singapore)

*Topic Sentiment Asset Pricing with DNN Supervised Learning*

14:00—14:30: Chung, Moo K. (University of Wisconsin-Madison)

*Topological data analysis on trees*

14:30—15:00: Eichler, Michael (Maastricht University)

*Causal learning with non-markovian constraints*

**Session (VI) 15:15 — 17:15**

15:15—15:45: Francq, Christian (ENSAE, France)

*Testing the existence of moments for GARCH processes*

15:45—16:15: Huang, Shih-Feng (National University of Kaohsiung)

*Modeling Financial Time Series with Soft Information*

16:15—16:45: Lin, Liang-Ching (National Cheng Kung University)

*Symbolic Interval-Valued Data Analysis based on Normality Assumption*

16:45—17:15: Monti, Anna Clara (University of Sannio)

*Ordinal responses: a robust approach*

**Session(VII) 20:00 — 22:00**

20:00—20:30: Preinerstorfer, David (University Libre Brussels)

*How to avoid the zero-power trap in testing for correlation*

20:30—21:00: Ronchetti, Elvezio (University of Geneva)

*Prediction and robustness*

21:00—21:30: von Sachs, Rainer (Université catholique de Louvain)

*Intrinsic Data Depth for Hermitian Positive Definite Matrices*

21:30—22:00: Wang, Yuan (University of South Carolina)

*Statistical signal analysis: A Topological Approach*

# March 3

**Session (VIII)**

9:30—11:00: Taniguchi, Masanobu

*A short history of international collaboration*

# *Abstracts*

## March 1

### Yuichi Goto

**Title: Random Effect ANOVA model for Time Series Data**

Abstract: Random effect ANOVA model has been studied by many authors since the early 1900s for independent observations. In this talk, we discuss the testing problem for the existence of random effect for random effect ANOVA model for time series data. First, we show the exact distribution of estimator for the variance of random effect. Then, we construct the chi-squared test for the existence of the random effect and show the consistency of the test.

### Kou Fujimori

**Title: The variable selection by the Dantzig selector for Cox's proportional hazards model**

Abstract: The proportional hazards model proposed by D. R. Cox in a high-dimensional and sparse setting is discussed. The regression parameter is estimated by the Dantzig selector, which will be proved to have the variable selection consistency. This fact enables us to reduce the dimension of the parameter and to construct asymptotically normal estimators for the regression parameter and the cumulative baseline hazard function.

## Akitoshi Kimura

### Title: The asymptotic variance estimators of the correlation estimator between latent processes and their asymptotic properties

Abstract: In this talk, we treat a model in which the finite variation part of a two-dimensional semi-martingale is expressed by time-integration of latent processes. We propose a correlation estimator between the latent processes and show its consistency and asymptotic mixed normality. Moreover, we propose two types of estimators for asymptotic variance of the correlation estimator and show their asymptotic properties in a high frequency setting. We focus on the proof of the asymptotic properties. Our model includes doubly stochastic Poisson processes whose intensity processes are correlated It{\^o} processes.

## Fumiya Akashi

### Title: Self-weighted GEL method for heavy-tailed ARMA models and its application

Abstract: This talk considers a testing problem of linear hypothesis on the coefficients of possibly infinite variance ARMA processes. To overcome the difficulties brought by heavy-tails, this talk constructs the least absolute deviations regression and self-weighting-based generalized empirical likelihood (GEL) statistics for the testing problem. By the self-weighting and GEL, the proposed test statistic is shown to have a pivotal chi-squared limit distribution regardless of whether the model has finite variance or not. In the latter half of this talk, we apply the self-weighted GEL method to the test of causality of heavy-tailed time series models. (Joint work with M. Taniguchi and A.C. Monti)

## Yan Liu

### Title: Sphericity test for high-dimensional time series

Abstract: We consider the testing problem for the sphericity hypothesis regarding the covariance matrix of high-dimensional time series. Recently, test statistics for sphericity of independent and identically distributed high-dimensional random variables have been studied under the condition that both the sample size n and the dimension p diverge to infinity. A test statistic for sphericity has been proved to be well-behaved even when p is larger than n. We investigate the test statistic under the situations of high-dimensional time series. The asymptotic null distribution of the test statistic is shown to be standard normal when the observations come from Gaussian stationary processes. In the simulation study, we illustrate the properties of the test statistic for several high-dimensional time series models. In our empirical study, we apply the test to the portfolio selection problem. (Joint work with Yurie Tamura and Masanobu Taniguchi)

## Xiaoling Dou

### Title: EM algorithms for estimating B-spline copulas

Abstract: The B-spline copula is a generalization of the Bernstein copula. It is defined by replacing the Bernstein basis functions by B-spline basis functions. This change requires the copula parameters satisfy slightly different conditions, in spite of the copula form remains the same. Because the Bernstein copula can be considered as a finite mixture distribution for given marginals, we can use EM algorithm methods to estimate the Bernstein copula. Since this idea is also available for the B-spline copula, we propose to generate the existing EM algorithms of the Bernstein copula to estimate the B-spline copula by changing the basis functions and the parameter conditions. Illustrative examples are presented with real data sets.

# March 2

## Takayuki Shiohama

### Title: Non-stationary Time Series Classification using Topological Data Analysis

Abstract: Time series classification (TSC) is an important and challenging problem in data mining. There are hundreds of algorithms for TSC available with the increase of  time series data availability. For more details, we refer recent work of Bagnall et al. (2017) and review paper of Fawaz et al. (2019).  Some of the machine learning algorithms are based on the bug of patterns, and learning patterns of similarity are key feature extraction for time series data.  In this study, we employ feature extraction of time series using Topological Data Analysis (TDA). TDA  refers to a collection of methods for finding topological structure in data. Until recently, topological inference relied on deterministic approaches, and it is well known that these inference is easily affected by outliers and/or noisy datasets. Moreover, the high computational costs are required for computing persistence homology with complex datasets in time and space. To overcome these circumstances, two approaches are proposed in literature. The first approach is the bootstrap estimation for persistence diagrams and landscapes of Chazal et al. (2014). The second refers the methods of subsampling of Chazal et al. (2015). In this study, we introduce

an unsupervised classification learning for several non-stationary time series using topological features. The bootstrap and subsampling methods are implemented to compare the performance of classifications.

## Mike K P So

### Title: Modeling of Coevolution in Financial Networks and its Implications for Market Risk Prediction

Abstract: This paper studies a network-based model for capturing stock comovement dynamics. We consider stochastic actor-oriented models (SAOMs) to analyze the coevolution between stock correlation networks and market risk, controlling for market-specific and firm-specific variables. Specifically, we propose two mechanisms, namely the stock selection process and the stock influence process, to describe the relationship between financial networks and risk. SAOM is a continuous-time model that assumes the existence of a number of stochastic first-order Markov mini-steps between each discrete observable time steps. This assumption is valid as stock comovement occurs in a high-frequency manner. Besides, it formulates a stochastic model framework for the modeling of stock comovement and market risk, assuming that changes in network link and risk are conditionally independent of each other given their current state. We also consider model comparison of various network structures for market risk prediction. Using time-series return and volume data of stocks in S&P500, we find support to both stock selection and stock influence processes. For the stock selection process, we find that stocks are less likely to move with others if they are linked with stocks of high risk in both return and volume networks. This is also

true for stocks that are indirectly linked with others of high risk in the volume network. For the stock influence process, we find that stocks are influenced by others of similar risk, but not by stocks of high risk. There also exist significant cross-network triple effects that drive stock comovement.

## Junichi Hirukawa

**Title: Weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation**

Abstract: An integral transformation which changes a fractional Brownian motion to a process with independent increments has been given. A representation of a fractional Brownian motion through a standard Brownian motion on a finite interval has also been given. On the other hand, it is known that the partial sum of the discrete time fractionally integrated process ($I(d)$ process) weakly converges to a fractional Brownian motion in infinite interval representation. In this talk we derive the weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation. (Joint work with Kou Fujimori)

## Ying Chen

**Title: Topic Sentiment Asset Pricing with DNN Supervised Learning**

Abstract: We develop an innovative deep neural network (DNN) supervised learning approach to extracting insightful topic sentiments from analyst reports at the sentence level and incorporating this qualitative knowledge in asset pricing and portfolio construction. The topic sentiment analysis is performed on 113,043 Japanese analyst reports and the topic sentiment asset pricing model delivers superior predictive power on stock returns with adjusted R2 increasing from 1.6% (benchmark model without sentiment) to 14.0% (in-sample) and 13.4% (out-of-sample). We find that topics reflecting the subjective opinions of analysts have greater impact than topics of objective facts and justification of the quantitative measures. (Joint work with Hitoshi Iwasaki and Jun Tu)

## Moo K. Chung

### Title: Topological data analysis on trees

Abstract: We introduce the Hilbert space approach for analyzing a collection of trees (graphs with no loops). Due to the lack of one-to-one correspondence between trees at the node or edge level, building a coherent statistical model has been a challenge. It is unclear how one even set up a simple linear model across trees with no correspondence. We present an algebraic representation called the Weighted Fourier Series (Statistica Sinica 18:1269-1291) that enables the parametric representation of trees. Statistical models can be then set up across the parameters of the representation. The method further enables us to build the proper inner product space and define the inner product between trees.

## Michael Eichler

### Title: Causal learning with non-markovian constraints

Abstract: In systems that are affected by latent variables conditional independences are often insufficient for inference about the structure of the underlying system. One common example is a system in which four observed variables $X_1$, $X_2$, $X_3$, and $X_4$ are conditionally independent given a fifth unobserved variable $Y$. While there are no conditional independences among the observed variables, the tetrad representation theorem states that they must satisfy the so-called tetrad constraints $\corr(X_i,X_j) \,\corr(X_k,X_l) - \corr(X_i,X_l)\,\corr(X_k,X_j)=0$ for all permutations $i, j, k, l$ of the indices $1,2,3,4$ (e.g. Spirtes et al., 2001). In this talk, we discuss the extension of this result to the multivariate time series case. We derive a spectral version of the above tetrad constraints in terms of spectral coherences and present a test for vanishing tetrad constraints in the frequency domain. Critical values for the test are obtained by asymptotic results and by bootstrap techniques. We discuss the relevance of our results for causal learning from observational time series data.

## Christian Francq

**Title: Testing the existence of moments for GARCH processes**

Abstract: It is generally admitted that many financial time series have heavy tailed marginal distributions. When time series models are fitted on such data, the non-existence of appropriate moments may invalidate standard statistical tools used for inference. Moreover, the existence of moments can be crucial for risk management, for instance when risk is measured through the expected shortfall. This paper considers testing the existence of moments in the framework of GARCH processes. While the second-order stationarity condition does not depend on the distribution of the innovation, higher-order moment conditions involve moments of the independent innovation process. We propose tests for the

existence of high moments of the returns process which are based on the joint asymptotic distribution of the Quasi-Maximum Likelihood (QML) estimator of the volatility parameters and empirical moments of the residuals. A bootstrap procedure is proposed to improve the finite-sample performance of our test. To achieve efficiency gains we consider non Gaussian QML estimators founded on reparametrizations of the GARCH model, and we discuss optimality issues. Monte-Carlo experiments and an empirical study illustrate the asymptotic results. (Joint work with Jean-Michel)

## Shih-Feng Huang

**Title: Modeling Financial Time Series with Soft Information**

Abstract: A hysteretic autoregressive model with GARCH effects and soft information, denoted by SHAR-GARCH, is proposed to model financial time series. The soft information contained in daily news is extracted by the techniques of support vector machine and principal component analysis. A Markov Chain Monte Carlo algorithm is proposed for estimating model parameters. A corresponding risk-neutral SHAR-GARCH model is derived by Esscher transform for option pricing. The returns and options of the S&P500 index and the daily news posted on the website of Reuters are used for our empirical study. The numerical results indicate that the proposed model has satisfactory performances in depicting the dynamics of financial time series and in pricing deep-in-the-money options. (Joint work with Hui-Chiao Huang)

## Liang-Ching Lin

**Title: Symbolic Interval-Valued Data Analysis based on Normality Assumption**

Abstract: This study considers iid interval-valued data in which the underlying distribution is assumed to be normality. An approximate expectation formula of order statistics from normal distributions is used in the univariate case to estimate the mean and variance via the method of moment. In contrast, in the bivariate case, we derived the likelihood function based on the bivariate copulae and obtained the corresponding maximum likelihood estimator. Monte Carlo simulations are conducted to evaluate our methods of estimation, confirming their validity and superiority to other methods. Real data example is also carried out for air quality indices. (Joint work with Sangyeol Lee)

## Anna Clara Monti

**Title: Ordinal responses: a robust approach**

Abstract: Cumulative models are widespread in various fields to describe the dependence of an observed ordinal response on subjects' covariates. In this context, as well as in many other statistical models, anomalous data, in the form of outlying covariates or incoherent responses, may affect the reliability of the inferential analyses. The talk shows that by an appropriate choice of the link function, robust estimation and testing can be can be based on the likelihood function, and the fitted model retains a good predictive ability. (Joint work with Valentino Scalera and Maria Iannario)

## David Preinerstorfer

**Title: How to avoid the zero-power trap in testing for correlation**

Abstract: In testing for correlation of the errors in regression models the power of tests can be very low for strongly correlated errors. This counterintuitive phenomenon has become known as the “zero-power trap”. In this talk I discuss solutions to the problem, and I provide illustrations in a situation involving network-generated correlation.

## Elvezio Ronchetti

**Title: Prediction and robustness**

Abstract: We provide a general discussion of the robustness issues in a prediction framework and analyze their implications in different areas, including classification, insurance, and estimation in finite populations. We illustrate more specifically these issues in the prediction of nonlinear indices (such as inequality or poverty measures) for small areas and in the presence of outliers. We propose two approaches to calibrate for the bias of nonlinear functionals, such as the Gini index and when the so called representative outliers come from a skewed heavy tail distribution.
(Joint work with Setareh Ranjbar and Stefan Sperlich)

## Rainer von Sachs

**Title: Intrinsic Data Depth for Hermitian Positive Definite Matrices**

Abstract: Non-degenerate covariance, correlation and spectral density matrices are necessarily symmetric or Hermitian and positive definite. This paper develops statistical data depths for collections of Hermitian positive definite matrices by exploiting the geometric structure of the space as a Riemannian manifold. The depth functions allow one to naturally characterize most central or outlying matrices, but also provide a practical framework for inference in the context of samples of positive definite matrices. First, the desired properties of an intrinsic data depth function acting on the space of Hermitian positive definite matrices are presented. Second, we propose two pointwise and integrated data depth functions that satisfy each of these requirements and investigate several robustness and efficiency aspects. As an application, we construct depth-based confidence regions for the intrinsic mean of a sample of positive definite matrices, which is applied to the exploratory analysis of a collection of covariance matrices in a multicenter clinical trial.

## Yuan Wang

### Title: Statistical signal analysis: A Topological Approach

Abstract: Classical signal processing methods typically transform signal in the time domain into another domain through Fourier and wavelet transforms. An alternative approach to signal analysis has recently been advanced through topological data analysis (TDA), which captures the topological changes of connected components and holes in data through a multiscale descriptor of persistent homology (PH). The PH algorithm reveals the changes of topological structures across different temporal and spectral scales of signals that are unobservable through the standard fixed-scale statistical approach. This talk briefly introduces some recent results of brain signal analysis through the topological approach.

## Masanobu Taniguchi

### Title: Recent Development for Collaborative Reserch based on JSPS Fundings

Abstract: This talk delivers a short history of development for collaborative research over 15 years based on JSPS fundings (M.Taniguchi). The related researchers are Puri, M.L., Hallin, M., Garderen, K.J., Lee, S., Veredas, D., DiCiccio, T., Monti, A.C., Petkovic, A., Patilea, V., Giraitis, L., Taqqu, M., Chen, C.W.S., Pewsey, A. etc. Integrating these collaborative research, we have published four English books. Some future view will be provided in relation to Waseda symposium and Otsu seminar.