WASEDA INTERNATIONAL SYMPOSIUM

High Dimensional Statistical Analysis for Time Spatial Processes & Quantile Analysis for Time Series (March 2016, Waseda)

Date: February 29 (Mon.) - March 2 (Wed.), 2016

Venue: Waseda University, Nishi-Waseda Campus

Building 55S, 2nd Floor, Room 4

Organizer: Masanobu Taniguchi

(Research Institute for Science & Engineering, Waseda Univ.)

Supported by Kiban(A-15H02061) & Houga(26540015)

Waseda International Symposium

High Dimensional Statistical Analysis for Time Spatial Processes & Quantile Analysis for Time Series

(March 2016, Waseda)

February 29 (Mon.) - March 2 (Wed.), 2016
Waseda University, Nishi-Waseda Campus
Building 55S, 2nd Floor, Room 4
(Access map: http://www.sci.waseda.ac.jp/eng/access/)

Organizer: Masanobu TANIGUCHI

(Research Institute for Science & Engineering, Waseda University)

Supported by

(1) Kiban (A-15H02061)

M.Taniguchi, Research Institute for Science & Engineering, Waseda University

(2) Houga (26540015)

M.Taniguchi, Research Institute for Science & Engineering, Waseda University

Program

(* Speaker)

February 29

Session (I): 13:30 -15:20: chaired by R. Davis

13:30 - 13:40: Masanobu Taniguchi (Waseda Univ.)
Opening

13:40 - 14:30: Fumiya Akashi* (Waseda Univ.) and Xiaofeng Shao (UIUC)
LAD-based empirical likelihood method and its local asymptotic power

14:30 - 15:20: Ryo Yoshida (Institute of Statistical Mathematics)
Bayesian approach towards data science driven materials discovery

15:20 - 15:30: Break

Session (II): 15:30 -17:10: chaired by M. Hallin

15:30 - 16:20: William T.M. Dunsmuir (Univ. of New South Wales)
Long Longitudinal Data Modelling

16:20 - 17:10: Murad Taqqu (Boston Univ.)
Self-similar processes with stationary increments on Wiener chaos

March 1

Session (III): 9:30 - 12:00: chaired by L. Giraitis

9:30 - 10:20: Yan Liu*, Yujie Xue and Masanobu Taniguchi (Waseda Univ.),
Minimax extrapolation error of stationary processes

10:20 - 11:10: Shu Hui Yu (National Univ. of Kaohsiung)
Prediction errors in unit-root models

11:10 - 12:00: Kazuyoshi Yata* and Makoto Aoshima (Univ. of Tsukuba)
Inference on high-dimensional covariance structures with fewer observations than the dimension

12:00 - 13:00: Lunch

Session (IV): 13:00 - 15:30: chaired by W. Dunsmuir

13:00 - 13:50: Hiroko Kato Solvang* (Institute of Marine Research Sundstgt.) and Masanobu Taniguchi
Microarray analysis using rank order statistics for ARCH residual empirical process

13:50 - 14:40: Satoshi Kuriki* (Institute of Statistical Mathematics), Tomoyuki Shirai and Trinh Khanh Duy
Some distributions associated with the cone of positive semidefinite matrices and their applications

14:40 - 15:30: Ching Kang Ing (Institute of Statistical Science, Academia Sinica)
On model selection from a finite family of possibly misspecified models

15:30 - 15:40: Break

Session (V): 15:40 - 17:20: chaired by M. Taqqu

15:40 - 16:30: Richard Davis (Columbia Univ.)
On Consistency/Inconsistency of MDL Model Selection for Piecewise Autoregression

16:30 - 17:20: Marc Hallin (Université libre de Bruxelles)
Monge-Kantorovich Ranks and Signs

18:20- Dinner

March 2

Session (VI): 9:30 - 12:00: chaired by C.K. Ing

9:30 - 10:20: Hiroaki Ogata* (Tokyo Metropolitan Univ.), Toshihiro Abe, Takayuki Shiohama and Hiroyuki Taniai
A circular autocorrelation of stationary circular Markov processes

10:20 - 11:10: Gwo Dong Lin (Institute of Statistical Science, Academia Sinica)
Recent Developments on the Moment Problem

11:10 - 12:00: Liudas Giraitis* (Queen Mary Univ. of London), V. Dalla and P.C.B. Philips
Testing for stability of the mean of heteroskedastic time series

Abstracts

Akashi, F.

LAD-based empirical likelihood method and its local asymptotic power

Abstract: In this talk, we construct the least absolute deviation-based empirical likelihood (LAD-EL) test statistic for testing problem of unknown parameters of linear regression models. It is shown that the test statistic has the standard chi-square limit distribution, which is pivotal. As a result, we can carry out a testing procedure without estimating any unknown quantities of the model, such as density functions of error terms. Furthermore, the limit distribution of LAD-EL and classical LAD-based test statistics under local contiguous alternatives are elucidated. Based on the result, the asymptotic local power of the proposed test is compared with that of the classical one, and it is shown that proposed test has higher power than classical one in some special cases. We also investigate empirical power of the proposed test by simulation experiments, and our approach is shown to have advantages in many senses.

Yoshida, R.

Bayesian approach towards data science driven materials discovery

Abstract: Computational design of small organic molecules involves a multi-objective combinatorial optimization that it is impractical to fully explore a vast landscape of the structure-property relationship. The chemical space of interest consists of more than 1060 potential structures. The objective is to identify as-yet-undiscovered compounds that exhibit various types of desired properties. The basis of our method is founded on a Bayesian perspective of QSPR (quantitative structure-property relationship) and inverse-QSPR. The proposed method begins with obtaining a set of QSPR models that predict target properties of input compounds. Instead of exploring a best single model, an ensemble learning technique is employed to derive much higher prediction accuracy from an integration of different molecular descriptors and constituent models. Substituting the QSPR models and a prior distribution into the Bayes

law, we derive a posterior distribution for the inverse-QSPR, which is conditioned by the desired properties. Monte Carlo computation is conducted to produce new compounds from the posterior. The prior plays a vital role in putting reality into the computationally generated compounds in which frequent patterns of chemical fragments are compressed to a language model learned from SMILES strings of existing compounds. The methodological basis and the great potential of our algorithm, e.g. detection capability and computation, are illustrated through a computational design of new dye materials and drug compounds.

Dunsmuir, W.
Long Longitudinal Data Modelling

Abstract: The terminology “long longitudinal data analysis” is used by us for multiple independent time series in which impacts of covariates are modelled using fixed and random effects mixed models with serial dependence. Given covariates, random effects on these covariates and a random process, the observed responses for each individual series are independent with an exponential family distribution. The structure of these models is similar to that of longitudinal data models with random effects. However, in contrast to that setting, in which there are many cases and few to moderate observations per case, the long longitudinal data setting has many observations per series and a few to moderate number of series. In examples such as these, for which shared regression effects need to be tested for equality across multiple series, use of random effects to capture between series regression variation can be useful. Also, unlike the usual assumption in longitudinal data analysis, the form and strength of serial dependence needs to vary between series in the applications we encounter. Two types of models for the serially dependent random processes are considered. Observation driven models use past values of the observed series to induce serial dependence. Parameter driven models use a latent unobserved process to induce dependence. For the observation driven class we present a simple and easily implementable approach to estimation of the mixed model based on adaptive Gaussian quadrature and the Laplace approximation used in conjunction with the existing R-package glarma software for fitting observation driven models for univariate discrete valued time series. For the

parameter driven class, high dimensional integrals need to be estimated along with the lower dimensional random effects integrals and for these Laplace approximation augmented by importance sampling is used to obtain the likelihood. Extensions to a more general model in which variations in serial dependence parameters across series is specified using additional random effects will be discussed. Two applications will illustrate the models and methods. The first, with Poisson responses, reassesses the impact on single vehicle nighttime fatalities of lowering the legal BAC limit for drivers in 17 US states. A second example, with binary responses, arises from a panel of listeners responding to changing musical features. Both applications highlight the need for flexibility in specifying serial dependence across series.

Taqqu, M.

Self-similar processes with stationary increments on Wiener chaos

Abstract: Self-similar processes with stationary increments are important because they characterize the scaling limits of sums of stationary sequences. In this talk, we review the notions of self-similarity, short and long memory, provide a partial history of the subject and introduce a broad class of self-similar processes represented by a multiple stochastic integral, called the generalized Hermite processes. We show that the normalized sums of some nonlinear long-memory stationary sequences converges to these generalized Hermite processes. (Joint work with Shuyang Bai)

Liu, Y.

Minimax extrapolation error of stationary processes

Abstract: We consider minimax extrapolators in line with the seminal work by Huber, who introduced the minimax variance to the field of statistics. Extrapolation problem, as known as the extremal problem, can be regarded as a linear approximation on the unit circle in the complex plane. Although robust one-step ahead predictor and interpolator has already been considered separately in the previous literature, we will give a general framework from both

the point of linear approximation on the unit circle with different information set and the point of the error evaluated under the L^p norm. We show that under certain conditions, there exists a minimax predictor for the class of spectral distributions $¥epsilon$-contaminated by unknown spectral distributions. With examples of multiple step prediction and interpolation, our results also contain prediction problems that there exist several missing observations in the past. (Joint work with Y. Xue and M. Taniguchi)

Yu, S-H.
Prediction errors in unit-root models

Abstract: Assume that observations are generated from the first-order autoregressive (AR) models with/without linear time trend and the unknown model coefficients are estimated by least squares. For both cases, we develop an asymptotic expression for the mean squared prediction error (MSPE) of the least squares predictor. First, for the model without linear time trend, we show that the term of order 1/n in this error, where n is the sample size, is twice as large as the one obtained from the AR(1) model satisfying the stationary assumption. Moreover, while the correlation between the squares of the (normalized) regressor variable and normalized least squares estimator is asymptotically negligible in the stationary AR(1) model, we have found that the correlation has a significantly negative value in the random walk model. Secondly, for the model with linear time trend, we develop an asymptotic expression for MSPE in the presence of a unit root. As a by-product, we also obtain a connection between the MSPE and the growth rate of the Fisher information. The key technical tool used to derive these results is the negative moment bound for the minimum eigenvalue of the normalized Fisher information matrix.

Yata, K.
Inference on high-dimensional covariance structures with fewer observations than the dimension

Abstract: In this talk, we consider testing the correlation coefficient matrix between two subsets of high-dimensional variables. The test is a very important tool of pathway analysis or graphical modeling for high-dimensional data. We produce a test statistic by using the extended cross-data-matrix (ECDM) methodology and show the unbiasedness of ECDM estimator. We also show that the ECDM estimator has the consistency property and the asymptotic normality in high-dimensional settings. We propose a test procedure by the ECDM estimator and evaluate its asymptotic size and power theoretically and numerically. We give several applications of the ECDM estimator. Finally, we demonstrate how the test procedure performs in actual data analyses by using a microarray data set. (Joint work with M. Aoshima)

Kato, S.H.

Microarray analysis using rank order statistics for ARCH residual empirical process

Abstract: Statistical two-group comparisons are widely used to identify the significant differentially expressed (DE) signatures against a therapy response for microarray data analysis. We applied a rank order statistics based on an Autoregressive Conditional Heteroskedasticity (ARCH) residual empirical process to DE analysis. This approach was verified by simulation study and was considered for publicly available datasets and compared with two-group comparison by original data and Autoregressive (AR) residual. The significant DE genes by the ARCH and AR residuals were reduced by about 20-30% to these genes by the original data. Almost 100% of the genes by ARCH are covered by the genes by the original data unlike the genes by AR residuals. GO enrichment and Pathway analyses indicate the consistent biological characteristics between genes by ARCH residuals and original data. ARCH residuals array data might contribute to refine the number of significant DE genes to detect the biological-feature as well as ordinal microarray data. (Joint work with M. Taniguchi)

Kuriki, S.

Some distributions associated with the cone of positive semidefinite matrices and their applications

Abstract: Let A be a standard Gaussian random matrix in the space Sym(n) of n $\times$ n symmetric (or Hermitian) matrices. Let PD(n) be the cone of positive semidefinite matrices in Sym(n). In this talk, we derive the distribution of the squared distance between the random matrix A and the cone PD(n). This distribution appears as the null distribution of the likelihood ratio criterion for testing multivariate variance components. In real and complex normal population cases, the distributions are mixtures of chi-square distributions with weights expressed in terms of the Pfaffian and the determinant, respectively. Moreover, when the size n of the matrix goes to infinity, by modifying Johansson's (1998) central limit theorem for eigenvalues of random matrices, the limiting distribution is proved to be Gaussian. This property of limiting Gaussianity was conjectured in previous literature (e.g., Amemiya, Anderson and Lewis, 1990). (Joint work with Tomoyuki Shirai and Trinh Khanh Duy)

Ing, C-K.

On model selection from a finite family of possibly misspecified models

Abstract: Model selection problems are usually classified into two categories according to whether the data generating process (DGP) is included among the family of candidate models. The first category assumes that the DGP belongs to the candidate family, and the objective of model selection is simply to choose this DGP. The second category assumes that the DGP is not one of the candidate models. In this case, one of the top concerns is to choose the model having the best prediction capability. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a key point of contention. In this article, we propose a misspecification-resistant information criterion (MRIC) to rectify this difficulty under the fixed-dimensional framework, which requires that the set of candidate models is fixed with the sample size. We prove the asymptotic efficiency of MRIC regardless of whether the true model belongs to the candidate family or not. We also illustrate MRIC's

finite-sample performance using Monte Carlo simulation.

Davis, R.

On Consistency/Inconsistency of MDL Model Selection for Piecewise Autoregression

Abstract: The Auto-PARM (Automatic Piecewise AutoRegressive Modeling) procedure, developed by Davis, Lee, and Rodriguez-Yam (2006), uses the minimum description length (MDL) principle to estimate the number and locations of structural breaks in a non-stationary time series. Consistency of this model selection procedure has been established when using conditional maximum (Gaussian) likelihood variance estimates. In contrast, the estimate of the number of change-points is inconsistent in general if Yule-Walker variance estimates are used instead. This surprising result is due to an exact cancellation of first-order terms in a Taylor series expansion in the conditional maximum likelihood case, which does not occur in the Yule-Walker case. (Joint work with Stacey Hancock and Yi-Ching Yao)

Hallin, M.

Monge-Kantorovich Ranks and Signs

Abstract: Unlike the real line, the real space $\mathbb{R}^K$, $K \geqq 2$ is not “naturally" ordered. As a consequence, such fundamentals univariate concepts as quantile and distribution functions, ranks, signs, all order-related, do not straightforwardly extend to the multivariate context. Since no universal pre-existing order exists, each distribution, each data set, has to generate its own—the rankings behind sensible concepts of multivariate quantile, ranks, or signs, inherently will be distribution-specific and, in empirical situations, data-driven. Many proposals have been made in the literature for such orderings—all extending some aspects of the univariate concepts, but failing to preserve the essential properties that make classical rank-based inference a major inferential tool in the analysis of semiparametric models where the density of some underlying noise remains unspecified: (i) exact distribution-freeness, and (ii) asymptotic semiparametric

efficiency, see Hallin and Werker (2003). Ranks and signs, and the resulting inference methods, are well understood and well developed, essentially, in two cases: one-dimensional observations, and elliptically symmetric ones. We start by establishing the close connection, in those two cases, between classical ranks and signs and measure transportation results, showing that the rank transformation there actually reduces to an empirical version of the unique gradient of convex function mapping a distribution to the uniform over the unit ball. That fact, along with a result by McCann (1995), itself extending the celebrated *polar factorization Theorem* by Brenier (1991), is then exploited to define fully general concepts of ranks and signs—called the *Monge-Kantorovich ranks and signs* coinciding, in the univariate and elliptical settings, with the traditional concepts, and enjoying under completely unspecified (absolutely continuous) *d*-dimensional distributions, the essential properties that make traditional rank-based inference an essential part of the semiparametric inference. (Joint work with Victor Chernozhukov, Alfred Galichon, and Marc Henry)

Ogata, H.

A circular autocorrelation of stationary circular Markov processes

Abstract: The stationary Markov process is considered and its circular autocorrelation function is investigated. More specifically, a transition density of the stationary Markov circular process is defined by two circular distributions, and we elucidate structure of the circular autocorrelation when the one distribution is uniform and the other is arbitrary. The asymptotic properties of parametric and nonparametric estimators of the circular autocorrelation function are derived. Some numerical studies, simulation results and application to the real data will be given. (Joint work with T. Abe, T. Shiohama and H. Taniai)

Lin G.D.

Recent Developments on the Moment Problem

Abstract: We consider univariate distributions with finite moments of all positive

orders. The moment problem is to determine whether or not a distribution is uniquely determined by the sequence of its moments. We first give a brief survey on this classical moment problem and then focus on the recent developments on the checkable criteria including Hardy's condition, Krein's condition and the rate of moment growth, which help us solve the problem more easily. Both Hamburger and Stieltjes cases are investigated. The former is concerned with distributions on the whole real line, while the latter deals only with distributions on the right half-line.

Keywords: Hamburger moment problem, Stieltjes moment problem, Cramer's condition, Carleman's condition, Krein's condition, Hardy's condition.

Giraitis, L.

Testing for stability of the mean of heteroskedastic time series

Abstract: Time series models are often fitted to the data without preliminary checks for stability of the mean and variance, conditions that may not hold in much economic and financial data, particularly over long periods. Ignoring such shifts may result in fitting models with spurious dynamics that lead to unsupported and controversial conclusions about time dependence, causality, and the effects of unanticipated shocks. In spite of what may seem as obvious differences between a time series of independent variates with changing variance and a stationary conditionally heteroskedastic (GARCH) process, such processes may be hard to distinguish in applied work using basic time series diagnostic tools. We develop and study some practical and easily implemented statistical procedures to test the mean and variance stability of uncorrelated and serially dependent time series. Application of the new methods to analyze the volatility properties of stock market returns leads to some unexpected surprising findings concerning the advantages of modeling time varying changes in unconditional variance.