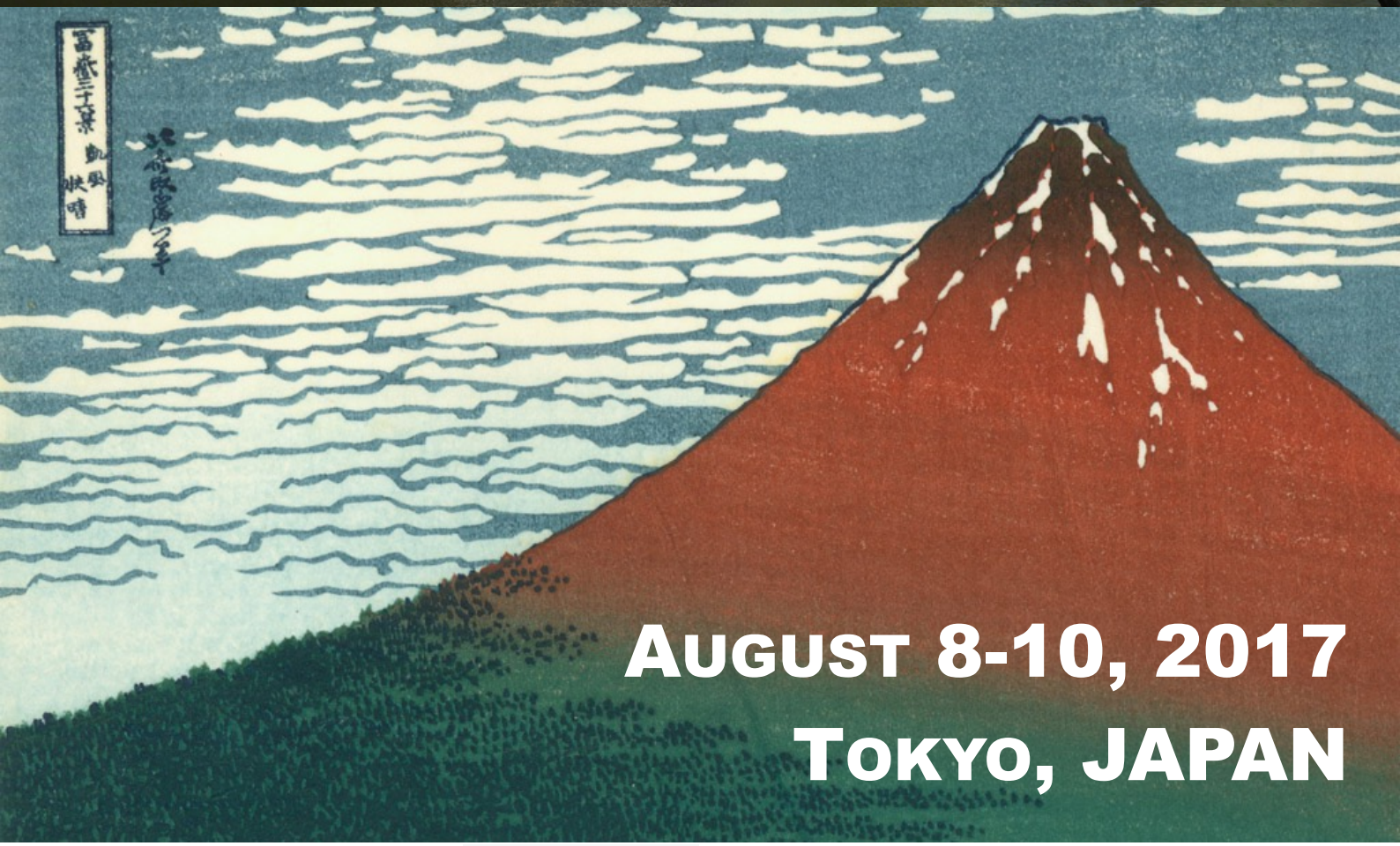
IFCS-2017

Conference Program And Book Of Abstracts

Conference of the International Federation of Classification Societies

The Challenge of Data Science in the Era of Big Data



August 8-10, 2017

Tokyo, JAPAN

Japanese Classification Society

In cooperation with Japan National Tourism Organization and The Institute of Statistical Mathematics

IFCS-2017
Conference Program and Book of Abstracts

Conference
of the International Federation
of Classification Societies

The Challenge of Data Science in the Era of Big Data

August 8-10, 2017
Tokyo, JAPAN

IFCS-2017 Organizing Team
Japanese Classification Society
Tokai University
International Federation of Classification Societies

Table of Contents

Sponsorship

Host

IFCS-2017 Organizing Team
Japanese Classification Society
Tokai University
International Federation of Classification Societies

Cosponsor

Institute of Statistical Mathematics

Cooperating Organization

Japan National Tourism Organization

Supporters' Organization

Japan Association for Consumer Studies
Japan Institute of Marketing Science
Japan Society for Fuzzy Theory and Intelligent Informatics
Japan Society of Kansei Engineering
Japan Marketing Research Association
Japanese Association for Mathematical Sociology
Japanese Association of Industrial / Organizational Psychology
Japanese Federation of Statistical Science Associations
Japanese Society of Applied Statistics
Japanese Society of Computational Statistics
The Behaviormetric Society
The Biometric Society of Japan
The Japan Statistical Society
The Japan Association for Research on Testing
The Japanese Psychological Association
The Japanese Society for Quality Control
The Operations Research Society of Japan

Organizing Team

Local Organizer

Tadashi Imaizumi

Scientific Program Committee

Tadashi Imaizumi, SPC Chair
Sadaaki Miyamoto, SPC Vice Chair, JCS
Yoshiro Yamamoto, LOC Chair
Akinori Okada, IFCS President
Maurizio Vichi, IFCS Past-President
Berthold Lausen, IFCS President-Elect
Christian Hennig, IFCS Secretary
Paul McNicholas, IFCS Publication Officer
Nema Dean, IFCS Treasurer
Vladimir Batagelj, SSS
Theodoros Chadjipantelis, GDSA
Sonya Coleman, IPRCS
José Gonclaves Dias, CLAD
Carlos Cuevas Covarrubias, SoCCCAD
Brian C. Franczak, CS
Salvatore Ingrassia, CLADAG
Hans Kestler, GfKl
Éva Laczka, HSA-CMSG
Sugnet Lubbe, SASA-MDAG
Fionn Murtagh, BCS
Mohamed Nadif, SFC
Heon Jin Park, KCS
Józef Pociecha, SKAD
Abderrahmane Sbihi, MCS
José Fernando Vera, SEIO-AMyC
Jeroen K. Vermunt, VOC
Patrick Groenen, IASC Delegate

Local Organizing Committee

Yoshiro Yamamoto, LOC Chair
Takafumi Kubota
Koji Kurihara
Masahiro Mizuta
Miki Nakai
Junji Nakano
Atsuho Nakayama
Makiko Oda
Takuya Ohmori
Kosuke Okusa
Fumitake Sakaori
Kumiko Shiina
Akinobu Takeuchi
Makoto Tomita
Yuki Toyoda
Hiroshi Yadohisa
Satoru Yokoyama
Christian Hennig (IFCS Secretary)

Secretariat

Fumitake Sakaori, Secretary-General
Satoru Yokoyama, Assistant Secretary-General
Makiko Oda, General Affairs

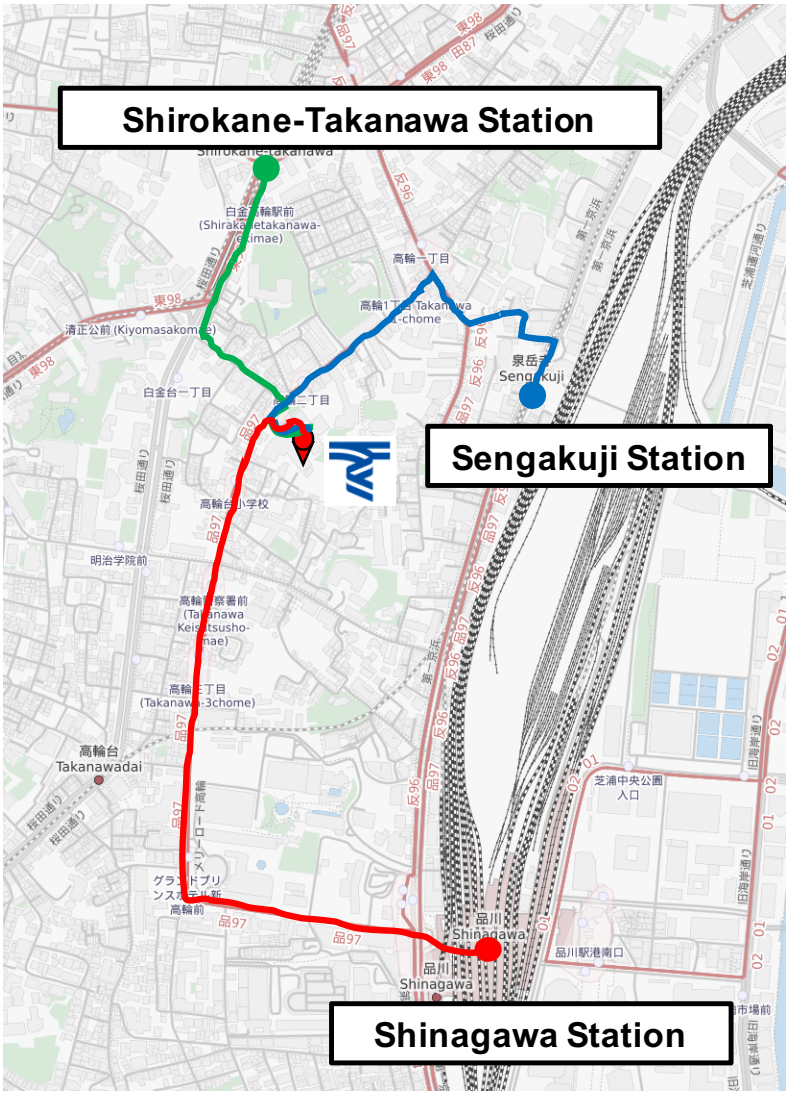
Conference Venue

The conference will be held on Takanawa campus of Tokai University. The campus is located in the commercial heart of Tokyo. In Takanawa, there is a famous Soto Zen Buddhist temple Sengaku-ji, where the famous Forty-seven Samurai and their lord are enshrined.

Takanawa is also located near Shinagawa, one of transport hubs in Tokyo in public transportation. It is very easy to come to downtown in Tokyo by train and subway from Shinagawa, as well as to Osaka and Kyoto by Shinkansen (high speed train). There are direct connections from both Narita and Haneda International Airports to Shinagawa.

From Shinagawa Station

About 15 minutes' walk (red route in the map right). Take the Toei bus for Meguro Station. Get off at "Takanawa Keisatsusho Mae" and then walk about 3 minutes from this bus stop.



From Sengakuji Station on the Toei Subway Asakusa Line

About 10 minutes' walk (blue route).

From Shirokane-Takanawa Station on the Tokyo-Metro Nanboku Line and Toei Subway Mita Line

About 8 minutes' walk (green route).

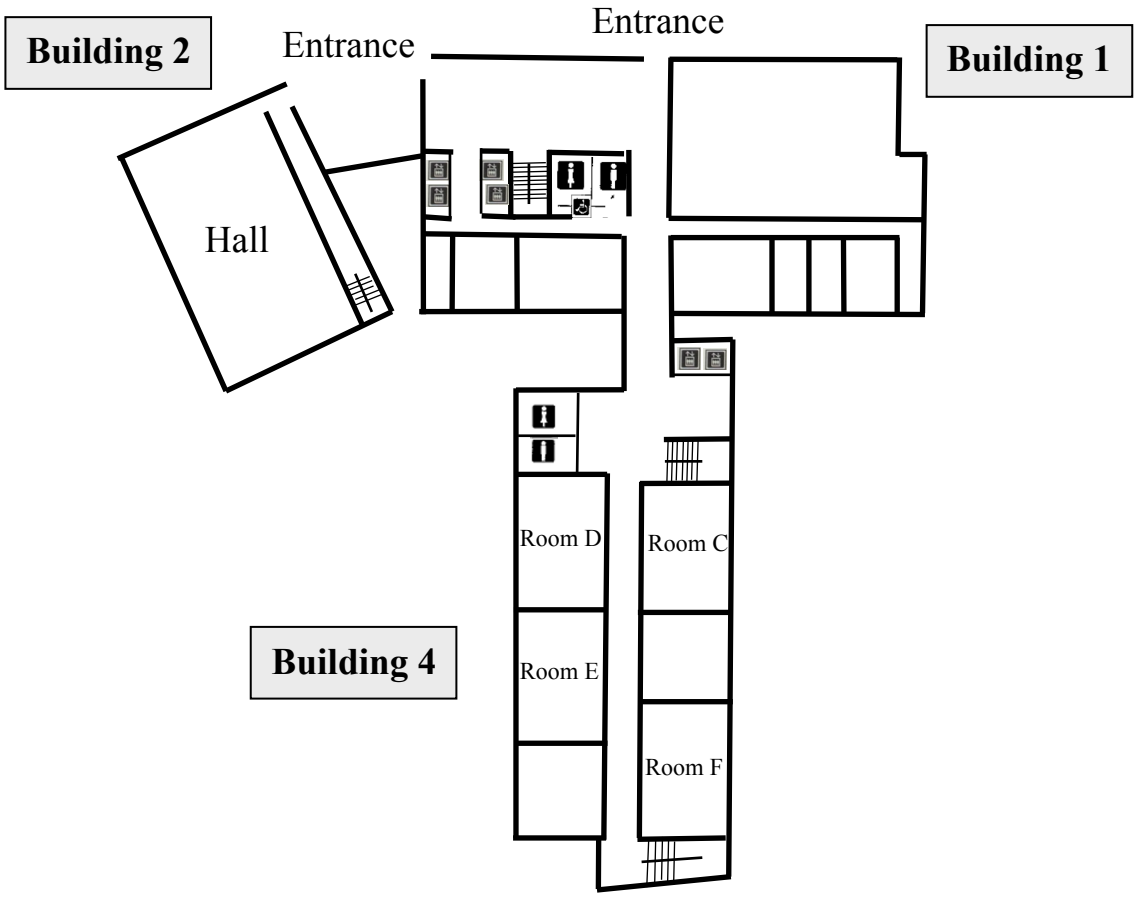
Conference Site

The conference will be held at the Building 1, 2 and 4 of Tokai University Takanawa campus. These buildings are connected so you can move without going outside. Specifically,

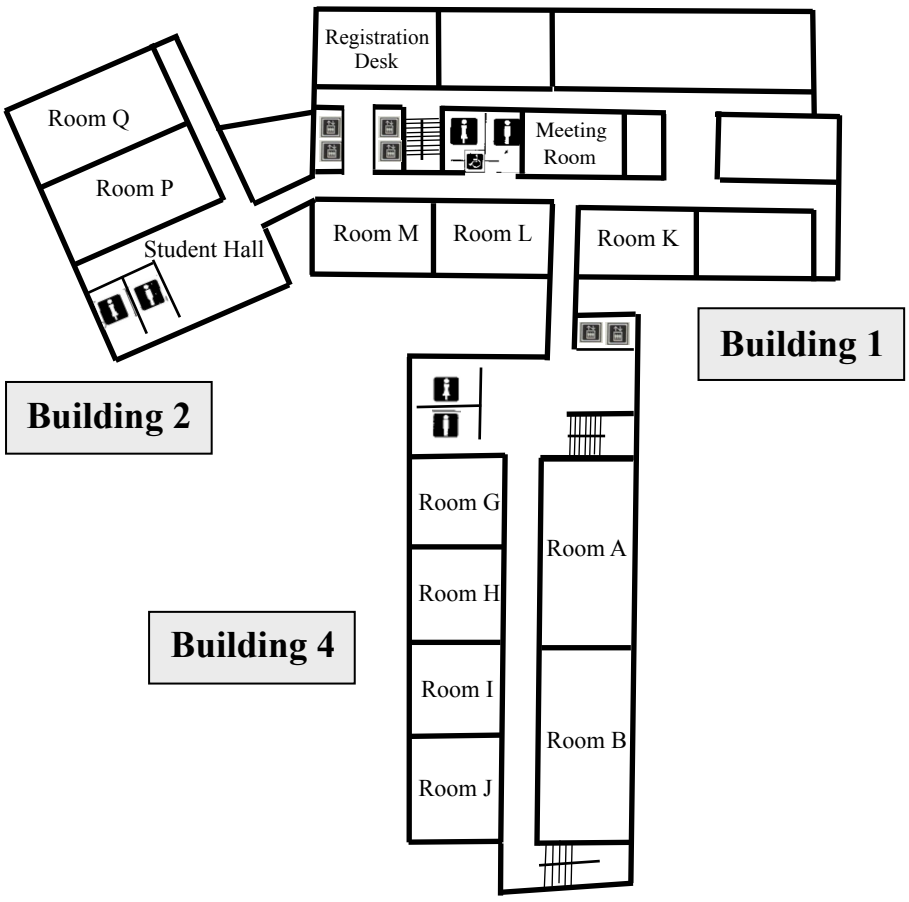
- Opening Ceremony, Plenary Invited sessions, Presidential Address, Award Session and Closing Ceremony will be held at the Hall, 1st floor of Building 2.
- Special sessions and Contributed sessions will be held at room A to L in the Building 1 or 2.
- Poster session will be held at the Student Hall, on the 2nd floor of Building 2.
- The coffee breaks will be served at the Student Hall and room P, on the 2nd floor of Building 2.
- Software demonstration and Book sales will be at room M. You can take a break in this room.
- You can take lunch at the Dining hall, on the 1st basement floor of Building 4. Please purchase the meal ticket at the entrance of Dining hall on the day.

Site Map

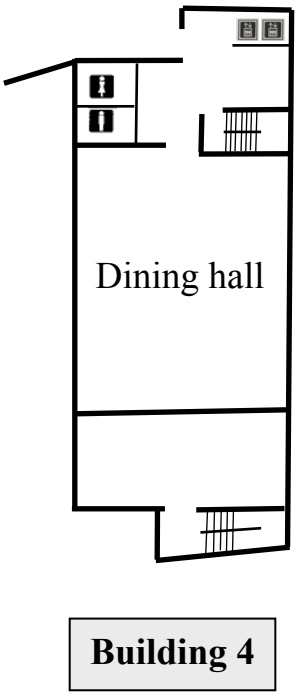
1st Floor



2nd Floor



1st Basement Floor



Social Program

Welcome Reception

The conference welcome reception will be held on August 8th (17:30) at the Dining Hall in the conference site (1st basement floor of Building 4, see the map at page 5). It is free of charge for the participants and accompanying persons.

Conference Dinner (Toll event)

The conference dinner will be held on August 9th at Meiji Kinenkan. Meiji Jingu's banquet hall Meiji Kinenkan is a green oasis in the heart of Tokyo located near Omotesando, Aoyama, and Akasaka in a separate part of the Meiji Jingu Gaien, which is the outer gardens of the Meiji Shrine. The serene garden is suitable for Shinto weddings and banquets.

It was the first wedding hall which offers a comprehensive range of services in Japan. The main hall is a place of historical interest since the drafts of the former Imperial Constitution and the Imperial House Act were discussed in the presence of the Meiji Emperor.

In the conference dinner, we will have a small concert of “Hana-Temari”, the sister duo with flute and koto, Japanese harp. You can’t help but be enchanted by the sound of Hana-Temari from the moment you listen to them because of its beautiful mixture combined with Japanese and European traditional music playing Japanese koto, and European flute.

Optional Tour to Asakusa

We will have an optional small tour to Asakusa before the conference dinner. Asakusa is the center of Tokyo's Shitamachi area, where the full atmosphere of an old town remains. Main attraction is a very popular Buddhist temple Sensoji built in the 7th century. The temple is approached via the Nakamise, a shopping street providing a variety of traditional snacks and tourist souvenirs.

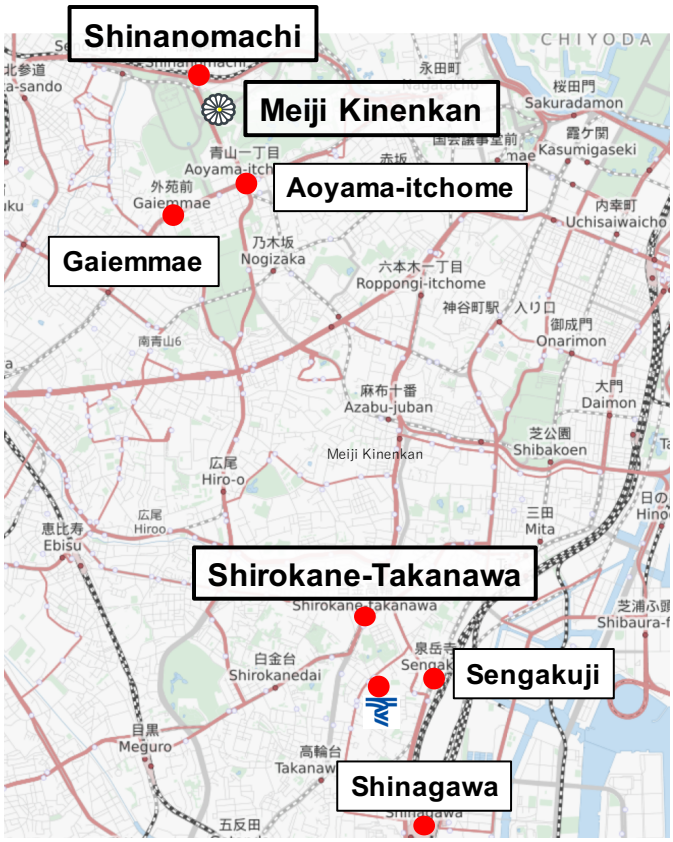
The tour buses will leave at 15:15 at the conference venue. After the end of the dinner, we will serve two buses to Shinagawa station for tour participants.

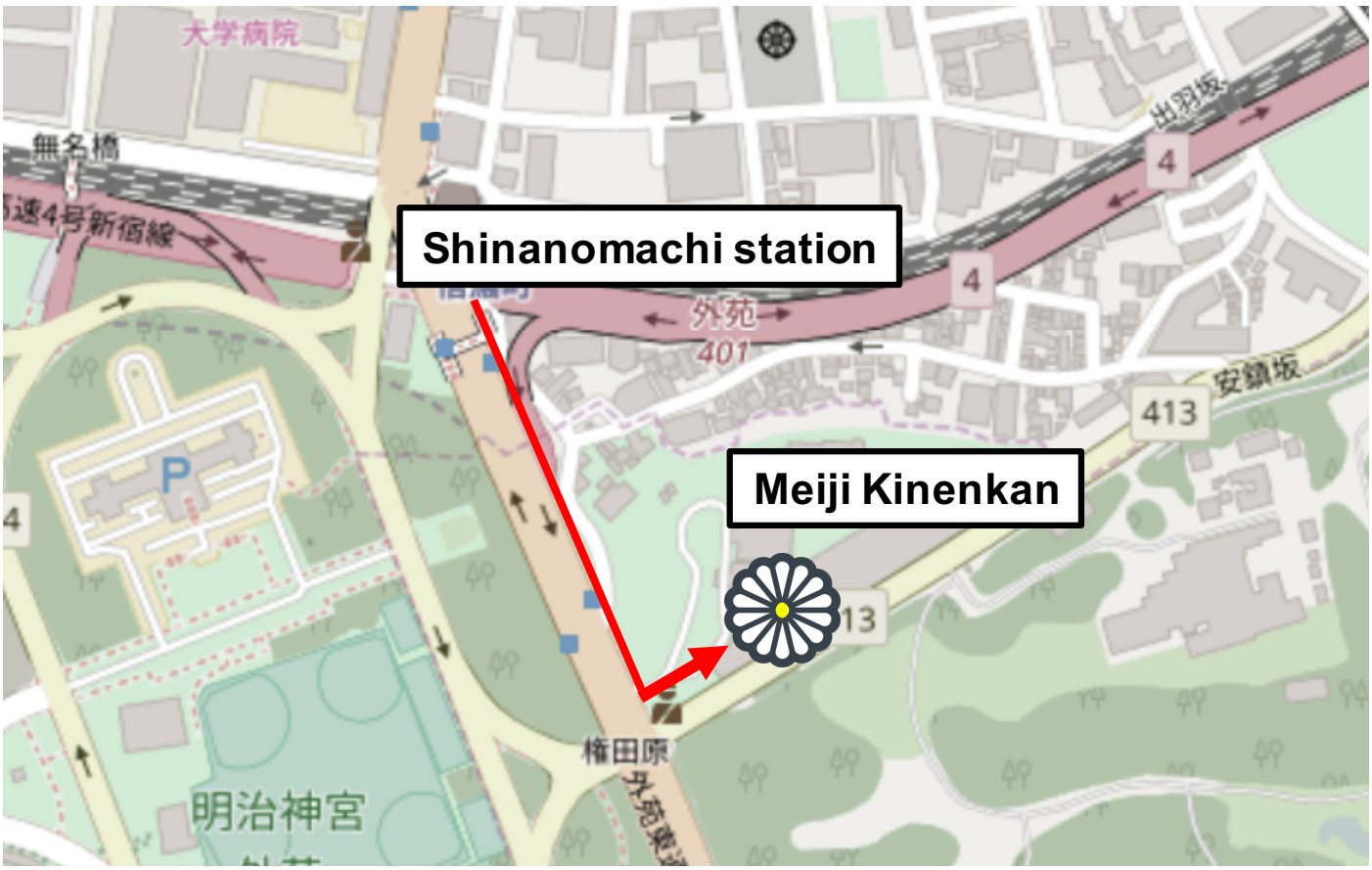
Direct Transportation to Meiji Kinenkan

Dinner participants who do not attend the small tour should go to the Meiji Kinenkan by public transportation. Meiji Kinenkan is located six kilometers to the north of the conference venue.

From the conference venue, the following route is recommended:

1. Takanawa campus of Tokai University
   — Shirokane-Takanawa Station
   (8 minutes on foot)
2. Shirokane-Takanawa Station
   — Yotsuya Station
   (10 minutes by the subway Nanboku Line)
3. Yotsuya Station
   — Shinanomachi Station
   (2 minutes by the JR Sobu Line)
4. Shinanomachi Station
   — Meiji Kinenkan (3 minutes on foot)
   For the walking direction, see the map below.





Also you can go to Meiji Kinenkan by 6 minutes walk from Aoyama-itchome station on the Ginza, Hanzomon, Oedo subway line, and Kokuritsu-kyogijo Station on the Oedo subway line.

Time	Room A	Room B	Room C	Room D	Room E	Room F	Room G	Room H	Room I	Room J	Room K	Room L
from 8:30	**Start Registration**											
9:00 - 9:30	**Opening Ceremony**											
	Hall, 1F Building 2											
9:30 - 10:15	**Plenary Invited 01** Chair: *Sadaaki Miyamoto*											
	Smart simulation and smart experimental design, *Tomoyuki Higuchi*											
	Hall, 1F Building 2											
10:15 - 10:45	**Coffee Break (**Student Hall, 2F Building 2)											
10:45 - 12:25	**SP01**	**SP02**	**SP03**	**SP04**	**SP05**	**SP06**	**SP07**	**SP08**	**SP09**	**SP10**	**CN01**	**CN02**
	Advances in latent variable modeling	**Challenges in visualisation and classification**	**Changing environment, new challenges, new responses**	**Clustering with mixture Models**	**Social implementation based on analytic results by latent classification methods for health management**	**Analysis of micro official statistics**	**Cluster validation: New developments and issues I**	**The cutting edge of biomedical big data - From bioinformatic methodologies to data analysis**	**Judgment and decision making**	**Functional data analysis**	**Contributed session 01**	**Contributed session 02**
	Organized by Salvatore Ingrassia - Chair: *Angela Montanar*	Chair: *Sugnet Lubbe*	Chair: *Éva Laczka*	Chair: *Paul McNicholas*	Organzied by Tetsuro Ogi and Michiko Watanabe - Chair: *Tetsuro Ogi*	Chair: *Yasumasa Baba*	Chair: *Christian Hennig*	Chair: *Yusuke Matsui, Masahiro Nakatochi , Yuichi Shiraishi*	Chair: *Kazuhisa Takemura*	Chair: *Yuko Araki*	Chair: *Atsuho Nakayama*	Chair: *Maura Mezzetti*
12:25 - 13:15	**Lunch Break**											
13:15 - 14:00	**Plenary Invited 02** Chair: *Józef Pociecha*											
	Mixture modelling of multivariate and / or longitudinal data: Arriving at insightful representations, *Marieke E. Timmerman*											
	Hall, 1F Building 2											

Time	Room A	Room B	Room C	Room D	Room E	Room F	Room G	Room H	Room I	Room J	Room K	Room L
14:00 - 14:10	**Short Break**											
14:10 - 15:30	**SP11**	**SP12**	**SP13**	**SP14**	**SP15**	**SP16**	**SP17**	**SP18**	**SP19**	**SP20**	**CN03**	**CN04**
	Cluster analysis and quantification	**Leading-edge research of clustering and its applications I**	**Classification models in finance and business I**	**Data mining and data visualization I**	**Marketing science I**	**Cluster Analysis of target data set in cluster benchmark data repository I**	**Enumeration algorithm and data science**	**Advances in sparse regression, dimension reduction and related computations**	**Recent problems of big data handling in official statistics**	**Causal inference and related topics**	**Contributed session 03**	**Contributed session 04**
	Orgainized by Shizuhiko Nishisato - Chair: *Pieter Kroonenberg*	Chair: *Yasunori Endo and Sadaaki Miyamoto*	Chair: *Jozef Pociecha*	Chair: *Yoshiro Yamamoto*	Chair: *Takeshi Moriguchi*	Chair: *Iven Van Mechelen*	Chair: *Masahiro Mizuta and Shin-ichi Minato*	Chair: *Yuichi Mori*	Chair: *Hiroe Tsubaki*	Chair: *Yutaka Kano*	Chair: *Kei Kurakawa*	Chair: *Mark De Rooij*
15:30 - 16:00	**Coffee Break (**Student Hall, 2F Building 2)											
16:00 - 17:20	**SP21**	**SP22**	**SP23**	**SP24**	**SP25**	**SP26**	**SP27**	**SP28**	**SP29**	**SP30**	**CN05**	**CN06**
	Perspectives of contingency table analysis	**Leading-edge research of clustering and its applications II**	**Classification models in finance and business II**	**Data mining and data visualization II**	**Marketing science II**	**Cluster Analysis of target data set in cluster benchmark data repository II**	**Bayesian inference and model selection**	**Robustness and active learning with logistic models**	**Statistical issues on clustering and classification in medical data analysis**	**Regularization methods**	**Contributed session 05**	**Contributed session 06**
	Organized by Shizuhiko Nishisato - Chair: *Eric Beh and Michel van de Velden*	Chair: *Yasunori Endo and Sadaaki Miyamoto*	Chair: *Jozef Pociecha*	Chair: *Koji Kurihara*	Chair: *Takeshi Moriguchi*	Chair: *Iven Van Mechelen*	Chair: *Yutaka Kano*	Chair: *Yuan-chin Ivan Chang*	Chair: *Taerim Lee*	Chair: *Patrick J.F. Groenen*	Chair: *Takashi Murakami*	Chair: *M Miftahuddin*
17:30 - 19:30	**Welcome Reception (**Dining Hall, B1F Building 4)											
18:00 - 20:00	**IFCS Council Meeting** (Meeting room, 2F Building 1)											

Time	Room A	Room B	Room C	Room D	Room E	Room F	Room G	Room H	Room I
9:00 - 9:45	**Plenary Invited 03** Chair: *Paul McNicholas*								
	An overview of hybrid latent variable models and how they can be used to address outliers, flexible distributions of latent variables and issues of data dimensionality, *Irini Moustaki*								
	Hall, 1F Building 2								
9:45 - 9:55	**Short Break**								
9:55 - 10:55	**SP31**	**SP32**	**SP33**	**SP34**	**SP35**	**SP36**	**CN07**	**CN08**	**CN09**
	Bayesian approach to classification	**Classification and representation for non metric and symbolic Data I**	**New topics related to classification problems I**	**Clustering and recommendation systems**	**BRCA data analysis**	**Issues in classification with complex data structures (CLADAG)**	**Contributed session 07**	**Contributed session 08**	**Contributed session 09**
	Chair: *Kazuo Shigemasu*	Chair: *J. Fernando Vera and Eva Boj del Val*	Chair: *Yoshikazu Terada, Michio Yamamoto and Hiroshi Yadohisa*	Chair: *Jean Diatta*	Organized by H. H. Bock - Chair: *Hans A. Kestler*	Chair: *Francesco Palumbo*	Chair: *Kunihiro Kimura*	Chair: *Daniel Baier*	Chair: *Yasumasa Baba*
10:55 - 11:25	**Coffee Break (**Student Hall, 2F Building 2)								
11:25 - 12:25	**SP37**	**SP38**	**SP39**	**SP40**	**SP41**	**CN10**	**CN11**	**CN12**	**CN13**
	Classification of (un)complex data (CLAD)	**Classification and representation for non metric and symbolic Data II**	**New topics related to classification problems II**	**Clustering and classification in innovative data science**	**Cluster analysis using non-gaussian mixtures**	**Contributed session 10**	**Contributed session 11**	**Contributed session 12**	**Contributed session 13**
	Chair: *José G. Dias*	Chair: *J. Fernando Vera and Eva Boj del Val*	Chair: *Yoshikazu Terada, Michio Yamamoto and Hiroshi Yadohisa*	Chair: *Fionn Murtagh*	Chair: *Brian Franczak*	Chair: *Łukasz Smaga*	Chair: *Shinobu Tatsunami*	Chair: *Seong-Keon Lee*	Chair: *Tsai-Hung Fan*
12:25 - 13:25	**Lunch Break**								
13:25 - 14:10	**Presidential Address** Chair: *Berthold Lausen*								
	Dissimilarity based on manner of dissimilarity/similarity to/from the others, *Akinori Okada*								
	Hall, 1F Building 2								
14:10 - 15:00	**Award Session** Chair: *Yasumasa Baba*								
	Hall, 1F Building 2								
15:15 - 21:00	**Short Tour & Conference Dinner**								

Thursday, August 10, 2017

Time	Room A	Room B	Room C	Room D	Room E	Room F	Room G	Room H	Room I	Room J
9:00 - 9:45	**Plenary Invited 04** Chair: *Heon Jin Park*									
	Classification by quantiles, *Cinzia Viroli*									
	Hall, 1F Building 2									
9:45 - 10:30	**Plenary Invited 05** Chair: *Yoshiro Yamamoto*									
	Decisions that are needed when using cluster analysis, and research that helps with making them, *Christian Hennig*									
	Hall, 1F Building 2									
10:30 - 11:00	**Coffee Break** (Student Hall, 2F Building 2)									
11:00 - 11:45	**Poster Session**									
	Student Hall, 2F Building 2									
11:45 - 12:35	**Lunch Break**									
12:35 - 14:15	**SP42**	**SP43**	**SP44**	**SP45**	**SP46**	**SP47**	**SP48**	**CN14**	**CN15**	**CN16**
	Advanced technique for analyzing (big) multi-set/multi-subject data	**Survey data analysis**	**Analysis and clustering of complicated data**	**Text classification**	**Data Science, classification and clustering**	**Methods of data analysis and statistical measures in the social sciences**	**Cluster validation: New developments and issues II**	**Contributed session 14**	**Contributed session 15**	**Contributed session 16**
	Chair: *Tom Wilderjans*	Chair: *Ryozo Yoshino*	Chair: *Junji Nakano*	Organized by Masakatsu Murakami - Chair: *Yuichiro Kobayashi*	Chair: *Berthold Lausen*	Chair: *Theodore Chadjipantelis*	Organized by Christian Hennig - Chair: *Hans-Joachim Mucha*	Chair: *Masahiro Mizuta*	Chair: *Noboru Wakayama*	Chair: *Guangliang Chen*
14:15 - 14:25	**Short Break**									
14:25 - 15:15	**Closing Ceremony**									
	Hall, 1F Building 2									

Monday, August 7, 2017

13:30 - 18:00	**Start Registration**

Tuesday, August 8, 2017

8/8 from 8:30	**Start Registration**
8/8 9:00 - 9:30	**Opening Ceremony** (Hall, 1F Building 2)
8/8 9:30 - 10:15	**Plenary Invited 01** Chair: *Sadaaki Miyamoto* (Hall, 1F Building 2)
	Smart simulation and smart experimental design, Tomoyuki Higuchi.
8/8 10:15 - 10:45	Coffee Break (Student Hall, 2F Building 2)
8/8 10:45 - 12:25	**SP01: Advances in latent variable modeling** - Organized by Salvatore Ingrassia - Chair: *Angela Montanar* (Room A)
	Latent markov factor analysis for exploring within-subject measurement model differences in experience sampling studies, Leonie V.D.E. Vogelsmeier, Jeroen K. Vermunt, and Kim De Roover.
	When factorial invariance fails: a new rotation approach for multigroup exploratory factor analysis to identify loading-specific differences, Kim De Roover and Jeroen K. Vermunt.
	Partial homogeneity models for temporal repeated cross-sectional latent class analysis, Brian Francis and Valmira Hoti.
	Hidden markov models for the analysis of longitudinal data through the LMest package, Francesco Bartolucci, Silvia Pandolfi and, Fulvia Pennoni.
	Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis, Vincent Vandewalle.
8/8 10:45 - 12:25	**SP02: Challenges in visualisation and classification** - Chair: *Sugnet Lubbe* (Room B)
	Determinants of survival risk factors for HIV/AIDS patients on ART in a developing country - accounting for clustering at facility level, Renette Julia Blignaut and Innocent Maposa.
	Challenges in visualising and imputing missing categorical data, Johané Nienkemper-Swanepoel, Sugnet Lubbe, and Niël le Roux.
	A visualization technique to identify critical variables in multivariate process monitoring, Niël Le Roux.
	Using the NSGA optimization algorithm for variable selection in spectral data classification: some lessons learnt, Martin Kidd and Martin Philip Kidd.
	Biplots based on principal surfaces, Raeesa Ganey.
8/8 10:45 - 12:25	**SP03: Changing environment, new challenges, new responses** - Chair: *Éva Laczka* (Room C)
	Estimation methods based on weighted cluster analysis, Roland Szilágyi, Beatrix Varga, and Renáta Géczi Papp.
	Using classification of regions based on the complexity of the global progress indices for supporting sustainable development, Anita Csesznák and Eva Sandor-Kriszt.
	Possibilities in welfare measures, Mónika Galambosné Tiszberger.
	A measuring and modelling way of multicollinearity with Petres' Red indicator, Peter Kovacs and Tibor Petres.
	Discussion, Masahiro Mizuta.

8/8 10:45 - 12:25 **SP04: Clustering with mixture Models** - Chair: *Paul McNicholas* (Room D)

Mixtures of common t-factor analyzers for high-dimensional data with missing information, Wan-Lun Wang.

Mixtures of skew-t factor analyzers with common factor loadings, Tsung-I Lin.

Clustering of multivariate categorical data with dimension reduction via nonconvex penalized likelihood maximization, Michio Yamamoto.

Subspace clustering with the multivariate-t distribution, Brian C Franczak.

Fractionally-supervised classification using the weighted likelihood, Irene Vrbik.

8/8 10:45 - 12:25 **SP05: Social implementation based on analytic results by latent classificatioin methods for health management -** Organzied by Tetsuro Ogi and Michiko Watanabe - Chair : *Tetsuro Ogi* (Room E)

Health management using digital signage and activity meter, Kenichiro Ito, Yusuke Kurita, and Tetsuro Ogi.

The effect of an immersive analytical tool on the exploratory analysis of big data, Iwane Maida, Kenichiro Ito, So Sato, Shunichi Nomura, Tetsuya Toma, and Tetsuro Ogi.

Classification of in-week and -day patterns in ambulatory activity and body composition change, Shunichi Nomura, Michiko Watanabe, and Yuko Oguma.

Attempt to set a step count target by applying latent class analysis, Kotaro Ohashi, Michiko Watanabe, and Yuko Oguma.

8/8 10:45 - 12:25 **SP06: Analysis of micro official statistics** - Chair: *Yasumasa Baba* (Room F)

Analysis of expenditure patters of virtual marriage households consisting of working couples synthesized by statistical matching method, Mikio Suga and Yasuo Nakatani.

Visualization and spatial statistical analysis for Vietnam household living standard survey, Takafumi Kubota.

The effects of natural disasters on household income and poverty in rural Vietnam : An analysis using Vietnam household living standards survey, Rui Takahashi.

The influence of household assets on choice to work, Shinsuke Ito.

The economic structure of rural areas in the mekong region countries: a comparative analysis of micro official statistics in Cambodia, Thailand, and Vietnam, Daisuke Sakata and Motoi Okamoto.

8/8 10:45 - 12:25 **SP07: Cluster validation: New developments and issues I** - Chair: *Christian Hennig* (Room G)

Non-parametric methods for clusters: Estimation of the number of clusters and a comparison test, Andre Fujita.

On model-based clustering under measurement uncertainty, Volodymyr Melnykov.

Cluster validation in multicriterion data clustering, Julia Handl and Emiao Lu.

Assessing model-based clustering methods with cytometry data sets, Gilles Celeux and Jean-Patrick Baudry.

8/8 10:45 - 12:25 **SP08: The cutting edge of biomedical big data - From bioinformatic methodologies to data analysis** - Chair: *Yusuke Matsui, Masahiro Nakatochi ,Yuichi Shiraishi* (Room H)

Classification of tree-valued data and its application to cancer evolutionary trees, Yusuke Matsui, Satoru Miyano, and Teppei Shimamura.

Constructing cancer characteristic-specific gene regulatory networks via L1-type regularized regression modeling, Heewon Park, Teppei Shimamura, Seiya Imoto, and Satoru Miyano.

Clinical classification of cancer subtypes based on gene mutations and expressions, Shuta Tomida.

Association analysis among myocardial infarction, cardiovascular disease-related single nucleotide polymorphisms, and DNA methylation sites utilizing the cluster analysis, Masahiro Nakatochi, Sahoko Ichihara, Ken Yamamoto, Tatsuaki Matsubara, and Mitsuhiro Yokota.

Latent probabilistic modeling for mutational signature in cancer genomes, Yuichi Shiraishi.

8/8 10:45 - 12:25 **SP09: Judgment and decision making** - Chair: *Kazuhisa Takemura* (Room I)

Classification and visualization of eye movement data in judgment and decision making studies, Masahiro Morii.

Image processing methods for gaze pattern analysis in marketing research, Jasmin Kajopoulos, Hajime Murakami, Keita Kawasugi, Marin Aikawa, and Kazuhisa Takemura.

Probability weighting function and time discounting function in decision making: Theory and experimental analysis, Kazuhisa Takemura and Hajime Murakami.

A utility model in intertemporal choice that is yielded by introducing psychological time duration, Yutaka Matsushita.

Eye movement analysis using image processing methods: for gaze data during decision making between a pair of product images, Hajime Murakami, Keita Kawasugi, Jasmin Kajopoulos, Marin Aikawa, Tokihiro Ogawa, and Kazuhisa Takemura.

8/8 10:45 - 12:25 **SP10: Functional data analysis** - Chair: *Yuko Araki* (Room J)

Selection of variables and decision boundaries in functional logistic regression model, Hidetoshi Matsui.

Variable selection for classification of multivariate functional data, Tomasz Górecki and Waldemar Wołyński.

Identifying changes in mean functions for a sequence of functional data, Jeng-Min Chiou.

Semi-supervised learning for functional data, Yoshikazu Terada.

8/8 10:45 - 12:25 **CN01: Contributed session 01** - Chair: *Atsuho Nakayama* (Room K)

Hybrid feature selection method for high-dimensional data sets, Asma Gul, Zardad Khan, Werner Adler, and Berthold Lausen.

Clustering by moving centroids using optimization heuristics, Mario A Villalobos-Arias, Juan José Fallas, and Jeffry Chavarria.

Non-hierarchical clustering for large data without recalculating cluster center, Atsuho Nakayama and Shinji Deguchi.

Supervised nested algorithm for classification based on k-means, Luciano Nieddu and Donatella Vicari.

Morphological characterization of 3D shape with curvature flow-based spin transformation and spherical harmonics decomposition, Ryo Yamada, Fujii Yosuke, Kohei Suzuki, Ayako Iwasaki, Takuya Okada, and Kazushi Mimura.

8/8 10:45 - 12:25 **CN02: Contributed session 02** - Chair: *Maura Mezzetti* (Room L)

Combining individual and aggregated data to investigate the role of socio-economic disparities on cancer burden in Italy, Maura Mezzetti and Francesca Dominici.

Classification of Japanese graduate schools: in terms of educational practices and the grown globalization competencies by the policies, Taiyo Utsuhara, Masaki Uto, Asana Ishihara, Atsushi Yoshikawa and Maomi Ueno.

Explanatory research of fashion behavior by bayesian network analysis, Keiko Yamaguchi and Hiroshi Kumakura.

A statistical model for supporting AirB&B hosts activity, Giulia Contu, Luca Frigau, and Claudio Conversano.

Information extraction and its visualization using the quantification method from Tweet information for a disaster, Takamitsu Funayama, Yoshiro Yamamoto, and Osamu Uchida.

8/8 12:25 - 13:15 Lunch Break

8/8 13:15 - 14:00 **Plenary Invited 02** Chair: *Józef Pociecha* (Hall, 1F Building 2)

Mixture modelling of multivariate and / or longitudinal data: Arriving at insightful representations, Marieke E. Timmerman.

8/8 14:00 - 14:10 Short Break

8/8 14:10 - 15:30 **SP11: Cluster analysis and quantification** - Organized by Shizuhiko Nishisato - Chair: *Pieter Kroonenberg* (Room A)

Bi-modal clustering and quantification theory: Flexible-filter clustering of contingency and response-pattern tables. Part 1. Historical overview, Shizuhiko Nishisato and Jose Garcia Clavel.

Bi-modal clustering and quantification theory: Flexible-filter clustering of contingency and response-pattern tables. Numerical demonstration of a proposed framework, Jose G Clavel and Shizuhiko Nishisato.

Clustering methods for preference data in the presence of response styles, Mariko Takagishi, Michel van de Velden, and Hiroshi Yadohisa.

Calculating neural reliability from EEG recordings of naturalistic stimuli, Pieter C. Schoonees and Niël J Le Roux.

8/8 14:10 - 15:30 **SP12: Leading-edge research of clustering and its applications I** - Chair: *Yasunori Endo and Sadaaki Miyamoto* (Room B)

Generative learning with emergent self-organizing neuronal networks, Alfred Ultsch.

Projection based clustering, Michael Christoph Thrun.

Agglomerative hierarchical clustering with automatic selection of the number of clusters using AIC and BIC, Sadaaki Miyamoto.

A convex analysis interpretation of fuzzy c-means, Yoshifumi Kusunoki.

8/8 14:10 - 15:30 **SP13: Classification models in finance and business I** - Chair: *Jozef Pociecha* (Room C)

Clustering sum of type loans using fuzzy c-means algorithm method, Khoiru Nissa Apriliya and Hanif Novrandhita.

Granular credit data classification with SVM based approaches, Ralf Werner Stecking.

ABC analysis in corporate bankruptcy prediction, Barbara Pawełek, Józef Pociecha, and Mateusz Baryła.

Benchmarking classification of stock performance by corporate performance measures - Insights from different modelling techniques, Karsten Luebke, Roland Wolf, and Sebastian Sauer.

8/8 14:10 - 15:30 **SP14: Data mining and data visualization I** - Chair: *Yoshiro Yamamoto* (Room D)

Classification and visualization by self-organizing maps and hierarchical cluster analysis, Yo Kameoka and Yoshiro Yamamoto.

Association rule analysis, and comparison of visualization by the quantification technique, Sanetoshi Yamada and Yoshiro Yamamoto.

Visualizing citation networks for assessment of research performance in academic institutions, Tomokazu Fujino, Keisuke Honda, Hiroka Hamada, and Yoshiro Yamamoto.

Credit scorecard development with support vector machines, Seohoon Jin and Marcel Eduard MBARGA.

8/8 14:10 - 15:30 **SP15: Marketing science I** - Chair: *Takeshi Moriguchi* (Room E)

Analyze the health consciousness of yogurt buyers, Saya Yamada and Yumi Asahi.

Compare of Taiwan import Japanese fruits and vegetables, Yukiya Suzuki and Yumi Asahi.

An exploratory study on the clumpiness measure of inter-transaction times: How is it useful for customer relationship management?, Yuji Nakayama and Nagateru Araki.

Proposal of the method to diagnose brand image and ability of a respondent by using reaction time and hierarchy model, Masao Ueda.

8/8 14:10 - 15:30 **SP16: Cluster Analysis of target data set in cluster benchmark data repository I** - Chair: *Iven Van Mechelen* (Room F)

Introduction to the IFCS cluster benchmark data repository, the two challenges connected with it, and the target data set for the second challenge, Iven Van Mechelen.

Clustering patients with non-specific low back pain, Hanneke van der Hoef.

Report on cluster analysis of target data set in Cluster Benchmark Data Repository, Michael Greenacre.

Medical back pain: A spectral clustering approach, Joey Fitch and Nazia Khan.

8/8 14:10 - 15:30 **SP17: Enumeration algorithm and data science** - Chair: *Masahiro Mizuta and Shin-ichi Minato* (Room G)

Decision diagram-based enumeration techniques and applications for statistical data analysis, Shin-ichi Minato.

Statistically significant pattern mining with applications to biology, Aika Terada.

Cluster detection using spatial scan statistic and its new development in large-scale scanning, Fumio Ishioka and Koji Kurihara.

Enumeration algorithms for political districting, Jun Kawahara, Takashi Horiyama, Keisuke Hotta, and Shin-ichi Minato.

8/8 14:10 - 15:30 **SP18: Advances in sparse regression, dimension reduction and related computations** - Chair: *Yuichi Mori* (Room H)

Cardinality-constrained multiple regression with comparisons to lasso regression, Kohei Adachi and Henk A. L. Kiers.

Tensor modeling and sliced inverse regression for tensor data, Su-Yun Huang.

Initial value selection for the alternating least squares algorithm, Masahiro Kuroda, Yuichi Mori, and Masaya Iizuka.

8/8 14:10 - 15:30 **SP19: Recent problems of big data handling in official statistics** - Chair: *Hiroe Tsubaki* (Room I)

A supervised multiclass classifier for the family income and expenditure survey, Kazumi Wada and Yukako Toko.

The IPUMS approach to harmonizing the world's population census data, Matthew Sobek.

A new statistical matching methodology using multinomial logistic regression and multivariate analysis, Isao Takabe and Satoshi Yamashita.

Components of classification on secure computation, Koji Chida, Satoshi Takahashi, Satoshi Tanaka, Ryo Kikuchi, Dai Ikarashi, Ryota Chiba, and Kiyomi Shirakawa.

8/8 14:10 - 15:30 **SP20: Causal inference and related topics** - Chair: *Yutaka Kano* (Room J)

A robust model selection criterion family and its application for the causal model, Sumito Kurata and Etsuo Hamada.

Causal analysis with cyclic structural equation models, Mario Nagase.

Identification and semiparametric adaptive estimation with nonignorable nonresponse data, Kosuke Morikawa and Jae-kwang Kim.

A high-dimensional location-dispersion model with dependent error and its applications to WTA data analysis, Ching-Kang Ing.

8/8 14:10 - 15:30 **CN03: Contributed session 03** - Chair: *Kei Kurakawa* (Room K)

The accelerated hyperbolic smoothing sum-of-distances clustering method: New computational results for solving very large instances, Adilson Elias Xavier and Vinicius Layter Xavier.

Similarity-reduced measures of diversity, François Bavaud.

*Tensor based author name disambiguation as a way of identifying author*s, Kei Kurakawa and Yasumasa Baba.

8/8 14:10 - 15:30 **CN04: Contributed session 04** - Chair: *Mark De Rooij* (Room L)

Classification based on dissimilarities towards prototypes, Beibei Yuan, Willem Heiser, and Mark de Rooij.

Three tales of linear regression, Mark De Rooij.

How to cross the river? – New distance measures, Andrzej Sokołowski, Małgorzata Markowska, Sabina Denkowska, and Dominik Rozkrut.

A data-driven method for deriving shorter DSM symptom criteria sets: The case for alcohol use disorder, Melanie Wall.

8/8 15:30 - 16:00 Coffee Break (Student Hall, 2F Building 2)

8/8 16:00 - 17:20 **SP21: Perspectives of contingency table analysis** - Organized by Shizuhiko Nishisato - Chair: *Eric Beh and Michel van de Velden* (Room A)

Correspondence analysis of overdispersed data, Eric Beh and Rosaria Lombardo.

Hybrid decomposition for three-way correspondence analysis with nominal-ordinal variables, Rosaria Lombardo, Eric J Beh, and Pieter M Kroonenberg.

Inverse CA: retrieving a contingency table from a CA solution, Michel van de Velden, Wilco van de Heuvel, Hugo Galy, and Patrick J.F. Groenen.

8/8 16:00 - 17:20 **SP22: Leading-edge research of clustering and its applications II** - Chair: *Yasunori Endo and Sadaaki Miyamoto* (Room B)

On robustness to outliers for two abstract fuzzy clustering optimization problems, Yuchi Kanzawa.

A comparative study on rough set-based k-means clustering, Seiki Ubukata, Keisuke Umado, Hiroki Kato, Akira Notsu, and Katsuhiro Honda.

A note on fuzzified even-sized clustering based on optimization, Kei Kitajima.

Two roles of cluster validity measures for clustering network data, Yukihiro Hamasuna, Ryo Ozaki, and Yasunori Endo.

8/8 16:00 - 17:20 **SP23: Classification models in finance and business II** - Chair: *Jozef Pociecha* (Room C)

Statistical analysis of the economic growth of the Central and Eastern Europe countries, Eugeniusz Gatnar.

Irrationality: A new type of error in econometric choice models?, Andreas Geyer-Schulz, Tino Fuhrmann, Marvin Schweizer, and Peter Kurz.

Robust clusterwise autoregressive conditional heteroskedasticity, Pietro Coretto, Giuseppe Storti, and Michele La Rocca.

Tail risk measurement and its application in finance, Krzysztof Jajuga.

8/8 16:00 - 17:20 **SP24: Data mining and data visualization II** - Chair: *Koji Kurihara* (Room D)

Detecting genetic association through shortest paths in a bidirected graph, Masao Ueki.

Statistical evaluation for spatial complexity based on echelon trees, Koji Kurihara, Shoji Kajinishi, and Fumio Ishioka.

R shiny implementation of lasagna plot: interactive manipulation and visualization of longitudinal data, Wataru Sakamoto and Marina Kaneda.

8/8 16:00 - 17:20 **SP25: Marketing science II** - Chair: *Takeshi Moriguchi* (Room E)

Generalization of relationship between tweets and products sales in case of new beverage products, Hiroyuki Tsurumi, Junya Masuda, and Atsuho Nakayama.

Effects of retargeting ads in the upper and lower purchase funnel, Takeshi Moriguchi, Guiyang Xiong, and Xueming Luo.

A hierarchical topic model for the e-commerce purchase behavior, Sotaro Katsumata, Eiji Motohashi, and Akihiro Nishimoto.

8/8 16:00 - 17:20 **SP26: Cluster Analysis of target data set in cluster benchmark data repository II** - Chair: *Iven Van Mechelen* (Room F)

K-means clustering on multiple correspondence analysis coordinates, Le Phan and Hongzhe Liu.

Identification of patient classes in low back pain data using crisp and fuzzy clustering methods, Alexandre Gondeau.

Cluster correspondence analysis and reduced K-means: A two-step approach to cluster low back pain patients, Fengmei Liu, Suchara Gupta, and Cristina Tortora.

Instigation of discussion, Christian Hennig.

8/8 16:00 - 17:20 **SP27: Bayesian inference and model selection** - Chair: *Yutaka Kano* (Room G)

Online ensemble learning using hierarchical bayesian model averaging, Makoto Okada, Keisuke Yano, and Fumiyasu Komaki.

Applied feasible generalized ridge regression estimation to linear basis function models, Masayuki Jimichi.

Bayesian interpretation of the ℓ_0 penalized linear regression estimator, Ryunosuke Tanabe.

Order selection for high-dimensional non-stationary time series, Shu-Hui Yu.

8/8 14:10 - 15:30 **SP28: Robustness and active learning with logistic models** - Chair: *Yuan-chin Ivan Chang* (Room H)

Double robust asymmetric logistic regression model for estimation of global marine stock abundance, Osamu Komori.

Active learning with simultaneous variable and subject selection, Zhanfeng Wang.

A greedy selection approach for active learning, Ray-Bing Chen.

Application of the influential index to active learning procedures, Yuan-chin Ivan Chang and Bo-Shiang Ke.

8/8 16:00 - 17:20 **SP29: Statistical issues on clustering and classification in medical data analysis** - Chair: *Taerim Lee* (Room I)

A Cluster analysis using gridded temperatures and precipitation data in Korea, Yung-Seop Lee, Hee-Kyung Kim, Youngho Lee, Myungjin Hyun, Jae-Won Lee.

False assignment error rate in discrete latent variable models, Donghwan Lee and Youngjo Lee.

Incorporating family disease history in risk prediction models with large-scale genetic data substantially dissolves unexplained variability, Sungho Won and Jungsoo Gim.

Clinical decision support system for hepato cellular carcinoma, Taerim Lee.

8/8 16:00 - 17:20 **SP30: Regularization methods** - Chair: *Patrick J.F. Groenen* (Room J)

Multiclass classification and meta-learning with bayes error estimates, Gerrit J.J. van den Burg and Alfred O. Hero.

Regularized generalized canonical correlation analysis 2.0, Arthur Tenenhaus, Michel Tenenhaus, and Patrick J.F. Groenen.

Sparse principal covariates regression for (ultra-)high-dimensional data: some computing issues, Katrijn Van Deun, Elise Crompvoets, and Eva Ceulemans.

8/8 16:00 - 17:20 **CN05: Contributed session 05** - Chair: *Takashi Murakami* (Room K)

Layered multivariate regression with its applications, Naoto Yamashita and Kohei Adachi.

Identifying common and distinctive components using sparse simultaneous components analysis, Niek Cornelis de Schipper and Katrijn van Deun.

Generalized orthonormal polynomial principal component analysis, Takashi Murakami.

Regularization method using gini index for core array of Tucker3 model, Jun Tsuchida and Hiroshi Yadoshisa.

8/8 16:00 - 17:20 **CN06: Contributed session 06** - Chair: *M Miftahuddin* (Room L)

Temporal based discriminant analysis of hyperspectral measurements, Nontembeko Dudeni-Tlhone.

Simulation and analysis by growth model reflecting non-uniformity, Kazuhide Namba.

Time series classification and clustering using dissimilarities between lagged distributions, Pablo Montero-Manso.

Modelling sea surface temperature in time effects perspective, M Miftahuddin, Asma Gul, Zardad Khan, Benjamin Hofner, Andreas Mayr, and Berthold Lausen.

8/8 17:30 - 19:30 **Welcome Reception** (Dining Hall, B1F Building 4)

8/8 18:00 - 20:00 **IFCS Council Meeting** (Meeting room, 2F Building 1)

Wednesday, August 9, 2017

8/9 9:00 - 9:45 **Plenary Invited 03** Chair: *Paul McNicholas* (Hall, 1F Building 2)

An overview of hybrid latent variable models and how they can be used to address outliers, flexible distributions of latent variables and issues of data dimensionality, Irini Moustaki.

8/9 9:45 - 9:55 Short Break

8/9 9:55 - 10:55 **SP31: Bayesian approach to classification** - Chair: *Kazuo Shigemasu* (Room A)

Generalized extended procrustes post-processing of MCMC samples in bayesian multidimensional scaling, Kensuke Okada and Shin-ichi Mayekawa.

A bayesian "confirmatory" factor analysis applied to WISC data, Kazuo Shigemasu and Masanori Kono.

Bayesian dynamic topic modeling with stable topics over time periods, Keisuke Takahata and Takahiro Hoshino.

8/9 9:55 - 10:55 **SP32: Classification and representation for non metric and symbolic Data I** - Chair: *J. Fernando Vera and Eva Boj del Va*l (Room B)

Model-based clustering of count data based on the logistic-normal-multinomial distribution, Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, and Javier Palarea-Albaladejo.

Robust clustering for functional data based on trimming and constraints, Luis Angel García-Escudero, Diego Rivera-García, Joaquin Ortega, and Agustín Mayo-Iscar.

Cluster distance-based regression, José Fernando Vera and Eva Boj.

8/9 9:55 - 10:55 **SP33: New topics related to classification problems I** - Chair: *Yoshikazu Terada, Michio Yamamoto and Hiroshi Yadohisa* (Room C)

Asymptotic theory of the sparse group lasso, Benjamin Poignard.

Selective inference for high dimensional classification, Yuta Umezu.

Semi-supervised learning for normal populations: A perspective from statistical missing data analysis, Kenichi Hayashi.

8/9 9:55 - 10:55 **SP34: Clustering and recommendation systems** - Chair: *Jean Diatta* (Room D)

Directional model-based clustering with social-network information for recommendation, Aghiles Salah and Mohamed Nadif.

A clustering model for decision, Pascal Préa, Célia Châtel, and François Brucker.

8/9 9:55 - 10:55 **SP35: BRCA data analysis** - Organized by H.H. Bock - Chair: *Hans A. Kestler* (Room E)

Using generalized additive models for CNV detection on multi gene panels, Corinna Ernst, Eric Hahnen, Andreas Beyer,and Rita K. Schmutzler.

8/9 9:55 - 10:55 **SP36: Issues in classification with complex data structures (CLADAG)** - Chair: *Francesco Palumbo* (Room F)

Discriminant analysis with categorical predictors: A NSCA-based approach, Alfonso Iodice D'Enza, Angelos Markos, and Francesco Palumbo.

Classifier selection and variable importance in random projection ensemble classification, Francesca Fortunato, Laura Anderlucci, and Angela Montanari.

8/9 9:55 - 10:55 **CN07: Contributed session 07** - Chair: *Kunihiro Kimura* (Room G)

Acquiescence or deep processing? latent class regression for modeling response styles, Kunihiro Kimura.

A two-step approach to latent class analysis for models with complex dependencies. detecting and modeling residual dependencies, Zsuzsa Bakk.

Latent class trees, Mattis van den Bergh.

8/9 9:55 - 10:55 **CN08: Contributed session 08** - Chair: *Daniel Baier* (Room H)

Association rules and community detection on bi-partite network, Giuseppe Giordano.

Improving the ordering and delivery processes in online-fashion shops: New approaches to integrate the voice of the customer, Daniel Baier and Alexandra Rese.

A study on individual small-area differences clustering of residential characteristics by any selection from differences scaling, Mitsuhiro Tsuji, Mayumi Ueda, and Toshio Shimokawa.

8/9 9:55 - 10:55 **CN09: Contributed session 09** - Chair: *Yasumasa Baba* (Room I)

Prediction by regression models with missing in covariates, Fumiya Sano, Kaito Takano, Shintaro Tomatsu, and Manabu Iwasaki.

*Advances in sequential automatic search of subset of classifier*s, Luca Frigau, Giulia Contu, and Francesco Mola.

Application of pseudo data based on continuous-discrete transformation to multivariate analysis, Yasumasa Baba.

8/9 10:55 - 11:25 Coffee Break (Student Hall, 2F Building 2)

8/9 11:25 - 12:25 **SP37: Classification of (un)complex data (CLAD)** - Chair: *José G. Dias* (Room A)

Hidden links, known figures - clustering co-authorship networks across scientific fields, Paula Brito, Conceição Rocha, Fernando Silva, and Alípio M. Jorge.

l1norm SVMS for binary regression, Pedro Duarte Silva.

Multilevel typologies: Classic and symbolic data approaches, José G. Dias.

8/9 11:25 - 12:25 **SP38: Classification and representation for non metric and symbolic Data II** - Chair: *J. Fernando Vera and Eva Boj del Val* (Room B)

Clustering of histograms using a simulated annealing algorithm, Alejandro Chacon and Javier Trejos.

Classification with distance-based logistic regression taking into account prediction error, Eva Boj del Val and Teresa Costa Cor.

External logistic biplot for nominal and ordinal data, Jose Luis Vicente-Villardon.

8/9 11:25 - 12:25 **SP39: New topics related to classification problems II** - Chair: *Yoshikazu Terada, Michio Yamamoto and Hiroshi Yadohisa* (Room C)

Perfect simple structure estimation via extension of quartimin criterion, Kei Hirose.

Screening procedure for auxiliary variables in the gaussian mixture model, Shinpei Imori.

Dimension reduction clustering based on constrained centroids, Kensuke Tanioka.

8/9 11:25 - 12:25 **SP40: Clustering and classification in innovative data science** - Chair: *Fionn Murtagh* (Room D)

Linear time visualization and search in big data using pixellated factor space mapping, Fionn Murtagh.

Core clustering as a tool for tackling noise in cluster labels, Renato Cordeiro de Amorim, Vladimir Makarenkov, and Boris Mirkin.

8/9 11:25 - 12:25 **SP41: Cluster analysis using non-gaussian mixtures -** Chair: *Brian Franczak* (Room E)

The multiple scaled contaminated Gaussian distribution, Cristina Tortora and Antonio Punzo.

Addressing overfitting in model-based clustering with the t-distribution, Jeffrey Andrews.

On clustering longitudinal data, Paul D. McNicholas.

8/9 11:25 - 12:25 **CN10: Contributed session 10** - Chair: *Łukasz Smaga* (Room F)

Inference for general MANOVA based on ANOVA-type statistic, Łukasz Smaga.

Spherization of multiway tables - A linear algebraic method for multiple testing problems using maximum statistics and concept of distance, Kazuhisa Nagashima, Tapati Basak, Satoshi Kajimoto, and Ryo Yamada.

Feature description of truncated normal distributions of additive tests for multiple testing p-value correction – An application to GWAS SNP data, Tapati Basak, Kazuhisa Nagashima, Satoshi Kajimoto, and Ryo Yamada.

8/9 11:25 - 12:25 **CN11: Contributed session 11** - Chair: *Shinobu Tatsunami* (Room G)

Usefulness of factor analyses for validity evaluation of a foot and ankle related quality of life outcome instrument, Shinobu Tatsunami, Takahiko Ueno, Naoki Haraguchi, and Hisateru Niki.

Data quality management of chain stores based on outlier detection, Linh Nguyen and Tsukasa Ishigaki.

Proposal of efficient defensive formation and classification the batters it valid, Kazuki Konda and Yoshiro Yamamoto.

8/9 11:25 - 12:25 **CN12: Contributed session 12** - Chair: *Seong-Keon Lee* (Room H)

Decision tree approaches for interval valued symbolic data response, Seong-Keon Lee.

Random forests followed by ABC analysis as a feature selection procedure for machine-learning, Jorn Lötsch, and Alfred Ultsch.

A flexible approach to identify interaction effects between moderators in meta-analysis, Xinru Li, Elise Dusseldorp, and Jaqueline J. Meulman.

8/9 11:25 - 12:25 **CN13: Contributed session 13** - Chair: *Tsai-Hung maqz* (Room I)

*Optimal treatment regime estimation: a clustering problem with a well-defined, pragmatic optimality criterio*n, Iven Van Mechelen and Aniek Sies.

Type I censored life test of a two-component series system under bivariate Weibull lifetime distribution, Tsai-Hung Fan, She-Kai Ju, and N. Balakrishnan.

Copula selection in random coefficient degradation models, Shuen-Lin Jeng.

8/9 12:25 - 13:25 Lunch Break

8/9 13:25 - 14:10 **Presidential Address** Chair: *Berthold Lausen* (Hall, 1F Building 2)

Dissimilarity based on manner of dissimilarity/similarity to/from the others, Akinori Okada.

8/9 14:10 - 15:00 **Award Session** Chair: *Yasumasa Baba* (Hall, 1F Building 2)

8/9 15:15 **Bus Departure for Short Tour** (Optional Tour)

8/9 19:00 - 21:00 **Conference Dinner at Meiji Kinenkan**

Thursday, August 10, 2017

8/10 9:00 - 9:45 **Plenary Invited 04** Chair: *Heon Jin Park* (Hall, 1F Building 2)

Classification by quantiles, Cinzia Viroli.

8/10 9:45 - 10:30 **Plenary Invited 05** Chair: *Yoshiro Yamamoto* (Hall, 1F Building 2)

Decisions that are needed when using cluster analysis, and research that helps with making them, Christian Hennig.

8/10 10:30 - 11:00 Coffee Break (Student Hall, 2F Building 2)

8/10 11:00 - 11:45 **Poster Session** (Student Hall, 2F Building 2)

Sufficient dimension reduction via random-partition for large-p-small-n problem, Hung Hung.

Model based clustering for tensor-valued data strcutures, Kohei Uno.

Extensions of Pearson's inequality between skewness and kurtosis to multivariate cases, Haruhiko Ogasawara.

Leveraging local data structure for multi-view correlation analysis with graph-structured associations, Akifumi Okuno and Hidetoshi Shimodaira.

Multi-criteria classifications in regional development modelling, Beata Bal-Domańska.

Data analysis of the cross-resistance rate between antibiotic drugs by nonmetric asymmetric multidimensional scaling, Yasutoshi Hatsuda, Syou Maki, Yasuhiro Ishimaki, Katsuhito Nagai, Sachiko Omotani, Junji Mukai, Masahiko Taguchi, Michiaki Myotoku, and Tadashi Imaizumi.

Spatial diversification of employment structures vs. educational capital in the European Union, Elżbieta Sobczak and Beata Bal-Domańska.

Penalized multidimensional unfolding of asymmetric data with self-distances constrained to be short, Gaku Tsujii and Kohei Adachi.

Motion-blurred image restoration, Huey-Miin Hsueh.

Fused sliced average variance estimation, Sungmin Won, Hyoin Ahn, and Jae Keun Yoo.

Permutation dimension test of fused sliced inverse regression, YooNa Cho and Jae Keun Yoo.

The use of the robust regresion estimator to estimate small trade firms, Grażyna Dehnel.

Using k-means classification method within GREG estimation to assess small enterprises in Poland, Grazyna Dehnel, Marek Walesiak, and Jacek Kowalewski.

Multi-category classification model of internet game and alcohol addiction based on resting-state EEG data, MinJi Cho, Yeonjung Hong, Sohyun Lee, Ji Yoon Lee, Yeon Jin Kim, Jung-Seok Choi, and Donghwan Lee.

A study on predictors of internet and smartphone addiction in adolescents via structural equation modeling, Sohyun Lee, Eunmin Park, and Donghwan Lee.

Nonlinear factor analysis for NCT test score linking, Tatsuo Otsu.

Recursive sparse path analysis via lasso type penalty, Ji Yao Li.

8/10 11:45 - 12:35 Lunch Break

8/10 12:35 - 14:15 **SP42: Advanced technique for analyzing (big) multi-set/multi-subject data** - Chair: *Tom Wilderjans* (Room A)

Estimating cross-source relationships from big data using component- and networks-analysis, Pia Tio, Lourens Waldorp, and Katrijn Van Deun.

Integrating customer data sets using topic models, Takuya Satomura.

Detecting disease subtypes by means of cluster independent component analysis (C-ICA) of multi-subject brain data, Jeffrey Durieux, Serge A.R.B. Rombouts, and Tom Frans Wilderjans.

Quantifying similarity in brain responses based on multi-subject eeg data: a simulation study to compare the inter-subject correlation (ISC) measure to novel methods, Tom F Wilderjans and Wouter D Weeda.

8/10 12:35 - 14:15 **SP43: Survey data analysis** - Chair: *Ryozo Yoshino* (Room B)

The longitudinal and cross-national values survey: Cultural manifold analysis of national character, Ryozo Yoshino.

A trial of diagnostic cut-off point selection in three-class classification for health questionnaire, Kazue Yamaoka, Yoshinori Nakata, Mutsuhiro Nakao, Kei Asayama, Mariko Inoue, and Toshiro Tango.

Cross-national comparison on collective characteristics of cultural symbols in Asia-Pacific area, Yuejun Zheng.

Analysis of visit record in interviewer mediated surveys: A case study using the survey on the Japanese national character and others, Tadahiko Maeda.

Cross-national studies of trust: searching for characteristics of trsut based on comparative surveys among eight nations, Fumi Hayashi and Masamichi Sasaki.

8/10 12:35 - 14:15 **SP44: Analysis and clustering of complicated data** - Chair: *Junji Nakano* (Room C)

A new clustering algorithm via graph connectivity, Ying-Chao Hung, Yu-Feng Li, and Liang-Hung Lu.

Kernel methods for symbolic data, Woncheol Jang.

Interactive visualization of characteristics of groups, Yoshikazu Yamamoto, Junji Nakano, and Nobuo Shimizu.

Dissimilarity by chi-squared statistic for aggregated symbolic data with continuous and categorical variables, Nobuo Shimizu, Junji Nakano, and Yoshikazu Yamamoto.

A study on distance between fields and fusion of different fields by using co-authored information of article, Yuji Mizukami.

8/10 12:35 - 14:15 **SP45: Text classification** - Organized by Masakatsu Murakami - Chair: *Yuichiro Kobayashi* (Room D)

The stylometry on Japanese political documents, Yohei Ono.

Quantitative storyline structure of "The Tale of Utsuho", Gen Tsuchiyama.

A comparative evaluation of feature selection methods, Wanwan Zheng and Mingzhe Jin.

The changes over time in Kōji Uno's writing style, Xueqin Liu.

Automated speech scoring: A text classification approach, Yuichiro Kobayashi.

8/10 12:35 - 14:15 **SP46: Data Science, classification and clustering** - Chair: *Berthold Lausen* (Room E)

Collinear by design. The econometrics of product configuration data, Andreas Geyer-Schulz.

Distances, bayes errors, and classification method performance in high dimensions, Claus Weihs, and Tobias Kassner.

New approaches for brand confusion analysis based on ad similarities, Daniel Baier.

Exhaustive relabeling experiments for biomarker selection, Ludwig Lausser, Alexander Groß, and Hans A. Kestler.

8/10 12:35 - 14:15 **SP47: Methods of data analysis and statistical measures in the social sciences -** Chair: *Theodore Chadjipantelis* (Room F)

Optimal model-based clustering with multilevel data, Fulvia Pennoni, Francesco Bartolucci, and Silvia Bacci.

A comparison of different applications of functional linear discriminant analysis, Sugnet Lubbe.

Changes in the gendered division of labor and women's economic contributions within Japanese couples, Miki Nakai.

Determining the similarity index in electoral behavior analysis: An issue voting behavioral mapping, Theodore Chadjipantelis and Georgia Panagiotidou.

8/10 12:35 - 14:15 **SP48: Cluster validation: New developments and issues II** - Organized by Christan Hennig - Chair: *Hans-Joachim Mucha* (Room G)

Evaluating fuzzy clustering results, Elena Rihova.

Understanding external validity indices, Matthijs J. Warrens.

Validation of sum of squared error clustering: Bootstrapping versus subsampling in view of the influence of multiple points, Hans-Joachim Mucha.

Decomposing information-theoretic validity indices, Hanneke van der Hoef.

8/10 12:35 - 14:15 **CN14: Contributed session 14** - Chair: *Masahiro Mizuta* (Room H)

An integrative tool for visualization of gene set analysis, Chen-An Tsai.

Functional data analysis approach for fatty acid concentration changes, YiFan Chen, Rojeet Shrestha, Ken-ichi Hirano, Shu-Ping Hui, Hitoshi Chiba, Yuriko Komiya, and Masahiro Mizuta.

RNAseq data clustering: comapartive study, Smail Jamail, Abderrahmane Sbihi, and Ahmed Moussa.

Extension of sinkhorn method: optimal movement estimation of agents under law of inertia, Daigo Okada, Naotoshi Nakamura, Yosuke Fujii, Takuya Wada, Ayako Iwasaki, and Ryo Yamada.

8/10 12:35 - 14:15 **CN15: Contributed session 15** - Chair: *Noboru Wakayama* (Room I)

Optimal doubling burn-in policy based on tweedie processes with applications to degradation data, Chien-Yu Peng.

Critical thinking ability scale development on item response theory, Noboru Wakayama, Yoshimitsu Miyazawa, Shinji Kajitani, and Maomi Ueno.

Nonparametric semi-supervised classification with application to signal detection in high energy physics, Alessandro Casa and Giovanna Menardi.

OASW foundation and some discussion on results, Fatima Batool and Christian Hennig.

8/10 12:35 - 14:15 **CN16: Contributed session 16** - Chair: *Guangliang Chen* (Room J)

Joint clustering of batch data in the presence of inter-sample variations, Sharon Lee.

Privacy preserving em learning for model-based clustering, Kaleb Leemaqz and Sharon Lee.

Novel geometrically adaptive scaling techniques for multiscale gaussian-kernel SVM, Guangliang Chen.

8/10 14:15 - 14:25 Short Break

8/10 14:25 - 15:25 **Closing Ceremony** (Hall, 1F Building 2)

Abstracts

Contents

Cardinality-constrained multiple regression with comparisons to lasso regression

Kohei Adachi and Henk A. L. Kiers

Abstract Sparse multiple regression (MR) refers to the modified MR procedures which provide the regression coefficient vectors including a number of zero elements with their locations also optimally estimated. In the existing sparse MR procedures, among which Lasso is popular, the sum of the MR loss function and a penalty function multiplied by a tuning parameter is minimized. Here, the penalty causes a number of elements in the regression coefficient vector to become zero, and the tuning parameter controls the cardinality of the vector (i.e, the number of its non-zero elements). The difficulties in such penalized approaches are that the correspondence of a tuning parameter value to the resulting cardinality is unknown in advance, and the domain of the tuning parameters consists of a continuum real values, which hence cannot be considered exhaustively. Here, we propose a sparse MR procedure without such difficulties. In this procedure, instead of using a penalty function, the MR loss function is minimized subject to the constriant that a prespecified number of elements of the regression coefficient vector becomes 0. In other words, the cardinality of the coefficient vector is constrained to be a pre-specified integer. We call this procedure cardinality-constrained MR (CCMR), in which the resulting cardinality is fixed and the domain of the cardinality consists of the limited number of integers, all of which can be considered. Kiers' (2002) iterative majorization algorithm is used for obtaining the CCMR solution. We also present a procedure for selecting the best cardinality with an information criterion. In simulation studies, CCMR is compared against Lasso in the goodness of recovering the true coefficients.

Keywords: Sparse Regression: Unpenalized Approach

Kohei Adachi
Osaka University e-mail: adachi@hus.osaka-u.ac.jp

Henk A. L. Kiers
University of Groningen e-mail: h.a.l.kiers@rug.nl

Addressing overfitting in model-based clustering with the t-distribution

Jeffrey Andrews

Abstract We introduce a bootstrap augmented EM-style algorithm for mixtures of multivariate t-distributions which simultaneously addresses two problems that arise during traditional parameter estimation via the EM algorithm: convergence to poor local maxima and overconfidence in clustering probabilities.

Keywords: Clustering; Parameter Estimation; Bootstrap; Multivariate t Distribution;

Jeffrey Andrews
University of British Columbia | Okanagan Campus e-mail: jeff.andrews@ubc.ca

Clustering sum of type loans using fuzzy c-means algorithm method

Khoiru Nissa Apriliya and Hanif Novrandhita

Abstract Economic development is an absolute prerequisite for the countries in the world, in order to minimize the distance the catch in economics and prosperity society of the advanced industrial countries. Shortcut of capital resources, many government in the world tried to bring in resources from abroad various types of loans. One solution to solve problem is carry out mapping the area to give an advice many government for their policy. Fuzzy C-Means clustering method is a clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. This research to cluster 120 all developing countries (based on World Bank Data) countries from 6 variables, first variable is external debt stocks, debt outstanding and disbursing, Disbursements, Interest payments, principal repayments, net debt inflows in 2014. The results analysis cluster to 4 groups using Xie and Beni index 0.0007595654. Iteration until converged, 49 iteration, with error 0.1 e-10. Based on cluster center, group 1 average cluster center high, with 3 countries. Group 2 average cluster center middle, with 10 countries. Group 3 average cluster center low with 103 countries, Group 4 average cluster center very high, with 3 countries.

Keywords: Fuzzy Clustering Means, Xie and Beni, Loan, Developing Countries, World Bank, Cluster

Khoiru Nissa Apriliya
Islamic University of Indonesia e-mail: khoiru.nissa4@gmail.com
https://www.facebook.com/chairu.aprileaStatistics

Hanif Novrandhita
Bank Woori Saudara e-mail: hanifnovrandhita@gmail.com
https://www.facebook.com/hanif.novrandhita?fref=ts-

Application of pseudo data based on continuous-discrete transformation to multivariate analysis

Yasumasa Baba

Abstract As one example, in social research surveys, income questions are framed in terms of income classes, and age is sometimes similarly coded. The processes of coding or categorizing are the same as those of transforming continuous data into discrete data. Such transformation from original continuous observations to discrete data is useful for condensing data if the information loss caused by the transformation is small.

In previous research, we obtained the surprising result that even if continuous quantities were each classified into only 5 categories, almost the same result as with the original data could be obtained in a principal component analysis (Baba 2010). The reason for this is that the correlation structures were only changed slightly by such transformation. This was because the correlation between two variables is a quadratic function of the variables. If L is the range of the data and c is the width of each class, then the order of difference between the variance of the original and discrete data is $(c/L)^2$. In principal component analysis, superior components come from the dominant parts of the variance-covariance matrix and the difference between the matrices calculated from the original and discretized data is of the same order, $(c/L)^2$.

Therefore, even if we used 5 categories data, similar results were obtained. These results prompt us to apply continuous-discrete transformation to conceal personal information represented as continuous data. After transforming the original continuous data into discrete data, we add random numbers and thereby obtain pseudo data. Such data have desirable features in a statistical sense. In this paper, the similarity of multivariate analysis results of continuous data to those after continuous-discrete-continuous transformation is discussed.

Reference

Baba, Y. (2010): Continuous-discrete Transformation and Multivariate Analysis. Proceedings of JKCS-2010. 185-186.

Keywords: categorical data, multivariate Analysis, concealment

Yasumasa Baba
School of Statistical Thinking, Professor Emeritus The Institute of Statistical Mathematics e-mail: baba@ism.ac.jp

Improving the ordering and delivery processes in online-fashion shops: New approaches to integrate the voice of the customer

Daniel Baier and Alexandra Rese

Abstract Nowadays, many online-fashion shops are permanently confronted with the question how to improve their ordering and delivery processes: Should they invest in new additional customer services like virtual-try ons, curated shopping, same-day delivery, or fixed delivery time or not? Will these new quality attributes effectively improve service quality in the eyes of the customer and consequently sales? Or are they just another nice-to-have gimmick without effects? Ideally, for cost and budget reasons, these questions should be answered before the new services are implemented. The traditional approach to answer such questions is the use of customer surveys based on the Kano model. This well-known data collection and analysis approach was developed by Noriaki Kano in the 1980s for service engineering purposes and describes how five types of quality attributes can be distinguished according to their relationship between functionality and contribution to customer satisfaction: must-be, indifferent, one-dimensional, attractive, and reverse. Over the years, the model has proven its usefulness in many applications and currently is frequently used to evaluate possible improvements of e-commerce sites. However, recently, revised Kano models have been introduced and their superiority compared to the traditional approach has become apparent. In this paper we discuss these revisions and test them in a real-world e-commerce site improvement application: A random sample of 5,162 customers of a large German online-fashion retailer participated in a survey based on the Kano model. They rated 14 quality attributes according to functional and dysfunctional aspects. The collected data were analyzed using the Kano model and its revisions resulting in valuable implications for online-fashion shop improvements.

Keywords: Kano model; online shopping; e-commerce site improvement

Daniel Baier
University of Bayreuth Chair for Innovation and Dialogue Marketing Universitätsstraße 30 D-95447 Bayreuth Germany
e-mail: daniel.baier@uni-bayreuth.de

Alexandra Rese
University of Bayreuth Chair for Innovation and Dialogue Marketing Universitätsstraße 30 D-95447 Bayreuth Germany
e-mail: alexandra.rese@uni-bayreuth.de

New approaches for brand confusion analysis based on ad similarities

Daniel Baier

Abstract Brand confusion occurs when a consumer is exposed to an advertisement (ad) for brand A but believes that it is for brand B. If more consumers are confused in this direction than in the other one (assuming that an ad for B is for A), this asymmetry is a disadvantage for A. Consequently, the confusion potential and structure of ads has to be checked: A sample of consumers is exposed to a sample of ads. For each ad the consumers have to specify their guess about the advertised brand. Then, the collected data are aggregated and analyzed using, e.g., MDS or two-mode clustering.Recently, a new approach where image data analysis and classification is applied for this purpose have been proposed and compared to these traditional approaches (Baier, Frost 2017): The confusion potential and structure of ads is related to featurewise distances between ads and – to model asymmetric effects – to the strengths of the advertised brands. The approach has been successfully tested in the German beer market. In this paper we further develop this new approach and apply it to further datasets. The results are encouraging.

Keywords: Marketing; Image Data Analysis and Classification; Confusion Data

Daniel Baier
Full professor, University of Bayreuth Chair of Innovation and Dialog Marketing e-mail: daniel.baier@uni-bayreuth.de http://www.innodialog.uni-bayreuth.de

A two-step approach to latent class analysis for models with complex dependencies. detecting and modeling residual dependencies

Zsuzsa Bakk

Abstract Latent class analysis is an approach widely used in the social sciences for classifying individuals based on a set of charcteristics called indicators (such as symptoms of depression, attitude toward different minorities) into a set of unobserved, latent groups. In most instances the classification is only the first step. The interest of scientist is in defining antecedents of the clustering, for example to model how attitudes vary based on education and nationality.While traditionally these relationship is modeled using a three step approach, this has multiple downsides, namely it introduces a classification error that later needs to be corrected and it is impossible to model direct effects between the indicators of the latent class variable and the antecedents.Hereby I propose a novel two step approach that makes it possible to model the direct effects of interest while keeping the measurment model fixed. The approach does not involve the unnecessary classification step, that creates porblems in the three-step approach. I present the results of a simulation study that compares the two step approach to exisitng approaches focusing on parameter bias and efficiency.

Keywords: latent class analysis, stepwise estimators, efficiency, bias

Zsuzsa Bakk
Department of Methodology and StatisticsAssistant Professor., Leiden University e-mail: z.bakk@fsw.leidenuniv.nl

Multi-criteria classifications in regional development modelling

Beata Bal-Domańska

Abstract The article presents the discussion regarding the influence of taking selected approaches to the classification of regions on estimation results of Solow-Swan growth models. As the existing modelling effects, obtained in the area of regional development, show the estimation results of the production function parameters' assessment, based on Solow-Swan growth model or beta-convergence models, are divided depending on the period and units covered by the study (countries, regions). The article presents modelling results of the development of economies in the European Union regions at NUTS 2 level, depending on the development factors in line with the extended MRW (Mankiw-Romer-Weil) growth model in the selected period for various groups of regions. Each time the econometric analysis was conducted in the groups of NUTS 2 regions, separated in terms of structure of economy or the level of smart specialization. However, every time, along with the change in the grouping method, the dividing boundary of the inclusion of regions into a given class was changing. It allowed the assessment of estimations stability depending on the inclusion of the selected regions into particular classes of smart specialization. To sum up, the conducted analysis allowed for presenting conclusions regarding the stability of the obtained estimations in the groups of the European Union regions characterized by a different level of smart specialization, and also for evaluating the relationship between growth factors and the level of economic development in the discussed groups.

Keywords: Classification, Econometric Modelling, regions

Beata Bal-Domańska
Wrocław University of Economics e-mail: beata.bal-domanska@ue.wroc.pl

Hidden markov models for the analysis of longitudinal data through the LMest package

Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni

Abstract We illustrate the main functions of the R package LMest (available from http://CRAN.R-project.org/package=LMes), which can be used to estimate the basic Latent Markov (LM) model and its extended formulations (Bartolucci *et al.*, 2017). The model is tailored for the analysis of longitudinal categorical data; for a detailed illustration of the model from a methodological point of view. we refer to the book of Bartolucci *et al.* (2013) and the discussion paper of Bartolucci *et al.* (2014).
The LMest package has several distinguishing features over the existing R packages for similar models. In particular, it is designed to deal with longitudinal data, that is, with (even many) i.i.d. replicates of (usually short) sequences of data, and it can be used with univariate and multivariate categorical outcomes. The package also allows us to deal with missing responses, including drop-out and non-monotonic missingness, under the missing-at-random assumption (Little and Rubin, 2002). Moreover, standard errors for the parameter estimates are obtained by exact computation of the information matrix or through reliable numerical approximations of this matrix. When the parameter space of the LM model is bounded, standard errors may also be obtained through parametric bootstrap (Davison, and Hinkley, 1997). Finally, computationally efficient algorithms are implemented for estimation and prediction of the latent states, also by relying on certain Fortran routines.
More in detail, we illustrate how, through the main functions of the LMest package, we can estimate the basic LM model and how to estimate the effect of time-fixed and time-varying covariates on the conditional distribution of the responses given the latent variable (measurement model). Covariates may also affect the initial and transition probabilities of the latent process (latent model), which is assumed to follow a first-order Markov chain.
When the covariates are included in the measurement model, through an LM model we can account for the unobserved heterogeneity, that is, the heterogeneity between individuals that cannot be explained on the basis of the observed covariates. The advantage with respect to a standard random-effects or latent class model with covariates (Lazarsfeld and Henry, 1968; Vermunt *et al.*, 1999; Pennoni, 2014) is that the unobservable heterogeneity is allowed to be time varying. In formulating the model, the conditional distribution of the responses given the latent variables and the covariates is assumed to be multinomial and it is parametrized by means of global logits. We also illustrate how to explore the effect of covariates on the latent trait underlying the responses, which is supposed to change over time and is represented by the Markov chain. In this context, the individual covariates are assumed to affect the initial and transition probabilities of the latent process through a multinomial logit parameterization. We finally show how to deal with clustered sample units, by employing the mixed LM model (van de Pol, and Langeheine, 1990). In this case, the initial and transition probabilities of the latent process are allowed to vary across different latent subpopulations defined by an additional discrete latent variable.

Francesco Bartolucci
Department of Economics, University of Perugia e-mail: francesco.bartolucci@unipg.it

Silvia Pandolfi
Department of Economics, University of Perugia e-mail: silvia.pandofi@unimib.it

Fulvia Pennoni
Department of Statistics and Quantitative Methods,University of Milano-Bicocca e-mail: fulvia.pennoni@unimib.it
http://www.statistica.unimib.it/utenti/pennoni/

Maximum likelihood estimation of the model parameters is based on a forward recursion (Baum *et al.* 1970), which is used to compute the manifest distribution of the response variables given the observed covariates. Then, the Expectation-Maximization (EM; Dempster *et al.* 1977) algorithm is used to maximize the likelihood function. This algorithm is based on a suitable forward-backward recursion adapted by the first proposal of Baum *et al.* (1970) and Welch (2003). The problems of model selection and multimodality of the likelihood function are jointly dealt with by allowing for different sets of starting values for the EM algorithm, which are found on the basis of a deterministic and a random starting rule, and then comparing the value of the likelihood function at convergence. Two main criteria are provided to select the number of latent states: the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978).
Finally, the LMest package also allows us to perform decoding, namely predicting the overall sequence of latent states for a certain sample unit on the basis of the data observed for this unit. In particular, local decoding is related to the prediction of the latent state of each unit in the sample at each time occasion, and is provided by maximizing the estimated posterior probabilities for any covariate and response configuration observed at least once. On the other hand, global decoding is aimed at tracking the latent state of a subject across time, on the basis of the *a posteriori* most likely sequence of states. The global decoding is performed by an adaptation of the Viterbi algorithm (Viterbi, 1967, Juan and Rabiner, 1991), which uses a forward and backward recursion having a complexity similar to the recursions adopted for maximum likelihood estimation of the parameters of the LM model.
To illustrate the potential of the package and for a correct interpretation of the latent states, we consider real data applications in different research fields. In particular, we consider data on self-reported health status that are derived from the American social survey related to the Health and Retirement Study (HRS) conducted by the University of Michigan. We also consider data on job satisfaction deriving from the Russia Longitudinal Monitoring Survey and data on conviction histories related to a cohort of offenders in England and Wales.

Main References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, Csáki. F. (eds.), *Second International Symposium on Information Theory*, 267-281. Akadémiai Kiado, Budapest.

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: An R Package for latent Markov models for longitudinal categorical data. *To appear on Journal of Statistical Software*. Bartolucci, F., Farcomeni, A., and Pennoni. F. (2013*). Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press, Boca Raton.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates (with discussion). *TEST*, **23**, 433-465.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, MA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd edition. John Wiley, New York.

Pennoni, F. (2014). *Issues on the Estimation of Latent Variable and Latent Class Models*. Scholars' Press, Saarbücken.

Keywords: Expectation-maximization algorithm; forward-backward recursions; time-varying unobserved heterogeneity

Feature description of truncated normal distributions of additive tests for multiple testing p-value correction – An application to GWAS SNP data

Tapati Basak, Kazuhisa Nagashima, Satoshi Kajimoto, and Ryo Yamada

Abstract Multiple testing of different genetic models for contingency tables in case-control studies are often performed in genetic association study. Failure to compensate for multiple testing can have important real-world consequences of producing false positives. This feature is common for dealing with data sets from the techniques such as microarrays, evolutionary genomics etc. So, the nominal p-values must be corrected to control type I error. In this study, we apply a Spherization based linear algebraic method which handles multiple marker tests in the context of geometric statistics. We correct the nominal p-values using truncated normal distribution [Botev, Z. I., 2017] that gives the extensive way to simulate from truncated multivariate normal distribution in higher dimensions.
We apply this method to GWAS SNP data for additive tests. We evaluate for three types of SNP subsets considering: simple SNP sets having same numbers of SNPS that are located physically on genome, LD blocks that are defined somewhere over the genome, SNPS located in gene-locus. The probabilities are calculated for multiple threshold points of truncated normal distribution for SNP subsets. Finally, we extract the features of our truncated normal distribution approach using GWAS real data.

Keywords: Multiple testing; Spherization; Truncated Normal Distribution; Additive Tests

Tapati Basak
Graduate Student, Unit of Statistical Genetics, Kyoto University,Kyoto, Japan, Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan e-mail: tapati@genome.med.kyoto-u.ac.jp

Kazuhisa Nagashima
Research Fellow, Unit of Statistical Genetics, Kyoto University, Kyoto, Japan, Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan e-mail: kazuhisa.nagashima@genome.med.kyoto-u.ac.jp

Satoshi Kajimoto
Medical Doctor, Kitano Hospital, Medical Doctor, Kitano Hospital e-mail: kajimoto.satoshi@gmail.com

Ryo Yamada
Professor, Unit of Statistical Genetics, Kyoto University, Kyoto, Japan, Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan e-mail: ryamada@genome.med.kyoto-u.ac.jp

OASW foundation and some discussion on results

Fatima Batool and Christian Hennig

Abstract Usually finding the sensible clustering solution for a given real data application of interest and estimation of appropriate number of clusters are dealt with separately. Many clustering methods require determination of the correct number of clusters beforehand. Rousseeuw (1987) has proposed average silhouette width (ASW) to access cluster quality based on cluster tightness and separation concepts. This criterion can be used to validate the relative quality of clusters and to estimate the number of clusters. A new clustering algorithm is proposed here based on optimization of the average silhouette width in a partitional clustering framework. The algorithm just needs pairwise distances between objects to estimate the number of clusters and produce a clustering against this number. The algorithm needs to be initialized. For this various already existing initialization methods and four new methods has been considered for comparisons. The new proposed method is tested on various simulated data sets including, elongated, spherical, clusters with different spreads, compact, different sized clusters, relatively small clusters in presence of bigger clusters and clusters coming from different distributions. Some of the results will be discussed in the talk with an indication of potential further work and challenges. Also, some data sets are indicated for which the method is not suitable to use for clustering.

References

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Keywords: Clustering; hierarchical models; average silhouette width; Estimation of number of clusters

Fatima Batool
2nd year PhD student at Department of Statistical Science University College London. Lecturer in Statistcs at Department of Statistics CIIT Lahore Pakistan., University College London e-mail: ucakfba@ucl.ac.uk

Christian Hennig
Senior Lecturer Department of Statistical Science UCL e-mail: c.hennig@ucl.ac.uk

Similarity-reduced measures of diversity

François Bavaud

Abstract Usual measures of diversity (such as the Shannon, Rényi or Tsallis entropies) consider the relative abundance of the observed categories. Taking also into account the possible similarities between categories, as sometimes proposed in quantitative ecology, aims at reducing the diversity measure. This contribution presents a new measure of similarity-reduced diversity, called effective entropy, formally defined as the optimal solution of a regularized, origin-constrained transportation problem, where origins and destinations consist of the categories under consideration. Equivalently, effective entropy results from a soft clustering problem with fixed centroids, and can be computed iteratively. The present framework also provides a basic confusion mechanism, where initial source categories (the stimuli) are mapped into target categories (the percepts). Increasing the value of the regularization parameter or temperature produces a series of phase transitions, successively merging all clusters into a single prototypical percept encompassing all the initial stimuli. Effective entropy is concave, that is increases under weighted pooling of abundance distributions. It reduces in the low-temperature limit to the Shannon entropy on distinct categories. Also, in the case of bivariate features with additive dissimilarities, the effective entropy is bounded above by the sum of univariate effective entropies, with equality if and only the bivariate features are independently distributed. The theory is illustrated on linguistic data (varieties of English), ecological data (trees variety) and supervised classification (numerals confusion matrix).

Keywords: Diversity measures; Soft clustering; Regularized optimal transportation; Confusion matrices; Information theory; Transformed histograms

François Bavaud
University of Lausanne e-mail: fbavaud@unil.ch

Correspondence analysis of overdispersed data

Eric Beh and Rosaria Lombardo

Abstract Traditionally, simple correspondence analysis (CA) applied to a two-way contingency table is performed by decomposing a matrix of standardized residuals using singular value decomposition (SVD) where the sum-of-squares of these residuals is Pearson's chi-squared statistic. Such residuals, which are treated as being asymptotically normally distributed, arise by assuming that the cell frequencies are Poisson random variables so that their mean and variance are the same. However, an examination of this issue – including those of Haberman (1973) and Agresti (2002, pg 81) – suggest that the variance of the residuals is less than one. Thus, we observe overdispersion in the table. Various strategies can be undertaken to study, and deal with, overdispersion. These include performing CA by stabilising the variance, by applying an SVD on the adjusted residuals (Beh, 2010) or by assuming the cell frequencies are from a generalised Poisson distribution (Consul, 1989) or the Conway-Maxwell Poisson distribution (Conway & Maxwell, 1962). We shall provide an overview of some of these issues and explore the consequences of adopting such strategies.

References

Agresti, A. (2002). *Categorical Data Analysis*, (2^{nd} edn.), Wiley.Beh, E. J. (2012), Simple correspondence analysis using adjusted residuals, *Journal of Statistical Planning and Inference*, 142, 965 – 973.
Consul, P. C. (1989), *Generalized Poisson Distributions: Properties and Applications*, Marcel Dekker, Inc.
Conway, R. W.; Maxwell, W. L. (1962), A queuing model with state dependent service rates, *Journal of Industrial Engineering, 12, 132 – 136.*
Haberman, S. J. (1973), The analysis of residuals in cross-classified tables, *Biometrics*, 75, 457 – 467.

Keywords: Correspondence analysis; overdispersion; standardised residual; adjusted residual; generalised Poisson distribution; Comway-Maxwell Poisson distribution

Eric Beh
University of Newcastle, Australia e-mail: eric.beh@newcastle.edu.au

Rosaria Lombardo
University of Campania e-mail: rosaria.lombardo@unicampania.it

Determinants of survival risk factors for HIV/AIDS patients on ART in a developing country - accounting for clustering at facility level

Renette Julia Blignaut and Innocent Mapoa

Abstract Introduction :
The burden of HIV/AIDS morbidity and mortality is very high in the developing countries with the Sub-Saharan Africa bearing the greater share of the disease. Antiretroviral treatment has given a lot of hope for survival to infected individuals yet their life span is still short compared to those un-infected. In order to design more effective HIV/AIDS treatment programs, this study tries to understand the risk factors that influence the survival of patients on antiretroviral treatment. This study was conducted on adult patients who were enrolled in antiretroviral treatment at the Katutura hospital in Windhoek, Namibia.
Study objectives: The main aim of this study was to understand which risk factors could impact survival rates of patients on antiretroviral treatment for HIV/AIDS.
Data collection: A retrospective cohort study design was used to collect information on adults (15 years or older) who had HIV/AIDS and who started on antiretroviral treatment between 01 January 2006 and 31 December 2010 at the Katutura hospital in Namibia. A sample of 500 adults were randomly selected from the records of HIV/AIDS patients. Due to missing information only 473 patients could be included in the analyses. Ethical approval to utilize the data was obtained from the Ministry of Health and Social Services (MoHSS), Namibia. It should be noted that the database did not contain any number or reference that could be traced back to a specific patient. The data were first described and then survival analysis techniques such as the log-rank test, Kaplan-Meier curves and proportional hazard models were applied to answer the research aim.
Results: Of the 473 patients, 62% were female. The majority of the sample were aged 25 to less than 35 (52%) and were single (83%). Ninety-two percent had a functional status as "working". Forty percent were at an early stage of the disease (stage 1), 25% were at a slightly more advanced stage (stage 2) and the remaining patients were at an advanced stage 29% (stage 3) and 7% (stage 4). Seventeen percent contracted tuberculosis whereas 46% had some opportunistic infection whilst on treatment. Slightly more than a quarter indicated that they suffered some side-effects due to the medication. A total of 111 (24%) patients died during the time of the study.
Concluding remarks: The product-limit survival estimates revealed that patients who started treatment in the early stages of the disease (stages 1 and 2) had a greater probability of survival compared to the more advanced stages (stages 3 and 4). Factors that influence the chances of survival were the CD4 count and weight when starting treatment and whether the patient had a treatment supporter who assisted with medication administration.

Keywords: HIV/AIDS; Log-rank; survival; Kaplan-Meier curves; proportional hazard models

Renette Julia Blignaut
Professor in Statistics, Department of Statistics and Population Studies, University of the Western Cape., University of the Western Cape MDAG South Africa e-mail: rblignaut@uwc.ac.za

Innocent Maposa
the Namibian University of Science and Technology e-mail: Imaposa@nust.na

Hidden links, known figures - clustering co-authorship networks across scientific fields

Paula Brito, Conceição Rocha, Fernando Silva, and Alípio M. Jorge

Abstract The way research collaboration is organized differs across scientific fields. In this work, we focus on the comparison and clustering of different scientific fields based on their collaboration networks. To this purpose, we use co-authorship networks, where nodes are researchers and the edges indicate the existence of at least one common publication. The data come from the University of Porto, covering 22 scientific fields and publications ranging from 1972 to 2016, organized in different time periods: before and after 2000, and per year after 2000.
Following previous work by Giordano and Brito (2014), we represent each network by the distributions of statistical indicators, such as degree, eigenvector centrality, or closeness. Since data cannot be considered uniform by intervals, the representation of distributions by histogram-valued variables is not appropriate. Therefore, we use a representation by quantile vectors (Ichino, 2008). Distributions are compared using the discrete version of the Mallows distance between quantile functions (Levine, Bickel, 2001). Given that network parameters cannot be considered independent, the (discrete version of the) Mahalanobis-Wassertein distance (Verde, Irpino, 2008) is used.
Hierarchical clustering is then applied, using the Ward aggregation index, leading to a typology of scientific fields for each time period. A separate analysis of co-authorship networks of a given field in different years allows for a better insight of the evolution of research collaboration along time, identifying stable periods and moments of change in the collaborations structure.
Acknowledgments This work is partially funded by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF) within project "TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020"

Keywords: Co-authorship networks; Distributional data; Network clustering; Symbolic data analysis

Paula Brito
FEP & LIAAD - INESC TEC, University of Porto e-mail: mpbrito@fep.up.pt
http://www.fep.up.pt/docentes/mpbrito

Conceição Rocha
LIAAD - INESC TEC, University of Porto e-mail: conceicao.n.rocha@inesctec.pt

Fernando Silva
Fac. Ciências & CRACS - INESC TEC, Universty of Porto e-mail: fds@fc.up.pt

Alípio M. Jorge
Fac. Ciências & LIAAD - INESC TEC, University of Porto e-mail: amjorge@fc.up.pt

Nonparametric semi-supervised classification with application to signal detection in high energy physics

Alessandro Casa and Giovanna Menardi

Abstract
Since the early Sixties, the Standard Model has represented the state of the art in High Energy Physics. It describes how the fundamental particles interact with each others and with the forces between them, giving rise to the matter in the universe. Despite its empirical confirmations, there are indications that the Standard Model does itself not complete our understanding of the universe. Model independent search aims to explain the shortcomings of this theory by empirically looking for any possible signal which behaves as a deviation from the background process, representing, in turn, the known physics.From a statistical perspective, this problem can be in principle formulated within an unsupervised framework of clustering. However, while the the signal, if present, is unknown, the background process is always present and well-known, so that a virtually infinite sample of data can be simulated from the latter process with Montecarlo techinques. Hence, available data have two different sources: an unlabelled sample which might include observations from both the processes, and an additional labelled, sample from the background only. A semisupervised approach can be particularly suitable in this context, for discriminating the two class labels; semisupervised classification techniques lie between unsupervised and supervised ones, sharing some characteristics of both the approaches. In this work we propose a procedure where additional information, available on the background, is integrated within a nonparametric clustering framework to detect deviations from known physics. Also, we propose a variable selection procedure that allows to work on a reduced subspace.The effectiveness of the whole methodology is shown via its application on a set of data related to a simulated experiment of a proton-proton collision.

Keywords: High energy physics; Nonparametric clustering; Semisupervised classification

Alessandro Casa
Department of Statistical Science, University of Padova e-mail: casa@stat.unipd.it

Giovanna Menardi
Department of Statistical Science, University of Padova e-mail: menardi@stat.unipd.it

Assessing model-based clustering methods with cytometry data sets

Gilles Celeux and Jean-Patrick Baudry

Abstract
There is no universally relevant clustering method, some appear to be useful for each specific task. Thus there is a need to describe the features of each method to help to decide its relevance for a given field of application. We attempt to do so for model-based clustering by applying it to a situation for which the task is clearly identified. High-dimensional flow and mass cytometry allow to describe many cells in great detail. Hence the need for clustering methods to detect groups of cells with similar protein marker expression profiles. A common task is to predict patients disease status. This can be done based on characteristics of the cells clusters of each patient. A challenging characteristic is that relevant groups of cells are typically rare populations. This supervised task provides us with a clear benchmark to assess the relevance of the clustering methods.

Keywords: Model-based Clustering; Cytometry; Unsupervised and Supervised Classification

Gilles Celeux
Inria Saclay-Île-de-France e-mail: Gilles.Celeux@inria.fr

Jean-Patrick Baudry
LSTA, UPMC e-mail: Jean-Patrick@upmc.fr

Clustering of histograms using a simulated annealing algorithm

Alejandro Chacon and Javier Trejos

Abstract
In this research, the clustering of histogram symbolic data is explored. The clustering is performed using a method based on partitions that uses a simulated annealing algorithm as a tool to control the search of the best partition. The Mallows L2 dissimilarity measure is used as the criterion to define the distance between elements of the dataset. This dissimilarity measure is used, among other reasons, because it allows to define a prototype or barycenter for the clusters under consideration, this is important because then it is possible to use the concepts of total inertia decomposition through the use of a Huygens'-type Theorem. The implementation of this dissimilarity measure follows closely the proposal of Billard and Kosmelj. In the simulated annealing algorithm, an exponential cooling schedule is used for the reduction of the temperature parameter. The number of transitions at a given temperature, that is the length of the Markov chain considered at that temperature, depends on the magnitude of the temperature parameter.At each transition, an object (histogram) is transferred from one cluster into another and the change in the within inertia is evaluated using the Metropolis rule.The algorithm is applied to two real datasets. The first one contains the monthly average temperatures for each of the 48 contiguous states of USA from 1895 to 2004. A total of 48 objects and 12 variables comprise the symbolic data set. The second dataset corresponds to maximum daily rainfall data registered during 40 years for 14 meteorological stations in Costa Rica. Each register is a histogram of rain fallen during a storm.The results are then compared with the clustering of the same data using the dynamic clustering algorithm as proposed by Irpino, Verde and Lechavalier. A better performance is obtained using the simulated annealing algorithm.

Keywords: Clustering; Symbolic Data; Histogram Data; Simulated Annealing; Mallows Dissimilarity

Alejandro Chacon
DEHC Ingenieros Consultores e-mail: achacon@dehc.cr http://www.dehc.cr

Javier Trejos
Dean, Faculty of ScienceFull Professor, School of MathematicsResearcher, Research Centre for Pura and Applied Mathematics, University of Costa Rica e-mail: javier.trejos@ucr.ac.cr http://www.cimpa.ucr.ac.cr/investigadores.html

Determining the similarity index in electoral behavior analysis: An issue voting behavioral mapping

Theodore Chadjipantelis and Georgia Panagiotidou

Abstract
The objective of the research is to propose an alternative similarity metrics method in electoral analysis, regarding voter's and candidate's opinions about issues.
Distance metrics have been basically used for the identification of similarities for the analysis of electoral behavior. However, in electoral behavior analysis several factors exist who create a need to reconsider the way we should count the proximity between two objects.
The similarity index should include behavioral and social factors that contribute to the formation of voter's opinion upon issues. A more efficient similarity index which will be able to translate the similarities or dissimilarities between voters and candidates based on issues is essential and should be researched.
For the purposes of this paper a dataset from Greek Helpmevote application (VAA) from 2015. The dataset contains the answers of voters and candidate parties on questions regarding issues on a Likert scale 1-5.
According to the original idea the aim is to classify all respondents and parties regarding their positions on issues. Subjects will be classified according the probability to participate in each group with discriminant analysis. The new variable will be clustered to create groups of respondents with similar attitude on issues and all clusters will be placed on the axis system.
Furthermore, clusters will be matched to issue variables with correspondence analysis. The analysis will match groups of already classified respondents according to their similarity on issues using probabilities instead of other metric systems (distances, x^2) and then match clusters with issues that seem to interest them.
The procedure is described as an issue voting behavioral mapping: ALl groups of respondents and issues will be positioned on an axis system, creating multiple behavioral contexts and depicting positions and distances among groups and issues. Issue voting behavioral mapping

Keywords: Classification; Discriminant;Correspondence;Clustering;Similarity metrics;Voting behavior;Electoral analysis;Political analysis;Issue voting;Behavioural mapping;

THEODORE CHADJIPANTELIS
Department of Political Sciences, ARISTOTLE UNIVERSITY OF THESSALONIKI e-mail: chadji@polsci.auth.gr

PANAGIOTIDOU
Department of Political SciencesLecturer on Contract, ARISTOTLE UNIVERSITY OF THESSALONIKI e-mail: gvpanag@polsci.auth.gr

Application of the influential index to active learning procedures

Yuan-chin Ivan Chang and Bo-Shiang Ke

Abstract
In this talk, we apply model-free influential indexes, which are develpoed based on classification score functions, to active learning procedures targeting at the area under receiver operating characteristic (ROC) curve. Becasue these novel indexes do not depend on the specific classification models used in active learning procedures, they can applied to modern complicated classification methods raised in machine learning literature such as the method of support vector machine. In this talk, the statsitical properties of these indexes will be discussed. Numerical results using synthesised data and some real examples will be presented.

Keywords: ROC curve, AUC, active learning

Yuan-chin Ivan Chang
Research Fellow, Institute of Statistical Science, Academia Sinica, Institute of Statistical Science, Academia Sinica e-mail: ycchang@sinica.edu.tw

Bo-Shiang Ke
PhD StudentDepartment of StatisticsInstitute of StatisticsNational Chiao Tung UniversityHsinchu, TAIWAN , Department of Statistics, Institute of Statistics, National Chiao Tung University, Hsinchu e-mail: naivete0907@gmail.com

Novel geometrically adaptive scaling techniques for multiscale gaussian-kernel svm

Guangliang Chen

Abstract
We address the problem of hyperparameter tuning in multiscale Gaussian-kernel SVM. In the single-scale setting, the classifier contains two parameters - C (tradeoff) and sigma (bandwidth) - that are often hard to tune. Many techniques have been developed to address that challenge, however, they all ignore the intrinsic geometry of the training data and always operate in a large, fixed region and thus are computationally inefficient. In our previous work, we introduced a simple, fast, accurate nearest-neighbor approach that directly infers the bandwidth parameter sigma from the training data (by exploiting its local geometry). In this work, we extend the previous approach in a natural way to construct multiscale Gaussian kernels by using a sequence of adaptive sigma values, each determined from a distinct training class. Our resulting algorithm has a clear geometric motivation and seems to produce very promising results in preliminary experiments.

Keywords: multiscale Gaussian-kernel SVM; hyperparameter tuning; nearest neighbor approach

Guangliang Chen
Assistant ProfessorDepartment of Mathematics and Statistics, San Jose State University e-mail: guangliang.chen@sjsu.edu http://www.math.sjsu.edu/ gchen/

A greedy selection approach for active learning

Ray-Bing Chen

Abstract
In this work, we are interested in active learning for the binary classification problems based on the logistic regression models. According to the concept of the active learning, a sequential design procedure is adopted as a subject selection procedure to select the unlabeled points into training set. In addition, we implement a greedy selection approach to identify proper classification model based on the current training set. Thus the proposed active learning algorithm sequentially iterates these two procedures to include more training points and then to update the classification model. Comparing with the classification results obtained by the whole training set with the full model, the simulation results show that the proposed algorithm can obtain the almost the same classification rates with much fewer training points based on a compact classification model. Finally a real example is also used to illustrate the performance of the proposed algorithm.

Keywords: Classification; logistic regression model; sequential design

Ray-Bing Chen
Department of Statistics, National Cheng Kung University e-mail: rbchen@mail.ncku.edu.tw

Functional data analysis approach for fatty acid concentration changes

YiFan Chen, Rojeet Shrestha, Ken-ichi Hirano, Shu-Ping Hui, Hitoshi Chiba, Yuriko Komiya, and Masahiro Mizuta

Abstract
We propose a method to analyze fatty acid concentration changes with Functional Data Analysis (FDA). The analysis on variations or changes of fatty acid concentration is a very essential part in the biological and pharmaceutical field; the changes depend on many conditions including age, sex, and drug dosage. People do research on the relationship between the concentration of fatty acid and conditions with the traditional statistic methods. FDA is a powerful approach in the field of data analysis; we can deal with target objects as functions. This means that FDA is good for analyzing the relationship. FDA has many advantages, e.g. we can adopt integration or deviations of the functions.
We assume the fatty acid concentration changes are described as functions on time t, and conditions, $f_i(t, condition_1, \cdots, condition_p)$, where $i = 1, \cdots, N$ and N is the size of data. We can get area under the curve (AUC) values with the integrations of the functions and can also get the peak values with the derivations of the functions. Furthermore, datasets of fatty acid can be analyzed using functional cluster analysis and functional principal components analysis.
We will show the results with actual fatty acid concentration change datasets.

Keywords: FDA; functional cluster analysis; functional principal components analysis

YiFan Chen
Hokkaido University e-mail: feifeiyifan@gmail.com

Rojeet Shrestha
e-mail: srojeet@hs.hokudai.ac.jp

Ken-ichi Hirano
e-mail: khirano@kb3.so-net.ne.jp

Shu-Ping Hui
e-mail: keino@hs.hokudai.ac.jp

Hitoshi Chiba
e-mail: chibahit@med.hokudai.ac.jp

Yuriko Komiya
e-mail: komiya@iic.hokudai.ac.jp

Masahiro Mizuta
e-mail: mizuta@iic.hokudai.ac.jp

Components of classification on secure computation

Koji Chida, Satoshi Takahashi, Satoshi Tanaka, Ryo Kikuchi, Dai Ikarashi, Ryota Chiba, and Kiyomi Shirakawa

Abstract
We introduce some important components of classification on a secure computation system and empirically evaluate them. The secure computation is a cryptographic technology that enables us to operate data while keeping the data encrypted. Due to the remarkable aspect, we can possibly construct a secure on-line analytical system based on the secure computation to protect against unauthorized access, computer virus and internal fraud. Moreover, the function of secure computation has a benefit for privacy. Suppose that a researcher would like to hold a statistical survey using highly confidential data of consumers stored in a smartphone or monitoring device. A potential solution to realize it while keeping privacy is that our confidential data can be encrypted in the smartphone or monitoring device and the encrypted data is analyzed without leaking individual data.In this research, we focus on a secure computation for classification. As a related study, Lindell and Pinkas proposed a secure computation protocol for decision-tree analysis in 2000. However, achieving general and practical secure computation system for classification available to big data is challenging task so far. Our approach is starting by making components of classification on a secure computation system. They include comparison, mapping and logarithmic calculation. A mapping operation can be used to categorize detailed data into summarized ones. We evaluate the components using an existing secure computation system that can be performed from statistical computing and graphics software "R". By combining some of the components using R script language, we can potentially make new functions for classification. Our future works include an implementation of a secure on-line analytical system based on the secure computation involving classification. A demonstration experiment to verify the practicality and scalability of the system in the field of official statistics is also in our scope.

Keywords: Secure Computation; Classification; Security; Privacy; Big Data; Official Statistics

Koji Chida
Secure Platform Laboratories, Senior Research Engineer, NTT Corporation e-mail: chida.koji@lab.ntt.co.jp

Satoshi Takahashi
Secure Platform Laboratories, Engineer, NTT Corporation e-mail: takahashi.s@lab.ntt.co.jp

Satoshi Tanaka
Secure Platform Laboratories, Researcher, NTT Corporation e-mail: tanaka.s@lab.ntt.co.jp

Ryo Kikuchi
Secure Platform Laboratories, Researcher, NTT Corporation e-mail: kikuchi.ryo@lab.ntt.co.jp

Dai Ikarashi
Secure Platform Laboratories, Research Engineer, NTT Corporation e-mail: ikarashi.dai@lab.ntt.co.jp

Ryota Chiba
Institute of Economic Research, Research Associate, Hitotsubashi University e-mail: rchiba@ier.hit-u.ac.jp

Kiyomi Shirakawa
Institute of Economic Research, Associate Professor, Hitotsubashi University e-mail: kshirakawa@ier.hit-u.ac.jp

Identifying changes in mean functions for a sequence of functional data

Jeng-Min Chiou

Abstract We propose a systematic approach to identifying changes in a sequence of functional data, where the number of change-points and their positions are unknown. The algorithm comprises the dynamic segmentation approach coupled with the backward elimination step. The method recursively searches for a pre-specified number of change-point candidates that are then further assured using hypotheses testing for statistical significance via a backward elimination procedure. We discuss the optimality property of the change-point estimates and show that the change-point candidates are consistent with the actual ones if they exist. We examine the finite sample performance of the algorithm via a simulation study and illustrate the method through an application to traffic flow analysis. This is a joint work with Yu-Ting Chen.

Keywords: changepoint analysis; functional data analysis; hypothesis test

Jeng-Min Chiou
Research Fellow / Professor Institute of Statistical Science, Academis Sinica, Taiwan
e-mail: jmchiou@stat.sinica.edu.tw

Multi-category classification model of internet game and alcohol addiction based on resting-state EEG data

MinJi Cho, Yeonjung Hong, Sohyun Lee, Ji Yoon Lee, Yeon Jin Kim, Jung-Seok and Donghwan Lee

Abstract It is known that the internet game addiction and alcohol addiction are similar in terms of the neurobiological mechanisms and psychological factors. To develop the classification model for distinguishing two addictions well, we use the resting-state quantitative electroencephalography (QEEG) data as well as clinical and neurocognitive data. To determine the final model, we compare the prediction performance of various existing supervised learning techniques such as support vector machine, random forest, sparse logistic regression and sparse generalized partial least squares. We also identify relevant predictors and their combinations in QEEG data to improve the classification accuracy. Furthermore, by proposing a two-stage procedure incorporating dimension reduction and feature selection, we try to provide a better characterization of QEEG for clustering subtypes within each addiction.

Keywords: Classification; Internet Addiction

MinJi Cho
Ewha Womans University e-mail: hie1031@naver.com

Yeonjung Hong
Ewha Womans University e-mail: yjhong00@naver.com

Sohyun Lee
Ewha Womans University e-mail: gusthgus@naver.com

Ji Yoon Lee
SMG-SNU Boramae Medical Center e-mail: idiyuni91@gmail.com

Yeon Jin Kim
SMG-SNU Boramae Medical Center e-mail: loveaj1220@nate.com

Jung-Seok Choi
SMG-SNU Boramae Medical Center e-mail: psychoresi@hanmail.net

Donghwan Lee
Ewha Womans University e-mail: donghwan.lee@ewha.ac.kr

Permutation dimension test of fused sliced inverse regression

YooNa Cho and Jae Keun Yoo

Abstract Rencetly, a fused approach of sliced inverse regression has been proposed to estimate the central subspace. It overcomes the deficit of sliced inverse regression, whose results depend on the numbers of slices. The fused approach is robust to the choices of the slices, but it does not proivde how to determine the true dimension of the central subspace. Here, a permutation approach is proposed to estimate the structural dimension of the fused sliced inverse regression. Also, we present numerical studies for the test through various simulation models. A real data analysis is presented for the illustration purpose.

Keywords: Fused approach; Permutation test; Sliced inverse regression; Sufficient dimension reduction

YooNa Cho
Graduate StudentDepartment of StatisticsEwha Womans University, Ewha Womans University e-mail: yoonacho92@gmail.com

Jae Keun Yoo
Associate ProfessorDepartment of StatisticsEwha Womans University, Ewha Womans University e-mail: peter.yoo@ewha.ac.kr

Bi-modal clustering and quantification theory: Flexible-filter clustering of contingency and response-pattern tables. Numerical demonstration of a proposed framework

Jose G Clavel and Shizuhiko Nishisato

Abstract Following Part 1, the present paper explains every detail of the proposed procedure, which starts with the construction of the condensed response-pattern table F, which is $mn \times (m+n)$ from the $m \times n$ contingency table C, then we will carry out analysis of F using dual scaling, yielding $(m+n-2)$ components. Using $(m+n-2) \times (m+n)$ matrix of coordinates, we will compute the within-column distance matrix, which is $(m+n) \times (m+n)$. From this square distance matrix, we will extract the between-variable distance matrix D_{xy}, which is $m \times n$. Note that this distance matrix is calculated, using $(m+n-2)$ dual scaling components. To study exhaustiveness of analysis, we will compare total information analysis based on C and that based on F. In this comparison, we should note that the distance between two variables assessed in smaller-dimensional space cannot be larger than the distance calculated in larger-dimensional space. This difference reflects the amount of information captured by analysis of C and that by F. Nishisato and Clavel (2017) used Heuer's suicide data which turned out to be difficult to handle by hierarchical clustering and partitioning clustering, while clustering with flexible filters, applied to the between-set distance matrix, proved to be useful for interpretation. Therefore, we will use the same example for the current paper. It is hoped that the numerical illustrations of the detailed procedure offer some insight into its use. Again, it will be shown that the framework of 'non-overlapping' clusters can be too restrictive to be useful for clustering of a large data set. The current study has an ambitious goal to derive some guidelines on how to use clustering with flexible filters for identifying clusters or characterizing clusters.

Keywords: expande multidimensional space; overlapping vs non-overlapping clusters; total vs between-set clustering

Jose G Clavel
ProfessorDepartment of Quantitative MethodsSchool of Economy and Business Administration, Universidad de Murcia, Spain e-mail: jjgarvel@um.es

Shizuhiko Nishisato
Professor Emeritus, University of Toronto, Canada, University of Toronto e-mail: shizuhiko.nishisato@utoronto.ca

Model-based clustering of count data based on the logistic-normal-multinomial distribution

Marc Comas-Cufí, Josep Antoni Martín-Fernández, Glòria Mateu-Figueras, and Javier Palarea-Albaladejo

Abstract The logistic-normal-multinomial (LNM) results from compounding a logistic-normal and a multinomial distribution. The logistic-normal models the multinomial probabilities as a random vector defined on a simplex, whereas the multinomial part connects them to the observed vectors of counts. Analogously to the use of multivariate normal mixtures for model-based clustering in real space, we investigate here the scope for considering mixtures of LNM distributions with the purpose of performing clustering analysis of multinomial count data. Although the parameters of a LNM distribution can be approximately estimated using different Monte Carlo Expectation-Maximization (MCEM) algorithms, the computational burden can be prohibitive in a general practical setting. In this work, we firstly discuss advantages and disadvantages of using a mixture of LNM distributions to model the variability within and between clusters. We then introduce an approach to identify grouping structure in the data based on the combination of a classification EM algorithm and a MCEM algorithm. The proposed methodology is illustrated in different scenarios and the results are finally compared to those obtained using current alternative approaches.

Keywords: Model-based Clustering; Mixture models; Count data; Compositional Data analysis

Marc Comas-Cufí
Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona
e-mail: mcomas@imae.udg.edu

Josep Antoni Martín-Fernández
Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona
e-mail: jamf@imae.udg.edu

Glòria Mateu-Figueras
Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona
e-mail: gloria@imae.udg.edu

Javier Palarea-Albaladejo
Biomathematics and Statistics Scotland e-mail: javier.palarea@bioss.ac.uk

A statistical model for supporting AirB&B hosts activity

Giulia Contu, Luca Frigau, and Claudio Conversano

Abstract In the last few years, new forms of hospitality have been developed. These new kinds of accommodation have come up beside at the traditional forms of hospitality. They are represented, for instance, by those based on sharing rooms and apartments. In time, they have become significant and have had an impact at economic and touristic level. One of the most famous examples of sharing accommodation is AirBnB. This company has joined together the owners of apartments and rooms who want to rent their space to improve their profits. In this paper we focus on hosts and their activities and use data about AirBnB hosts operating in the region of Sardinia (Italy) in 2016. First, we cluster hosts to understand their characteristics and to assess the role of variables characterizing each cluster. Then, we use the Vector Generalized Additive Models (VGAMs) to understand how different categories of hosts can improve their occupancy rates.

Keywords: AirBnB; VGAM; clustering,

Giulia Contu
University of Cagliari e-mail: giulia.contu@unica.it

Luca Frigau
University of Cagliari e-mail: frigau@unica.it

Claudio Conversano
University of Cagliari e-mail: conversa@unica.it

Core clustering as a tool for tackling noise in cluster labels

Renato Cordeiro de Amorim, Vladimir Makarenkov, and Boris Mirkin

Abstract Real-world data sets often contain mislabelled entities. This can be particularly problematic if the data set is being used by a supervised classification algorithm at its learningphase. In this case the accuracy of this classification algorithm, when applied to unla-belled data, is likely to suffer considerably. In this paper we introduce a clustering-based method capable of reducing the number of mislabelled entities in data sets. Our method can be summarised as follows: (i) cluster the data set; (ii) select the entities that have the most potential to be assigned to correct clusters; (iii) use the entities of the previous step to define the core clusters and map them to the labels using a confusion matrix;(iv) use the core clusters and our cluster membership criterion to correct the labels ofthe remaining entities. We perform numerous experiments to validate our method em-pirically using k-nearest neighbour classifiers as a benchmark. We experiment with both synthetic and real-world data sets with different proportions of mislabelled entities. Our experiments demonstrate that the proposed method produces promising results. Thus, it could be used as a pre-processing data correction step of a supervised machine learning algorithm.

Keywords: label noise; clustering; k -means; core clustering; Minkowski distance.

Renato Cordeiro de Amorim
School of Computer Science University of Hertfordshire Hatfield AL10 9AB UK e-mail: r.amorim@herts.ac.uk

Vladimir Makarenkov
Département d'informatique, Université du Qubec à Montréal, C.P. 8888 succ. Centre-Ville,Montreal (QC) H3C 3P8 Canada. e-mail: makarenkov.vladimir@uqam.ca

Boris Mirkin
Department of Data Analysis and Machine Intelligence, National Research University Higher School of Economics, Moscow, Russian Federation And, Department of Computer Science e-mail: mirkin@dcs.bbk.ac.uk

Robust clusterwise autoregressive conditional heteroskedasticity

Pietro Coretto, Giuseppe Storti, and Michele La Rocca

Abstract The aim of the work is twofold. First, we investigate the inhomogeneity of the cross-sectional distribution of realized stock volatilities. This inhomogeneity it is shown to be well captured by a finite Gaussian mixture model plus a uniform component that represents the "noise" generated by abnormal variations in returns. In fact, it is common that in a cross-section of realized volatilities there is a small proportion of stocks showing extreme behavior. The mixture model is used to cluster stocks into groups of homogeneous realized volatility. The clustering is based on constrained MLE, constraints impose a certain separation between clusters and noise. An EM-algorithm with iterative M-step is developed. We show that clustering information can be profitably used for specifying parsimonious state-dependent models for volatility forecasting. We propose novel GARCH-type model specifications whose parameters can vary through time as a function of the probability that, at a given time point, the stock belongs to a specific volatility cluster. Finally, the empirical performance of the proposed models is assessed by means of an application to a panel of U.S. stocks traded on the NYSE.

Keywords: GARCH; model-based clustering; robustness; mixture models

Pietro Coretto
Department of Economics and Statistics; Università degli Studi di Salerno (Italy) e-mail: pcoretto@unisa.it http://www.decg.it/pcoretto/

Giuseppe Storti
Department of Economics and Statistics; Università degli Studi di Salerno (Italy) e-mail: storti@unisa.it http://docenti.unisa.it/005005/home

Michele La Rocca
Department of Economics and Statistics; Università degli Studi di Salerno (Italy) e-mail: larocca@unisa.it http://docenti.unisa.it/001580/home

Using classification of regions based on the complexity of the global progress indices for supporting sustainable development

Anita Csesznák and Eva Sandor-Kriszt

Abstract This paper gives an overview of the applicability of recently published global indices for measuring social, environmental and economic progress at different regional levels. The critical investigation is based on determining the relevance and comparability of the referred data sources and aggregation techniques used for composing complex indicators. Advantages and barriers for appropriate geographical/regional classification in context of measuring achievement of globally recognized sustainable development goals, are presented and prioritized. Classification of sustainability measures related to the development of new business and non-profit models based on shared values provides new tools for evaluation and interpretation of both temporary and longer term regional differences in support of global and local development policies. By this way the current and desirable states of the key drivers for human well-being development, like innovation in higher education systems, can be better identified and explained for the policy-makers, especially in Central-Eastern Europe, where dramatic changes of current "business as usual" strategies and models and the well and fast adaption of globally successful innovative solutions are necessary at all overarching macro, medium and micro levels within the common universe of society, environment and economy.

Keywords: classification; sustainability; regional differencies

Anita Csesznák
Budapest Business School e-mail: orosznecsesznak.anita@uni-bge.hu

Eva Sandor-Kriszt
e-mail: kriszt.eva@uni-bge.hu

Three tales of linear regression

Mark De Rooij

Abstract Linear regression is one of the most often used statistical techniques in data analysis. We will discuss three different ways of using linear regression: a descriptive, an explanatory, and a predictive way. Each of these ways requires a different story, but we feel that the manner in which regression is used is often a hybrid way. We focus on differences in assumptions underlying the three stories in terms of sampling, model specification, and model testing. We point out that each of the three stories has its own advantages and disadvantages. Furthermore, we take the standpoint that in teaching regression the different stories should be clearly distinguished, in order to avoid confusion in the usage of regression models.

Keywords: Regression, Sampling, Inference, Bias-Variance tradeoff

Mark De Rooij
Leiden University Institute of Psychology Methodology and Statistics Department e-mail: Rooijm@fsw.leidenuniv.nl
Http://www.markderooij.info

When factorial invariance fails: a new rotation approach for multigroup exploratory factor analysis to identify loading-specific differences

Kim De Roover and Jeroen K. Vermunt

Abstract In the literature, multigroup exploratory factor analysis (EFA) has been gaining popularity to address measurement invariance. An important indicator of this fact is the emergence of exploratory structural equation modeling (Asparouhov & Muthén, 2009). Firstly, respecifying confirmatory factor analysis (CFA) models in an exploratory fashion is problematic (Browne, 2001; MacCallum, Roznowski, & Necowitz, 1992) and using EFA as a precursor to CFA has proven to be a better strategy (Gerbing & Hamilton, 1996). Secondly, fixed zero loadings are often too restrictive (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996).
When using multigroup EFA, the rotational freedom per group needs to be resolved (1) for attaining an interpretable simple structure (Dolan, Oort, Stoel, & Wicherts) and (2) to enable hypothesis testing for the loadings (Asparouhov & Muthén, 2009). With respect to the latter, Jennrich (1973, 2001, 2002) did some important work in obtaining optimally rotated maximum likelihood estimates. We extended this approach to better accommodate the multigroup case and the search for measurement invariance (or the lack thereof). Specifically, each group is rotated both to simple structure per group and agreement across groups and loading differences are disentangled from differences in the structural model (i.e., factor variances and correlations). In a simulation study, the new approach is applied with several rotational criteria and their performance is evaluated and compared regarding the identification of group-specific measurement models and loading-specific differences between groups.

Keywords: measurement invariance; factorial invariance; multigroup exploratory factor analysis; rotation identification; hypothesis testing

Kim De Roover
Assistant ProfessorDepartment of Methodology and StatisticsTilburg School of Social and Behavioral Sciences, Tilburg University e-mail: k.deroover@uvt.nl

Jeroen K. Vermunt
Full ProfessorDepartment of Methodology and StatisticsTilburg School of Social and Behavioral Sciences, Tilburg University e-mail: j.k.vermunt@uvt.nl

Identifying common and distinctive components using sparse simultaneous components analysis

Niek Cornelis de Schipper and Katrijn van Deun

Abstract Combining data from multiple sources on the same subjects to reveal an underlying structure is getting more common as data gets easier to collect. For example, the integration of traditional survey data with genetics data to study gene-environment interaction (Boyd et al., 2012). A popular framework to analyze these integrated data from multiple sources is Sparse Simultaneous Component Analysis (SSCA), a generalization of Sparse Principal Components Analysis (SPCA). However, it is challenging to reveal common components shared by all data sources and distinctive components unique to each data source using SSCA, as the underlying common and unique struc- ture could be very subtile. Furthermore, SSCA is not specifically aimed at revealing the common and distinctive components. We pro- pose an alternative method that imposes a predefined sparse structure on the components weights in order to reveal the common and distinc- tive components. In this paper we will present an efficient coordinate descent algorithm for SSCA that yields common and distinctive com- ponents. It includes SPCA as a special case. The performance of the method and the algorithm will be addressed in a simulation study. Preliminary results indicate that the underlying common structure can be recovered well using the algorithm.

Keywords: Sparse Simultaneous Component Analysis, Common and Distinctive Components

Niek Cornelis de Schipper
Methodology and Statistics: PhD Candidate, Tilburg University e-mail: n.c.deschipper@uvt.nl

Katrijn van Deun
Methdology and Statistics: Associate Professor, Tilburg University e-mail: k.vandeun@uvt.nl

Using k-means classification method within GREG estimation to assess small enterprises in Poland

Grazyna Dehnel, Marek Walesiak, and Jacek Kowalewski

Abstract Economic development places new demands on short-term business statistics. Statistical data are expected to be delivered in short intervals, with improved accuracy and coherence. One of the major problems involved in estimating information about economic activity across small domains is the inappropriate sample size and incompleteness of data sources. The distribution of enterprises by target variables tends to be considerably skewed to the right, with high variation, high kurtosis, and usually contains outliers. The pressure to produce accurate estimates at a low level of aggregation or substantially reduce sample sizes has increased the importance of looking for possibilities of applying new, more sophisticated estimation methods. The paper presents an attempt to estimate basic economic information about Polish small enterprises by applying k-means classification to one of the small area statistics methods – generalized regression *estimation* (*GREG*). The study was conducted at a low level of aggregation. The domain of interest was a unit resulting from joint cross-classification by province and economic activity category. The k-means method was used to identify similar groups of domains of interest. This made it possible to use a GREG estimator at domain level, based on an extended sample, consisting of similar domains of interest. Lagged variables from the administrative registers were used as auxiliary variables. Estimation precision was assessed using the bootstrap method. The purpose of the study was to increase the effectiveness of estimates and extend of the scope of information both in terms of the number of variables and the range of breakdowns available in statistical outputs.

References

Rao J.N.K., Molina I., 2015, *Small area estimation*. Wiley series in survey methodology, 2nd ed. Hoboken, Wiley, New Jersey.

Wu, J., 2012. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.

Keywords: GREG estimation, k-means classification method, business statistics

Grazyna Dehnel
Poznań University of Economics and Business e-mail: g.dehnel@ue.poznan.pl

Marek Walesiak
Department of Econometrics and Informatics, Wrocław University of Economics e-mail: marek.walesiak@ue.wroc.pl

Jacek Kowalewski
Statistical Office in Poznan/Poznań University of Economics and Business e-mail: j.kowalewski@stat.gov.pl

The use of the robust regression estimator to estimate small trade firms

Grażyna Dehnel

Abstract In the face of dynamic changes in the economy, there is a growing demand for multivariate statistics for cross-classified domains. In economic statistics this demand poses a particular challenge owing to the unique character of the population of enterprises, which is what motivates the search for estimation methods that can exploit administrative sources to a greater extent. The adoption of new solutions in this area is expected to increase the scope of statistical outputs and improve the efficiency of estimates. The purpose of the present study is to test the application of the regression estimator based on the LS method and least median of squares regression to estimate characteristics of Polish small trade firms. The estimation process is supported with delayed variables from administrative registers used as auxiliary variables. The variables of interest are estimated at the level of province (NUTS 2).

Keywords: robust estimation, business statistics, outliers

Grażyna Dehnel
Department of Statistics, Poznan University of Economics and Business e-mail: g.dehnel@ue.poznan.pl

Classification with distance-based logistic regression taking into account prediction error

Eva Boj del Val and Teresa Costa Cor

Abstract In Boj et al. (2017) it is showed a procedure for the estimation of prediction error, PE, the square root of mean squared error, MSE, of prediction for distance-based generalized linear models (Boj et al., 2016). Expressions are developed when the general cases of power families of error distributions and of links are used. As a first step, MSE is approximated with the sum of the process variance and of the estimation variance. The estimation variance can be estimated by applying the delta method and/or by using bootstrap. When using bootstrap one is able to obtain an estimation of the distribution of each predicted value. To help us in the knowledge of the randomness of the new predicted values, confidence intervals can be calculated by taking into account the bootstrapped distributions. Now, it is showed the expression of PE for the generalized linear model with Binomial error distribution and logit link function, the logistic regression. Its calculus and usefulness are illustrated to solve the problem of Credit Scoring, where policyholders are classified into defaulters and non-defaulters. They are analyzed two sets of real credit risk data for which the probabilities of default are estimated. Distance-based logistic regression models are fitted using the dbglm function of the dbstats package for R (Boj et al. 2014).

References

Boj, E., A. Caballé, P. Delicado, and J. Fortiana (2014). *dbstats: distance-based statistics (dbstats)*. R package version 1.0.4.

Boj, E., Delicado, P., Fortiana, J., Esteve A., Caballé, A. (2016). Global and local distance-based generalized linear models. *TEST* 25, 170–195.

Boj, E., Costa, T., Fortiana, J. (2017). Prediction error in distance-based generalized linear models. In: Palumbo, F., Montanari, A. and M. Vichi (eds.). *Data science. Innovative developments in data analysis and clustering.* Springer International Publishing, 2017. (to appear, DOI 10.1007/978-3-319-55723-6_15).

Keywords: Distance-based logistic regression; mean squared error; delta method; bootstrap; credit risk

Eva Boj del Val
Department of Mathematics for Economics, Finance and Actuarial Sciences. Faculty of Economics and Business., University of Barcelona e-mail: evaboj@ub.edu

Teresa Costa Cor
Department of Mathematics for Economics, Finance and Actuarial Sciences. Faculty of Economics and Business., University of Barcelona e-mail: tcosta@ub.edu

Multilevel typologies: Classic and symbolic data approaches

José G. Dias

Abstract The increasing amount of available data on the Internet together with new technologies that allow linking and integrating data from heterogeneous sources have put enormous pressure on the development of new analytic frameworks. New types of data may have many distinct characteristics that result from the integration of different layers of hierarchically structured complex systems from micro to macro levels, i.e., observed units are nested within units of higher levels. Classical examples are patients nested within hospitals, students within classrooms, children within families, or employees within firms. Other examples may deal with space and time dependencies. These multilevel structures have been extensively researched in the context of statistical modelling. A typical problem facing researchers using these complex data structures is to find a typology of individual observations or units. If these units were randomly selected from a given population, then traditional clustering (probabilistic or heuristic-driven algorithms) can be applied. In case these units are organized within upper-level units, the independence assumption is violated and results do not take the heterogeneity at the upper-level into account. Alternatively, symbolic data tend to be more focused on the upper-level and aggregate data at lower level. For instance, clustering schools may result from the analysis of the distribution of characteristics of students within. This aggregation ignores the nested structure and heterogeneity at the individual level. This paper discusses the clustering of nested structures. We compare multilevel latent class models that take the nested structure into account with alterative solutions. In particular, we emphasize the advantage of clustering upper-level units without aggregating lower-level units. Results using synthetic and empirical data provide important recommendations on how to handle these complex data sets.

José G. Dias
Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL e-mail: jose.dias@iscte.pt

l1norm SVMS for binary regression

Pedro Duarte Silva

Abstract Assume that a set of given objects partitioned into two well defined groups can be described by the random pairs $\mathbf{z} = (\mathbf{x}, y)$, where $\mathbf{x}$ is a set of attributes and y is a set of group labels. Then, given a training sample of l independent examples, drawn from some unknown but common distribution, and a rich enough Reproducing Kernel Hilbert Space, it follows that the Bayes rule minimizing the misclassification probability can be consistently estimated by the standard two-group classification SVM rule. However, if misclassification costs differ across groups or the training sample examples are drawn separately from the two classes with proportions that do not approach the true *a priori* probabilities, standard SVMs do not approximate Bayes rules and carry no information about the *a posterior* membership probabilities other than the sign of $P(y = +1|\mathbf{x})$–1/2. Nevertheless, non-standard SVMs that minimize a weighted misclassification loss plus a regularization penalty are consistent estimators of the minimal expected cost Bayes rule. Based on these insights, Wang, Shen and Liu (2008) proposed a *l2-norm regularization posterior probability* regression model, obtained by solving a succession of non-standard SVMs with varying weights in the loss function. Here, we will describe a l1-norm *regularization posterior probability SVM* specially adpted for Big Data problems. It is known that *l1-norm* based SVMs are able to handle efficiently much larger data sets than *l2-norm* based SVMs and often have better generalization abilities (Mangasarian and Thompson, 2008). The computational and statistical properties of this proposal will be illustrated by simulation experiments.

References:

Mangasarian, O.L. and Thompson, M.L. (2008), Chunking for massive nonlinear kernel classification. Optimisation Methods and Software. 23, 365-375.

Wang, J., Shen, X. and Liu, Y. (2008), Probability estimation for large margin classifiers. Biometrika 95, 149-167.

Keywords: Support Vector Machines; Categorical Regression; Big Data

Pedro Duarte Silva
Catolica Porto Business Scholl and CEGE Universidade Católica Portuguesa (Porto) e-mail: psilva@porto.ucp.pt

Temporal based discriminant analysis of hyperspectral measurements

Nontembeko Dudeni-Tlhone

Abstract This study explored the application of multilevel models (longitudinal growth models, in particular) to analyse temporal spectral measurements collected from the eight tree species of interest. The main focus was to identify relevant models that could be used to answer the key questions concerning chlorophyll variation in time over for the main subjects (leaves nested within trees across species types). Different growth models with varying complexity levels were fitted in order to answer the relevant research question. Some of the key results showed that variation in REP (chlorophyll concentration indicator) was significant from the onset, with an initial average REP exceeding 705nm (standard error=1.85). This variation increased significantly over time (weekly) by about 0.22 units. A suitable model that could be used as input into a discriminatory model for the species was, therefore, identified.

Nontembeko Dudeni-Tlhone
Council for Scientific an dIndustrial Research (CSIR) e-mail: ndudeni@gmail.com

Detecting disease subtypes by means of cluster independent component analysis (C-ICA) of multi-subject brain data

Jeffrey Durieux, Serge A.R.B. Rombouts, and Tom Frans Wilderjans

Abstract An emerging challenge in the study of brain diseases and mental disorders, like dementia and depression, consists of revealing systematic differences and similarities between subgroups of patients in functional connectivity patterns (FCPs), that is, coordinated activity across brain regions. As such, existing subtypes of the disease may be characterized in terms of FCPs and disease subtypes may get detected which transcend the current diagnostic boundaries and which show a differential development and prognosis of the pathology.
In order to obtain FCP's, researchers often collect resting-state functional Magnetic Resonance Imaging (rs-fMRI) data and analyze this data with Independent Component Analysis (ICA). ICA is a technique that decomposes a multivariate observed signal into a set of underlying independent source signals and a mixing matrix. In an fMRI context, the sources represent spatial maps, which corresponds to FCPs, and the mixing matrix contains the associated time courses.
Analyzing the brain data of each patient separately with ICA has as major drawback that each patient will be characterized by different FCPs, which makes it difficult to detect the systematic differences and similarities in FCPs between (groups of) patients. Therefore, we propose *Cluster Independent Component Analysis(C-ICA)*. The goal of this method is to cluster the patients into homogenous groups based on the similarities and differences in their FCPs. As such, patients allocated to the same cluster are assumed to have similar connectivity patterns, whereas patients belonging to different clusters will be described by different FCPs. This allows a data-driven detection of disease/disorder subtypes based on different FCPs'.
In this presentation, the C-ICA model is proposed, along with an alternating least squares type of algorithm to estimate its parameters. Further, the results of an extensive simulation study to evaluate the performance of the novel C-ICA method are presented. Lastly, the use of C-ICA is illustrated on empirical brain data.

Keywords: clustering; ICA; dementia; unsupervised learning; multi-subject; fMRI data

Jeffrey Durieux
PhD student, Methodology & Statistics Unit, Institute of Psychology, Faculty of Social and Behavioural Sciences, Leiden University, The NetherlandsLeiden Institute for Brain and Cognition, Leiden, The Netherlands e-mail: j.durieux@fsw.leidenuniv.nl

Serge A.R.B. Rombouts
Methodology & Statistics Unit, Institute of Psychology, Faculty of Social and Behavioural Sciences, Leiden University, The NetherlandsLeiden Institute for Brain and Cognition, Leiden, The Netherlands, Full professor e-mail: romboutssarb@fsw.leidenuniv.nl

Tom Frans Wilderjans
Assistant professor, Methodology & Statistics Unit, Institute of Psychology, Faculty of Social and Behavioural Sciences, Leiden University, The NetherlandsResearch Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational e-mail: t.f.wilderjans@fsw.leidenuniv.nl

Using generalized additive models for CNV detection on multi gene panels

Corinna Ernst, Eric Hahnen, Andreas Beyer,and Rita K. Schmutzler

Abstract We present an approach for copy number variation (CNV) detection which is tailored to the challenges of multi gene panel analysis. Our method relies on a generalized additive model (GAM), which models observed read count frequencies as a product of two smooth functions. Input data is assumed to consist of mapped reads originating from $m, m \geq 30$ samples which are captured on the same gene panel, and to be re-aligned around indels and filtered for duplicates. Inter-sample normalization occurs position-wise as proposed by Anders and Huber [1] for the aim of RNA-seq data normalization. Genome-wide generalized additive models (GAMs) have recently been shown to comprise a powerful tool for the identification of ChIP-Seq peaks and genomic regions of aberrant methylation [2]. We present a GAM which models the mean of observed read counts as a product of two smooth functions, namely, a generic background function that contributes to all m samples and a sample-specific smooth function. The latter function is used for final CNV calling. It is assumed to deviate significantly from zero in case a CNV exists.We validated our approach on the diagnostic TruRisk TM gene panel comprising 48 genes known or assumed to be implicated in hereditary breast and/or ovarian cancer. We compared the performance of our method to the performance of two other tools adapted to CNV analysis on targeted sequencing data, namely, CnvHunter and ExomeDepth. Evaluation revealed that our approach achieves sensitivities and specificities higher orclose to the values achieved by existing tools.References[1] S Anders and W Huber. Differential expression analysis for sequence count data. Genome Biology, 11(10), 2010.[2] G Stricker, A Engelhardt, D Schulz, M Schmid, A Tresch, and J Gagneur. GenoGAM: Genome-wide generalized additive models for ChIP-seq analysis. Bioinformatics, 2017. [Epub ahead of print]

Keywords: Generalized additive model; CNV detection

Corinna Ernst
Center for Familial Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of Cologne, Germany e-mail: corinna.ernst@uk-koeln.de

Eric Hahnen
Center for Familial Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of Cologne, Germany e-mail: eric.hahnen@uk-koeln.de

Andreas Beyer
Cellular Networks and Systems Biology (CECAD), University of Cologne, Cologne, Germany e-mail: andreas.beyer@uk-koeln.de

Rita K. Schmutzler
Center for Familial Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of Cologne, Germany e-mail: rita.schmutzler@uk-koeln.de

Type I censored life test of a two-component series system under bivariate Weibull lifetime distribution

Tsai-Hung Fan, She-Kai Ju, and N. Balakrishnan

Abstract A series system fails if any of its components fail. As these components are all from the same system, they may be correlated. In this paper, we consider a series system of two components having a joint bivariate Marshall-Olkin Weibull lifetime distribution under Type I censoring in life testing. It is common have masked data in which the component that causes failure of the system is not observed. In such a tgsituation, we apply the maximum likelihood approach via EM algorithm along with the missing information principle to estimate the parameters and the standard errors of the MLE. Statistical inference on the model parameters and the mean lifetimes and the reliability functions of the system as well as of the components are derived. The proposed method is confirmed by simulation study and applied to a real data set successfully.

Keywords: Type I censored life test, Marshall-Olkin Weibull distribution, series system, masked data, EM algorithm.

Tsai-Hung Fan
Professor, Graduate Institute of Statistics National Central University e-mail: thfanncu@gmail.com

She-Kai Ju
Graduate Student, Graduate Institute of Statistics National Central University e-mail: roger770109@gmail.com

N. Balakrishnan
Distinguished University Professor, Department of Mathematics and Statistics McMaster University e-mail: bala@univmail.cis.mcmaster.ca

Medical back pain: A spectral clustering approach

Joey Fitch and Nazia Khan

Abstract We use a Spectral Clustering algorithm to find clusters among lower back pain symptoms in medical patients and assess outcomes within each cluster. First, we process the different types of data separately. For categorical data, we use weighted disjunctive coding. For numerical data (continuous and ordinal), we rescale each variable individually to a [0,1] interval, where 0 represents the minimum value and 1 represents the maximum value. Binary data is kept in the original 0-1 format. In this way, all types of data are mapped to comparable dimensions. Next, we compute a similarity score between every pair of patients using an adaptation of Pearson correlation. This leads to a full n x n similarity matrix, guaranteed to be square, symmetric, and nonnegative. We then calculate the spectral (eigen) decomposition of this similarity matrix, reducing the dimensionality into an eigenvector subspace (with each dimension scaled by the corresponding eigenvalues). Finally, we perform k-means clustering in this new subspace to find four clusters, where the data should be well-separated by the second through fourth eigenvector dimensions. We highlight the identifying symptoms of each patient cluster by inspecting any variable whose within-cluster average is extraordinarily low or high, relative to the other clusters. Lastly, we explore the distinct recovery outlooks of each cluster, comparing the cluster means for each recovery assessment variable (after centering and variance standardization).

Keywords: Spectral Clustering; Similarity; Eigen Decomposition

Joey Fitch
Full-time M.S. Statistics student and part-time Mathematics lecturer at San Jose State University., San Jose State University e-mail: joseph.fitch@sjsu.edu

Nazia Khan
Master student, Department of Mathematics and Statistics, San Jose State University

Classifier selection and variable importance in random projection ensemble classification

Francesca Fortunato, Laura Anderlucci, and Angela Montanari

Abstract Random Projections (RP) ensemble classifiers allow to improve classification accuracy while extending to the high-dimensional context methods originally developed for low dimensional data. However, reducing *redundancy* and understanding the properties of the variable ranking induced by the RP ensemble classifier are still open issues. In fact, despite such classifiers highly improve the classification accuracy, they do not allow the identification of the variables with the highest discriminative power and their performance could still be enhanced by a suitable selection of a *good* subset of them. With the aim to identify both the most accurate subset of classifiers and the most discriminant input features, in this work we investigated two different directions. One one hand, combining the original idea of using the Multiplicative Binomial Distribution (MBD) as the reference model to describe and predict the ensemble accuracy and an important result on such distribution, we devised a simple forward-selection technique called Ensemble Selection Algorithm (ESA). On the other, inspired by the Random Forest (RF) process for feature selection, we adjusted the RP ensemble classifier so as to keep the information on variable importance. Specifically, we measured the relative importance of each input feature through a specific coefficient, called Variable Importance in Projection (VIP), and then we removed the variables that present the smallest values of such coefficient. Results of applying both the ESA and the VIP criterion in simulated and real data demonstrate that our proposal successfully controls the misclassification rate by using a very small number of individual classifiers and by ranking the features in terms of their discriminative power.

Francesca Fortunato
University of Bologna e-mail: francesca.fortunato3@unibo.it

Laura Anderlucci
University of Bologna e-mail: laura.anderlucci@unibo.it

Angela Montanari
Department of Statistical Sciences, University of Bologna e-mail: angela.montanari@unibo.it

Partial homogeneity models for temporal repeated cross-sectional latent class analysis

Brian Francis and Valmira Hoti

Abstract This talk addresses the problem of assessing change over time in cross-sectional surveys modelled by latent class analysis (LCA). We approach the problem by consideration of measurement invariance, which has produced a hierarchy of latent class models for multiple group LCA (complete homogeneity, structural homogeneity, partial homogeneity and complete heterogeneity) proposed by Clogg and Goodman(1984, 1985) and developed by Kankaras, Moors and Vermunt(2010). These models allow for both the class sizes and the conditional class probabilities to change over time. We extend and develop the hierarchy of measurement invariance models to the situation where the "group" is a continuous variable such as time. We show that the complete heterogeneity model is unrealistic for temporal data. Instead, researchers need to consider a variety of partial heterogeneity models suitable for cross-sectional temporal data and which are interpretable in terms of social change. These include linear partial heterogeneity, polynomial partial heterogeneity and changepoint partial heterogeneity. In the talk, we develop a range of models and apply them to 21 human value items from seven sweeps of the European Social Survey for the UK. Various models of profile change are considered The conclusion is that human values are indeed changing over time, with change focused on one or two specific items.

Keywords: Latent class analysis, repeated cross-sectional data, temporal change

Brian Francis
Professor of Social Statistics, Department of Maths and Statistics, Lancaster University e-mail: b.francis@lancaster.ac.uk http://www.lancaster.ac.uk/maths/about-us/people/brian-francis

Valmira Hoti
Postgraduate student, Lancaster University e-mail: V.Hoti@Lancaster.ac.uk http://www.lancaster.ac.uk/maths/about-us/people/valmira-hoti

Subspace Clustering with the Multivariate-t Distribution

Brian C Franczak

Abstract Clustering procedures suitable for the analysis of very high-dimensional data are needed for many modern data sets. One model-based clustering approach called high-dimensional data clustering (HDDC) uses a family of Gaussian mixture models to model the sub-populations of the observed data, i.e., to perform cluster analysis. The HDDC approach is based on the idea that high-dimensional data usually exists in lower-dimensional subspaces; as such, the dimension of each subspace, called the intrinsic dimension, can be estimated for each sub-population of the observed data. As a result, each of these Gaussian mixture models can be fitted using only a fraction of the total number of model parameters. This family of models has gained attention due to its superior classification performance compared to other families of mixture models; however, it still suffers from the usual limitations of Gaussian mixture model-based approaches. Herein, a robust analogue of the HDDC approach is proposed. This approach, which extends the HDDC procedure to include the multivariate-t distribution, encompasses 28 models that rectify one of the major shortcomings of the HDDC procedure. Our tHDDC procedure is fitted to both simulated and real data sets and is compared to the HDDC procedure using an image reconstruction problem that arose from satellite imagery of Mars' surface

Keywords: Finite Mixture Models; Classification; Clustering; Shifted Asymmetric Laplace

Brian C Franczak
MacEwan University e-mail: franczakb@macewan.ca

Advances in sequential automatic search of subset of classifiers

Luca Frigau, Giulia Contu, and Francesco Mola

Abstract In a classification problem a model (classifier) is used to discriminate observations among two or more classes of a response variable. In literature, several algorithms have been developed to handle the two-class problem. Nonetheless, they can be extended to solve the multi-class problem: examples of specific strategies such as one-against-all, one-against-one and Error-Correcting Output Codes have been developed. The first two methods consist, respectively, in comparing each class versus the rest, and all possible pair of classes individually. The latter strategy, instead, consists of performing many classifiers and applying a voting scheme to decide the correct class. Since probably no multi-class approach generally outperforms the others, the interpretation aspect plays an important role when choosing the classification method. In this paper we refer to a combined algorithm, called Sequential Automatic Search of a Subset of Classifiers (SASSC), that splits a classification problem among C classes into $K < C$ two-classes sub-problems. SASSC has been mainly developed to consider the trade-off between knowledge extraction, and interpretation of the relationships among predictors, and classification accuracy. The main contribution of this study can be summarized as follows: a) alternative criteria for the aggregation of classes and super-classes are considered; b) alternative criteria for the estimation of the response class for unseen (test-set) observation are investigated; c) the performance of SASSC is compared to that obtained from alternative methods used in the framework of multi-class learning using a set of multi-class response datasets.

Keywords: classification; decision trees; multi-class learning; SASSC

Luca Frigau
University of Cagliari e-mail: frigau@unica.it

Giulia Contu
University of Cagliari e-mail: giulia.contu@unica.it

Francesco Mola
University of Cagliari e-mail: mola@unica.it

Visualizing citation networks for assessment of research performance in academic institutions

Tomokazu Fujino, Keisuke Honda, Hiroka Hamada, and Yoshiro Yamamoto

Abstract Academic institutions are required to assess their research performance by using quantitative measures. Counting number of publications from the institutions is a simple index and commonly used in Japan. However, it is also important to assess the quality of the publications. The staff in a department dealing with IR (Institutional Research) need to understand the current situation of the publication from their institution. Visualizing citation networks could be a powerful tool for initially understanding such a situation including the quality of publications through the networks. In this talk, we propose the new framework to construct a visualization of citation networks by using data extracted from WoS (Web of Science), which is a major academic publication database. The framework consists of following three parts: (1) extracting publications released from a specific institution and publications which refer them from the database, (2) clustering publications by estimating their topics using a latent topic model, (3) visualizing citation networks with the topics, year of publication and the other attributes.Extracting publications including citation information must be more burdensome task if it needs publications which have the distance of two or more from the publications of the institution. To solve this problem, we use graph database called "Neo4j", which can store network structures and understand queries based on the relationship among nodes.In the institutions which have many researchers with different disciplines, it is important to perform clustering with publications by their contents because the number of publications and the structure of the citation network is different if the study of the fields is different. We applied a latent topic model for abstracts of each publication to estimate topics of them and used the topic information in a visualization of the citation network.

Keywords: Visualization; Institutional Research; Citation Network

Tomokazu Fujino
Fukuoka Women's University e-mail: fujino@fwu.ac.jp

Keisuke Honda
The Institute of Statistical Mathematics e-mail: khonda@ism.ac.jp

Hiroka Hamada
The Institute of Statistical Mathematics e-mail: hhamada@ism.ac.jp

Yoshiro Yamamoto
Tokai University e-mail: yama@tokai-u.jp

Non-parametric methods for clusters: Estimation of the number of clusters and a comparison test

Andre Fujita

Abstract Clustering is an important tool in biological data investigation. For example, in neuroscience, one major hypothesis is that the presence or not of a disorder can be explained by the differences in how neurons cluster. In molecular biology, genes may cluster in a different manner between controls and patients, or also among different stages or grades of a disease (e.g., cancer). In this context, methodologies to correctly estimate the number of clusters, and also compare the clustering structures become important. Thus, we present the slope statistic to determine the number of clusters, and the analysis of cluster structure variability (ANOCVA) to statistically test the equality of the clustering structures among two or more datasets. Both are data driven approaches based on the silhouette statistic, and use as input, the output of the majority of clustering algorithms. We show the results in intensive Monte Carlo simulations, and illustrate their applications in two large functional magnetic resonance imaging (fMRI) datasets, one of attention deficit hyperactivity disorder (ADHD) and another of autism spectrum disorder (ASD). Both methods are implemented in R and freely available at http://www.r-project.org (packages slope and anocva, respectively).

Keywords: slope statistic; number of clusters; statistical test; anocva

Andre Fujita
Associate Professor, Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo e-mail: fujita@ime.usp.br https://www.ime.usp.br/ fujita/

Information extraction and its visualization using the quantification method from tweet information for a disaster

Takamitsu Funayama, Yoshiro Yamamoto, and Osamu Uchida

Abstract In Japan, natural disasters occur frequently in recent years. SNS such as Twitter is used in the transmission and collection of information in the event of a disaster. In the Great East Japan Earthquake that occurred in March 2011, there were cases that a local government rescued victims by information posted on Twitter. In addition, when someone want to contact the family, we could not use the telephone. We could use the Internet at that time, so we contacted a family using SNS. Some local governments began to transmit information from Twitter by such a case. In this way, Twitter is utilized as a means of information transmission in the event of a disaster. Therefore, various information gathers in Twitter. We think that utilizing Tweet data is useful for decision making and collection of information for victims and local governments. In this research, we examined the method of extracting information at the time of a disaster using Tweet data at the Kumamoto Earthquake that occurred in April 2016. This data regularly was acquired Tweet to be included keywords such as “rescue”, “evacuation”, “misinformation” and so on using Twitter API. We needed high-spec computer because the size of collected the data was big. Therefore, we decided to set the information to be extracted in order to reduce the data size. In this research, we decided to extract information on evacuation. Thus, we analyzed using Tweet including “evacuation”. We did morphological analysis Tweet data, made co-occurrence matrix of nouns and adjectives, visualized the data using correspondence analysis and so on, and extracted words with relevance. Additionally, we extracted Tweet including words with high relevance, examined the information obtained from visualization and the contents of actual Tweet, considered validity of the visualization.

Keywords: Twitter; Information extraction; Text Mining;

Takamitsu Funayama
Graduate School of Science and Technology, Tokai University e-mail: tkmt.funayama@gmail.com

Yoshiro Yamamoto
School of Science, Tokai University e-mail: yama@tokai-u.jp

Osamu Uchida
School of Information Science and Technology, Tokai University e-mail: o-uchida@tokai.ac.jp

Biplots based on principal surfaces

Raeesa Ganey

Abstract Principal surfaces are smooth two-dimensional surfaces that pass through the middle of a p-dimensional data set. They minimize the distance from the data points, and provide on a non-linear summary of the data. The surfaces are non-parametric and their shape is suggested by the data. The formation of a surface is found using an iterative procedure which starts with a linear summary, typically with a principal component plane. Each successive iteration is a local average of the p-dimensional points, where an average is based on a projection of a point onto the surface of the previous iteration. Biplots are considered as extensions of the ordinary scatterplot by providing for more than three variables. When the difference between data points are measured using a Euclidean-embeddable dissimilarity function, observations and the associated variables can be displayed on a non-linear biplot. A non-linear biplot is predictive if information on variables is added in such a way that it allows the values of the variables to be estimated for points in the biplot. Prediction trajectories, which tend to be non-linear are created on the biplot to allow information about variables to be estimated.The goal is to extend the idea of nonlinear biplot methodology onto principal surfaces. The ultimate emphasizing will be on high dimensional data where the nonlinear biplot based on a principal surface will allow for visualization of samples and the predictive variable trajectories.

Keywords: Biplots; Principal surfaces; Nonparametric principal components; Multidimensional scaling

Raeesa Ganey
School of Statistics and Actuarial Science, Associate lecturer, University of the Witwatersrand e-mail: raeesa.ganey@wits.ac.za

Robust clustering for functional data based on trimming and constraints

Luis Angel García-Escudero, Diego Rivera-García, Joaquin Ortega, and Agustín Mayo-Iscar

Abstract Several procedures for curve clustering have been recently proposed. Unfortunately, they are not designed to deal with outlying curves and the presence of (even a very small fraction of anomalous curves) may be extremely harmful for them. Unfortunately, the existence of outlying curves is often the rule rather than exception in many real data analyses. With this idea in mind, we propose a robust model-based clustering method that relies on an "small-ball pseudo-density" for functional data. Robustness follows from the joint application of a data-driven trimming approach and constraints. Apart from trimming, constraints on the scatter parameters in each cluster are also important in order to avoid detecting spurious clusters. A computationally feasible algorithm is available for its practical implementation. An interesting feature of the proposed methodology is its ability to perform clustering and outlier detection simultaneously. The procedure is tested in a simulation study, and is also applied to real data sets.

Keywords: Clustering; Robustness; Functional Data

Luis Angel García-Escudero
Dpto. Estadística e I.O. Profesor Titular de Universidad, Universidad de Valladolid. ES Q4718001C e-mail: lagarcia@eio.uva.es

Diego Rivera-García
Centro de Investigación en Matemáticas, Guanajuato e-mail: driver@cimat.mx

Joaquin Ortega
Centro de Investigación en Matemáticas, Guanajuato e-mail: jortega@cimat.mx

Agustín Mayo-Iscar
Universidad de Valladolid. e-mail: agustinm@eio.uva.es

Statistical analysis of the economic growth of the Central and Eastern Europe countries

Eugeniusz Gatnar

Abstract In the paper some classification models have been used to find the drivers of the growth observed in the ten Central and Eastern Europe countries since joining the European Union. in the meantime five of them have adopted the common currency, i.e. Euro, therefore we have also compared the pace of economic growth of the two groups.

Keywords: classification, economic growth, Central and Eastern Europe

Eugeniusz Gatnar
University of Economics in Katowice, National Bank of Poland e-mail: eugeniusz.gatnar@nbp.pl

Irrationality: A new type of error in econometric choice models?

Andreas Geyer-Schulz, Tino Fuhrmann, Marvin Schweizer, and Peter Kurz

Abstract This contribution is motivated by the analysis of a large scale end consumer car configuration data set recently made available by TNS Infratest which resulted in the unexpected discovery of a rather large number of irrational configurations. Irrationality is formally defined as choice behavior which deviates from the behavior predicted byVon Neumann-Morgenstern's expected utility theory.The formal analysis of such deviations traditionally followed three approaches, namely1. the development of non-expected utility theory as e.g. Tversky and Kahneman's cumulative prospect theory which try to explain these paradoxes and deviations from the standard choice axioms,2. the development of non-deterministic approaches to choice under risk and uncertainty that address EU deviations as e.g. McFaddens random utility model, and last but not least,3. by extended models which contain an explicit representation of the process and context of decision-making (as e.g. interactions with social context, ...) of Ben-Akiva et al.In contrast, in the context of car configuration data, we investigate the possibility of modeling and estimatinga rational part worth utility model and to describe irrationality as deviation from rationality. We are especially interested in the nature of such deviations.

Keywords: choice models, preprocessing, irrationality,

Andreas Geyer-Schulz
Institut für Informationswirtschaft und MarketingAbteilung für Informationsdienste und elektronische MärkteKIT, Karlsruher Institut für Technologie (KIT) Institut für Informationswirtschaft und Marketing (IISM) Abteilung für Informationsdienste und elektronische Märkte e-mail: andreas.geyer-schulz@kit.edu

Tino Fuhrmann
Karlsruher Institut für Technologie (KIT) Institut für Informationswirtschaft und Marketing (IISM) Abteilung für Informationsdienste und elektronische Märkte e-mail: Tino.Fuhrmann@student.kit.edu

Marvin Schweizer
Karlsruher Institut für Technologie (KIT) Institut für Informationswirtschaft und Marketing (IISM) Abteilung für Informationsdienste und elektronische Märkte e-mail: Marvin.Schweizer@student.kit.edu

Peter Kurz
TNS Deutschland GmbH,Landsberger Str. 284,D-80687 München, e-mail: Peter.Kurz@tns-infratest.com

Collinear by design. The econometrics of product configuration data

Andreas Geyer-Schulz

Abstract Internet product configurators for end-consumers have been rolled out by all major car manufacturers in the last years. In this contribution we develop the econometric foundations for this new data source:Configuration data results from a closed world representation of the product space in which the consumer can choose his preferred alternative. A closed world model - when completely modelled -implies a collinear data set. Such a data set has a natural singular least squares solution for the manufacturer's price function which consists of price function relative to a "virtual configuration" and an affine linear operator. This solution represents an infinite number of linear dependent price functions. The price function determines the prices of all configurable product variants relative to the price of a default configuration given by the constant of the model. We show how the general solution is linked with the linear price for a specific default configuration for a concrete product and we indicate changes necessary for testing hypothesis on the parameters of the price function. Last, but not least, we discuss possible behavioral implications of selecting specific price functions as suggested by Kahneman's prospect theory.

Keywords: Least Squares, Price function, Product Configuration.

Andreas Geyer-Schulz
Institut für Informationswirtschaft und MarketingAbteilung für Informationsdienste und elektronische MärkteKIT, Karlsruher Institut für Technologie (KIT) Institut für Informationswirtschaft und Marketing (IISM) Abteilung für Informationsdienste und elektronische Märkte e-mail: andreas.geyer-schulz@kit.edu

Association rules and community detection on bi-partite network

Giuseppe Giordano

Abstract An important success-key for network-based approaches is the large availability of relational data arising from static and dynamic database, new data collection methods and on-line exchange of information among social actors. This continuous phenomenon of data recording calls for the emergence of data storage, newly defined statistical treatment, and knowledge extraction from huge datasets.In this paper we cope with the problem of extraction of association rules derived by a very large bi-partite network and derive a general framework to deal with such kind of data. We define a funnel strategy where raw relational data are analyzed as in an affiliation matrix (actors per events) holding transactional data. We show how association rules are derived and evaluated at macro level (the complete graph), meso level (sub-graphs) and micro level (egocentric networks).Association rules mining aims at discovering important relations between item sets in form of frequent items. For instance, an item set can be the set of all products sold in a discount, while the set of buyers is the active set that operate by selecting items from the products set (passive set). Such selections can be easily arranged in a table where each row holds the elements of the active set (customers) and in columns are the passive set items (products).We aim at putting together both elements of Transactional and Relational data in order to define a common framework of analysis that exploits the powerful properties of graph theory and social network analysis read to better extract and exploring association rules from transactional data.

Keywords: block model; graph; social network analysis

Giuseppe Giordano
Associate Professor in StatisticsDepartment of Economics and Statistics, Department of Economics and Statistics University of Salerno Italy e-mail: ggiordan@unisa.it

Identification of patient classes in low back pain data using crisp and fuzzy clustering methods

Alexandre Gondeau

Abstract We have performed a cluster analysis of the low back pain dataset in the framework of the IFCS-2017 data challenge. Because the original data matrix contained missing values, the first part of our analysis concerned the imputation of missing values using the Fully Conditional Specification model. The Local Outlier Factor method was then used to detect and eliminate the outliers. After the data normalization, we removed highly correlated variables from the transformed dataset and carried out a k-means clustering of the remaining variables based on their correlations, i.e., the variables with the highest mutual correlations were assigned to the same cluster. Once the variables were assigned to different clusters, one representative per cluster (i.e., the variable with the highest contribution score to the first principal component) was selected. Among the 13 selected variables, there are representatives of each of the 6 variable domains (contextual factor, participation, pain, psychological, activity and physical impairment), specified as important in the paper by Nielsen et al. (2016). Different clustering methods, including Discriminant Analysis of Principal Components (DAPC), k-means and k-medoids, were then carried out to cluster the reduced low back pain dataset. Consensus solutions, both crisp and fuzzy, were then calculated using the GV3 method. The obtained crisp consensus clustering including 5 classes was described in detail and compared to the meta-data annotation.

Keywords: Low Back Pain Data, Missing Value Imputation, Unsupervised Variable Selection, Variable Clustering, Consensus Clustering.

Alexandre Gondeau
Department of Computer Science – PhD student, Université du Québec à Montréal
 e-mail: gondeau.alexandre@courrier.uqam.ca

Variable selection for classification of multivariate functional data

Tomasz Górecki and Waldemar Wołyński

Abstract New variable selection method is considered in the setting of classification with functional data $\mathbf{X}(t)$ (Ramsay and Silverman (2005), Horvath and Kokoszka (2012)). The variable selection is a dimensionality reduction method which leads to replace the high dimensional process $\mathbf{X}(t)$, with a low-dimensional vector $\mathbf{Y}$ still giving a comparable classification error. The various classifiers appropriate for functional data are used. The proposed variable selection method is based on functional distance covariance (Szekely and Rizzo (2009, 2012)) and is a modification of the procedure given by Kong et al. (2015). The proposed methodology is illustrated on a real data example.

References

HORVATH, L. and KOKOSZKA, P. (2012): Inference for Functional Data with Applications. Springer. New York.

KONG, J., WANG, S. and WAHBA G. (2015): Using distance covariance for improved variable selection with application to learning genetic risk models. Statistics in Medicine 34, 1708-1720.

RAMSAY, J.O. and SILVERMAN, B.W. (2005): Functional Data Analysis, Second Edition. Springer. New York.

SZEKELY, G.J. and RIZZO, M.L. (2009): Brownian distance covariance. Annals of Applied Statistics 3(4), 1236-1265.

SZEKELY, G.J. and RIZZO, M.L. (2012): On the uniqueness of distance covariance. Statistical Probability Letters 82(12), 2278-2282.

Keywords: Multivariate Functional Data; Variable Selection; Functional Distance Covariance; Classification

Tomasz Górecki
Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan e-mail: tomasz.gorecki@amu.edu.pl

Waldemar Wołyński
Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan e-mail: wolynski@amu.edu.pl

Report on cluster analysis of target data set in Cluster Benchmark Data Repository

Michael Greenacre

Abstract The particular path taken here in clustering the data set of back-pain subjects is characterized by extensive use of correspondence analysis and multiple correspondence analysis. This approach has the advantage of treating this set of mixed-scale data at the categorical level, using fuzzy coding to transform the continuous variables into fuzzy categorical ones. Missing values are then easily dealt with as they simply represent additional categories of the variables and can be handled using so-called "subset correspondence analysis", preserving all the cases without missing value imputation. The clustering itself is performed using regular k-means clustering, choosing between different numbers of clusters using a criterion based on the incremental benefit in between-cluster sum-of-squares.

Keywords: Clustering, Correspondence Analysis, Dimension Reduction, Fuzzy Coding, k-Means, Multiple Correspondence Analysis, Subset Analysis

Michael Greenacre
Michael Greenacre is Professor of Statistics in the Economics and Business Department of the Universitat Pompeu Fabra, Barcelona. He has written six books and co-edited four multi-authored volumes around the themes of multivariate analysis, particularly data visualization, correspondence analysis and biplots. Most recently, in December 2016, the third edition of his book Correspondence Analysis in Practice has been published by Chapman & Hall / CRC Press. His present research is in statistical methods in ecology, and he collaborates in various projects in the Arctic., Universitat Pompeu Fabra e-mail: michael.greenacre@upf.edu http://www.econ.upf.edu/ michael

Hybrid feature selection method for high-dimensional data sets

Asma Gul, Zardad Khan, Werner Adler, and Berthold Lausen

Abstract Complex data, such as high-dimensional data sets, including a large number of non-informative features is challenging to be dealt with by standard classification methods. Combining multiple classifiers, known as ensemble methods, can give substantial improvement in prediction performance of classifiers while dealing with such type of data (Hothorn et al., 2004, Adler et al., 2016). We propose a technique, by extending Ensemble of Subset of kNN (ESKNN) classifiers (Gul et al., 2016), of hybrid feature selection to select a set of informative features. In the first step, we discuss to use criteria as Gini index and two-sample-statistics simultaneously to rank the features and a proportion of the ranked features are selected. In the next step, ESkNN is developed on this reduced feature set. In the ESkNN classifiers based upon their individual performance are selected and then are combined sequentially starting from the best model and assessed for their collective performance. We used bench mark data sets with some added non-informative features and microarray data sets for the evaluation of our method. The proposed method is compared with kNN, bagged kNN, random kNN, multiple feature subset method, random forest, support vector machines and random projection ensembles. Experimental comparisons reveal that our proposed method gives better classification performance than the usual kNN and its ensembles, and performs comparable to established methods. Moreover, the proposed embedded feature selection method improves the classification performance of random forest and support vector machines.

References:

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Troger, M. (2004): Bagging Survival Trees. *Statistics in Medicine*, 23(1), 77-91.

Adler, W., Gefeller, O., Gul, A., Horn, F. K., Khan, Z., & Lausen, B. (2016). Ensemble Pruning for Glaucoma Detection in an Unbalanced Data Set.*Methods of Information in Medicine*, 55(6), 557-563.

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W., & Lausen, B. (2016). Ensemble of a subset of kNN classifiers. *Advances in Data Analysis and Classification*, 1-14.

Gul, A., Khan, Z., Miftahuddin, M., Adler, W., Perperoglou, A., Lausen, B. (2016), Ensemble of k-nearest neighbour classifiers for class membership probability estimation. In: Wilhelm, A., Kestler, H. A. (eds.), Analysis of Large and Complex Data, European Conference on Data Analysis, Bremen, July, 2014, Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag Berlin, ISBN 978-3-319-25224-7, 411-421.

Asma Gul
Department of Statistics, Assistant Professor, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan. e-mail: gulasma24@gmail.com

Zardad Khan
Department of Statistics, *Abdul Wali Khan University, Mardan, Pakistan* e-mail: zardadkhan@awkum.edu.pk

Werner Adler
Department of Biometry and Epidemiology, *University of Erlangen-Nuremberg, Germany* e-mail: werner.adler@fau.de

Berthold Lausen
Department of Mathematical Sciences, *University of Essex, Colchester, UK* e-mail: blausen@essex.ac.uk

Keywords: High-dimensional data, Ensemble methods; Hybrid feature selection method

Two roles of cluster validity measures for clustering network data

Yukihiro Hamasuna, Ryo Ozaki, and Yasunori Endo

Abstract This paper discusses two roles of cluster validity measures for clustering network data: an evaluation measure for network partitions and a merging criterion in agglomerative hierarchical clustering procedure. Clustering network data such as community detection is one of the important topics in network analysis. Louvain method, which is based on local optimization for Modularity is state-of-the-art network clustering method. Modularity is one of the evaluation measures for network partitions and shows better results when edges within a community are dense and edges between communities are sparse. The words "community" and "cluster" are considered as the same concept in clustering literature. Cluster validity measures are used to evaluate cluster partitions in traditional clustering such as k-means and fuzzy c-means. Several cluster validity measures and their extensions have been proposed and actively studied until now. In this paper, the usefulness of cluster validity measures for clustering network data are studied from the two viewpoints based on traditional clustering. One is an evaluation measure for network partitions and the other is a merging criterion in agglomerative hierarchical clustering procedure. First, we show the experimental results that network partitions are evaluated by cluster validity measures. Next, we propose agglomerative hierarchical clustering method for network data based on local optimization for cluster validity measures. The proposed method is considered as similar algorithm by assuming the Louvain method as agglomerative hierarchical clustering based on local optimization for Modularity. The proposed method is able to estimate the optimal number of clusters by tracking cluster-merging process locally. Numerical experiments with several artificial and benchmark datasets are conducted to compare cluster validity measures with Modularity. The effectiveness of cluster validity measures based evaluation and algorithms is shown through numerical experiments.

Keywords: network clustering; cluster validity measures; hierarchical clustering; Modularity; Louvain method

Yukihiro Hamasuna
2009–2010 Research Fellow of the Japan Society for the Promotion of Science 2011–2014 Assistant Professor, Kindai University 2014– Lecturer, Kindai University, Department of Informatics, School of Science and Engineering, Kindai University e-mail: yhama@info.kindai.ac.jp

Ryo Ozaki
Graduate School of Science and Engineering, Kindai University e-mail: 1210370107b@gmail.com

Yasunori Endo
Faculty of Engineering, Information and Systems, University of Tsukuba, e-mail: endo@risk.tsukuba.ac.jp

Cluster validation in multicriterion data clustering

Julia Handl and Emiao Lu

Abstract The use of multicriterion approaches to data-clustering can be advantageous in situations where clustering solutions need to comply with multiple definitions of clustering quality. Concretely, possible applications of such an approach cover two fundamental types of problems:

- Those for which entities are described by multiple incommensurate features, and an aggregation of dissimilarity information prior to clustering is not desirable.
- Those in which a comprehensive description of cluster quality is regarded as important, and good clustering solutions are required to meet a number of criteria including compactness, separation and possible additional aspects of good cluster quality.

Multicriterion approaches to data-clustering typically embrace the principle of Pareto optimality and thus seek to identify the set of optimal trade-off partitions for a given selection of criteria. One of the challenges this introduces for the practical use of such techniques is the difficulty of selecting a final solution from the set of possible candidate solutions, in particular where this set is large.
This problem is clearly one of model selection and, specifically, one of cluster validation. However, the sets of solutions returned by multicriterion approaches obey particular types of properties, and this may be of value in identifying a final representative solution. Our current work considers possible approaches to the selection process that draw upon the various properties of the optimal trade-off sets returned by multicriterion clustering approaches. More concretely, we contrast the mechanisms offered by traditional cluster validation, with application-specific consideration, and solution selection approaches introduced in the field of multicriterion optimization. Here, we provide concrete results regarding the relative performance of a range of such approaches in a specific application context: the clustering of analogous time series, with a view to improving pooling techniques and their forecasting performance. We conclude by considering the possible implications of this work for cluster validation, in general.

Keywords: Cluster Validation; Multicriterion Clustering

Julia Handl
University of Manchester e-mail: julia.handl@mbs.ac.uk http://personalpages.manchester.ac.uk/mbs/julia.handl/Senior Lecturer in Decision Sciences

Emiao Lu
PhD Student, University of Manchester e-mail: emiao.lu@manchester.ac.uk

Data analysis of the cross-resistance rate between antibiotic drugs by nonmetric asymmetric multidimensional scaling

Yasutoshi Hatsuda, Syou Maki, Yasuhiro Ishimaki, Katsuhito Nagai, Sachiko Omotani, Junji Mukai, Masahiko Taguchi, Michiaki Myotoku, and Tadashi Imaizumi

Abstract We developed a web application system of Chans (Charts and antibiogram-making system), which has a capability to compute the cross-resistance rate (*CRR*) of antibiotic drugs through the data of drug sensitivity test donated. Considering the recent serious problems of the drug resistance in conjunction with antibiotic substances, we focused four major bacterial species of *Pseudomonas aeruginosa* (*P. aeruginosa*), *Escherichia coli* (*E. coli*), MRSA, and MSSA. After the data of *CRR* were calculated with the Chans, they were analyzed with the Nonmetric Asymmetric Multidimensional Scaling (NAMS) method. NAMS exhibits the data on a two-dimensional diagram. The data-to-data distance corresponds to the similarity of the data, and the circle size plotted represents the significance of the data. In common to those four bacterial species, antibiotic drugs in the same family were constellated in the positions close to each other. This feature suggests that the similarity of the action mechanism in the drugs be reflected upon the diagram. In case of *P. aeruginosa* and *E. coli*, explanatory axes were proposed. That is, fluoroquinolone family (inhibitor of DNA synthesis) was arranged in the center of the diagram. The families of aminoglycosides, tetracyclines, and macrolides (the inhibitor of protein synthesis) were arranged in the positive region of abscissa axis. The families of carbapenems, cephems, and penicillins (the inhibitor of cell wall synthesis) were arranged in the negative region of abscissa axis. In the data of MRSA and MSSA, some data was resulted in the degeneration on the diagram. Hence explanatory axes were not proposed. The degeneration was tangible when the data of *CRR* became 100 percent. The degeneration on the diagram was not improved simply by the exclusion of the data of 100 percent on the analysis.

Yasutoshi HATSUDA
Faculty of Pharmacy, Osaka Ohtani University e-mail: hatuday@osaka-ohtani.ac.jp

Syou MAKI
Faculty of Pharmacy, Osaka Ohtani University e-mail: makisyo@osaka-ohtani.ac.jp

Yasuhiro ISHIMAKI
Department of Pharmacy, Takanohara Central Hospital e-mail: ishimaki@takanohara-ch.or.jp

Katsuhito NAGAI
Faculty of Pharmacy, Osaka Ohtani University e-mail: nagaika@osaka-ohtani.ac.jp

Sachiko OMOTANI
Faculty of Pharmacy, Osaka Ohtani University e-mail: omotasati@osaka-ohtani.ac.jp

Junji MUKAI
Faculty of Pharmacy, Osaka Ohtani University, Faculty of Pharmacy, Osaka Ohtani University
e-mail: mukaijyun@osaka-ohtani.ac.jp

Masahiko TAGUCHI
Department of Pharmacy, Takanohara Central Hospital e-mail: ishimaki@takanohara-ch.or.jp

Michiaki MYOTOKU
Faculty of Pharmacy, Osaka Ohtani University e-mail: myoutom@osaka-ohtani.ac.jp

Akinori OKADA
Faculty of Sociology, Rikkyo University e-mail: okada@tama.ac.jp

Tadashi IMAIZUMI
School of Management & Information Sciences, Tama University e-mail: imaizumi@tama.ac.jp

Keywords: Cross-resistance; Multidimensional Scaling; Antibiotic Drugs

Cross-national studies of trust: searching for characteristics of trsut based on comparative surveys among eight nations

Fumi Hayashi and Masamichi Sasaki

Abstract Trust in interpersonal relations varies based on the specific comditions of cultures, nation, societies, and regions. Regarding the relationships of responses (using the "three-Item Rosenberg Scale" or "Misanthoropy Measures" composed of three questions and their responses), survey data among eight nations (Japan, Taiwan, U.S., Germany, Russia, the Czech Republic, Finland and Turkey) were analyzed to determine the common and unique characteristics of trust and their inherent relationships.

Keywords: Comparative survey; Trust; Correspondence analysis

Fumi Hayashi
Emeritus ProfessorVisiring Senior Research Fellow at the Institute of Social Sciences of Chuo University, Toyo Eiwa University e-mail: fumihayas@khe.biglobe.ne.jp

Masamichi Sasaki
Emeritus Professor Visiting Senior Reserach Fellow at the Institute of Social Sciences of huo Universiry , Hyogo Kyoiku University e-mail: masasaki@tamacc.chuo-u.ac.jp

Semi-supervised learning for normal populations: A perspective from statistical missing data analysis

Kenichi Hayashi

Abstract It is widely said that semi-supervised learning improves the prediction performance by the incorporation of unlabeled cases in classification problems. Most methods for semi-supervised learning assume that the unlabeled cases are randomly observed in the feature space. In other words, the missing mechanism of partially labeled data are missing completely at random (MCAR). Then the following question arises: does/how semi-supervised learning improve the prediction performance compared to supervised learning? We investigated the effect of mechanisms that generates unlabeled cases from the viewpoint of statistical missing data analysis. The asymptotic relative efficiency is derived and calculated to compare a classifier by semi-supervised learning to one by supervised learning. The numerical evaluations show that semi-supervised learning can improve prediction but harmful situations exist.

Keywords: Classification problem; Missing data analysis; Partially labeled data; Semi-supervised learning

Kenichi Hayashi
Keio University e-mail: hayashi@math.keio.ac.jp

Decisions that are needed when using cluster analysis, and research that helps with making them

Christian Hennig

Abstract There is a bewildering and relentlessly growing number of cluster analysis methods and every user of cluster analysis faces the hard task of choosing an appropriate one for the application of interest. There are more decisions to make. Many methods require tuning (most prominently but not exclusively the specification of the number of clusters), and results can depend strongly on pre-processing decisions such as dissimilarity design, transformation and selection of variables. A user may hope that such decisions can be made in some kind of "objective" manner based on the data alone, but this is an illusion because different characteristics of clusters are required in different applications. In this presentation I will discuss some issues and present some recent research results and open problems concerning potentially, among others, cluster benchmarking and the multivariate comparison of clustering methods, matching clustering methods to research aims, the role of stability, the role of axiomatic and consistency theory and "performance guarantees", and the role of probability models and parametric bootstrap simulations.

Christian Hennig
Department of Statistical Science, University College London e-mail: c.hennig@ucl.ac.uk

Smart simulation and smart experimental design

Tomoyuki Higuchi

Abstract An emergence of big data boosts a development of the machine learning technology (including modern statistics) that makes it easy to realize the new association-based findings. For an example, the homepage for Electronic Commerce site is modified at every moment in an attempt to draw much attentions from customers when some correlations between the detailed design and sales performance are found by the machine learning technique with some "criteria". That is an essential technology for digital marketing. Such a great success with big data analytics is quickly influencing ways of doing a scientific research, and results in a birth of new research fields: Materials informatics. Nowadays, to reduce the time and cost for discovering a new materials, the modern machine learning techniques such as Deep Neural Networks and Sparse modeling are intensively applied to diverse materials data in material science and engineering. It is true that some suggestions and insights can be achieved with machine learning technique, but more efforts to understand materials are definitely required because of an existence of first principle (governing equations). It should be remarked in general that many of information with simple usage of machine learning is nothing more than the association-based rule, i.e., correlation. Statisticians deeply understand that the causality differs from correlation. To employ information beyond existing data plays an important role in a scientific discovery. A natural way of its realization is to conduct many experiments and simulations, but they are usually expensive in time and cost. Then a smart way of conducting experiments/simulations is a key technology in terms of an efficient scientific discovery. Bayesian statistics can provide a platform to introduce heterogeneous information and integrate them. The new methodology within a framework of Bayesian computations, data assimilation and Bayesian optimization, can be regarded as to realize a smart simulation and a smart experimental design. In this talk, we give a brief introduction of data assimilation and Bayesian optimization, and would like to discuss a research direction in statistics to face with the era of big data and machine learning.

Tomoyuki Higuchi
The Institute of Statistical Mathematics e-mail: higuchi@ism.ac.jp

Perfect simple structure estimation via extension of quartimin criterion

Kei Hirose

Abstract In factor analysis model, the penalized maximum likelihood approach using the lasso-type penalty allows the sparse estimation of the factor loadings. Although the lasso-type penalty is useful in many situations, it cannot always estimate a loading matrix that possesses the simple structure. In this paper, we propose to use a quartimin criterion as a penalty of the factor loadings. It is shown that the quartimin criterion allows us to estimate the perfect simple structure when the value of tuning parameter is sufficiently large. Furthermore, some of the estimates of the factor loadings can result in exactly zero by extending the quartimin criterion.

Keywords: Factor analysis; clustering; L1 regularization

Kei Hirose
Associate Professor,Institute of Mathematics for Industry, Kyushu University e-mail: hirose@imi.kyushu-u.ac.jp
http://www.keihirose.com/

Motion-blurred image restoration

Huey-Miin Hsueh

Abstract Camera shake for object movement always lead to occurrence of motion blur in an image. A conventional model for the blurred intensity is expressed by the convolution of the original intensity and a point spread function. A motion-blurred image can be successfully restored if an adequate estimation of the point spread function is obtained. In this study, we consider a linear point spread function, in which there involve two parameters: angle and length. We develop an estimating procedure for the two parameters based on the power spectrum of the blurred image. Through intensive experiments, the proposed method is found to produce more accurate angle detection and more stable length estimation than existing methods.

Keywords: Image restoration, motion blur, point spread function, power spectrum, Radon transform

Huey-Miin Hsueh
Department of Statistics, National ChengChi University, Taipei, Taiwan e-mail: hsueh@nccu.edu.tw

Tensor modeling and sliced inverse regression for tensor data

Su-Yun Huang

Abstract When data collected are naturally tensor- or array-structured, such as images and videos, or when moments of data are considered, such as gene-gene, gene-environment and gene-gene-environment interactions, we need some modeling and analysis techniques that take the array structure of data into account. Simple vectorization of tensor data will lead to a very long vector in an extremely high dimension and will cause great difficulty due to the curse of dimensionality. In this talk I will focus on supervised tensor dimension reduction via the concept of sliced inverse regression. Numerical examples will be presented.

Keywords: dimension reduction; sliced inverse regression; tensor data analysis

Su-Yun Huang
I am a Research Fellow and Professor at Institute of Statistical Science, Academia Sinica, Taiwan. e-mail: syhuang@stat.sinica.edu.tw

Sufficient dimension reduction via random-partition for large-p-small-n problem

Hung Hung

Abstract Sufficient dimension reduction (SDR) is continuing an active research area nowadays. Conventional SDR methods can easily fail in the high dimension and low sample size setting. To overcome the problem of high-dimensionality, some works have been developed by projecting the covariates onto a lower dimensional envelope subspace, where conventional SDR methods can be directly applied. In this article, we propose a new method of envelope estimation. Our estimation scheme combines two popular ideas from high dimensional data analysis: (1) random sketching and integration of multiple sketches, and (2) random-partition of covariates in search for influential ones. We name the proposed SDR "random-partition SDR" (RP-SDR). Comparing with existing methods, RP-SDR is less affected by the selection of tuning parameters. Moreover, the estimation procedure of RP-SDR does not involve the determination of the structural dimension until at the last stage, which makes it more robust in estimating the target of interest.

HUNG HUNG
Institute of Epidemiology and Preventive Medicine, National Taiwan University e-mail: hhung@ntu.edu.tw

A new clustering algorithm via graph connectivity

Ying-Chao Hung, Yu-Feng Li, and Liang-Hung Lu

Abstract We introduce a new clustering algorithm based on the concept of graph connectivity, where each resulting sub-graph corresponds to a cluster with highly similar objects connected by edge. The proposed algorithm is flexible in the sense that the clustering scheme can be tuned by selecting the number of edges (for connecting objects with high similarity) and a set of singular points (for separating objects with low similarity) so as to optimize a well-designed quality index. Further, it has the time complexity of $O(nlogn)$ and space complexity of $O(n^2)$. We show that the proposed algorithm performs well for some benchmark data in terms of clustering accuracy.

Keywords: clustering; graph connectivity

Ying-Chao Hung
Department of Statistics, Professor, National Chengchi University, Taiwan e-mail: hungy@nccu.edu.tw

Yu-Feng Li
Department of Electrical Engineering, Student, National Taiwan University, Taiwan e-mail: chyautai2210@gmail.com

Liang-Hung Lu
Department of Electrical Engineering, Professor, National Taiwan University, Taiwan e-mail: lhlu@cc.ee.ntu.edu.tw

Screening procedure for auxiliary variables in the gaussian mixture model

Shinpei Imori

Abstract The Gaussian mixture model (GMM) plays an important role in the statistical analysis (e.g., cluster analysis). We focus on the estimation of unknown parameters in the GMM. Recently, we sometimes encounter a situation that auxiliary variables (or secondary variables) are observed in addition to the variables of interest (called the primary variables). It is known that estimation accuracy of the unknown parameters for the primary variables can be improved by utilizing the auxiliary variables that have the same labels as the primary variables have. On the other hand, if the auxiliary variables have different labels from those of the primary variables, there is a possibility that the estimation accuracy worsens. Hence, it is required for improving the estimation accuracy to select the best subset of the useful auxiliary variables, and an information criterion for selecting the auxiliary variables has been already proposed in a previous study. However, when the number of the auxiliary variables is not so small, a large number of candidate models can be considered. Therefore, it is infeasible to select the best subset by information criteria among all candidate models. In this paper, we propose a new screening procedure for selecting the useful auxiliary variables, and consider its theoretical properties.

Keywords: Variable Screening; Auxiliary Variables; Gaussian Mixture Model

Shinpei Imori
Osaka University e-mail: imori@sigmath.es.osaka-u.ac.jp

A high-dimensional location-dispersion model with dependent error and its applications to WTA data analysis

Ching-Kang Ing

Abstract Over the past decade, variable selection in high-dimensional regression models has mainly focused on independent data with homogeneous variances, which regrettably preclude many engineering or economic data that not only exhibit evident time series features but also have variances changing over time. Motivated by wafer acceptance test (WAT) data, in this work, we pay attention to a high-dimensional location-dispersion model with short- or long-range dependent error, which simultaneously takes time dependence and heteroscedasticity into account. We propose a new high-dimensional method generalized from the one used by Ing and Lai (2011) and obtain its selection consistency in situations where the location and dispersion components of the model obey some strong sparsity conditions. Numerical simulation studies and real data.

Keywords: High-dimensional location-dispersion model; Selection consistency; Short- and long-memory time Series; Strong sparsity; WTA data

Ching-Kang Ing
Institute of Statistics, National Tsing Hua University e-mail: cking@stat.nthu.edu.tw

Discriminant analysis with categorical predictors: A NSCA-based approach

Alfonso Iodice D'Enza, Angelos Markos, and Francesco Palumbo

Abstract Discriminant analysis when all the predictors are categorical has peculiarities and traditional methods may perform poorly. In particular our attention is focused on the the case of many categorical predictors. In such a case a method that jointly performs dimensionality reduction and discriminant analysis can ensure better performance with respect to an approach based on independent steps.Methods combining dimension reduction and clustering for categorical data typically involve a combination of correspondence analysis (CA). However, this proposal aims to demonstrate that a variant of the non-symmetric correspondence analysis can be profitably adapted to the discriminant analysis problem. In particular, it can be shown that the criterion being maximized is the difference between the class margin distribution and the class distribution conditional to each categorical attribute. The performance of the proposed approach is appraised on both synthetic and real data sets.

Keywords: Discriminant Analysis, Non symmetric Correspondence Analysis, Categorical Data

Alfonso Iodice D'Enza
Universita degli Studi di Cassino e-mail: iodicede@unicas.it

Angelos Markos
Democritus University of Thrace e-mail: amarkos@eled.duth.gr

Francesco Palumbo
Università di Napoli Federico II e-mail: fpalumbo@unina.it http://www.docenti.unina.it/francesco.palumbo

Cluster detection using spatial scan statistic and its new development in large-scale scanning

Fumio Ishioka and Koji Kurihara

Abstract It is very basic and important issue to identify the location where a problem happens geographically, such as the status of infection occurrence at each county or a hazard map of natural disasters. That would be an effective tool to devise measures for safety management and to clarify the cause. In recent years, the spatial scan statistic proposed by Kulldorff(1997) that tests whether there are significant hotspot or coldspot clusters in a specific area are widely used. Because this statistical approach employs a method of finding a cluster by moving variable size window on the whole study area, it is very difficult to test all the potential cluster areas unless the number of study regions is extremely small. Kulldorff used using a circular shaped window to scan the potential cluster area, but it is pointed that a non-circular shaped cluster, such as along a river or a road cannot be detected. To overcome this problem, various methods for efficiently finding them have been proposed so far. We have proposed an echelon's spatial scan statistic that can detect arbitrarily shaped cluster. In this work, we describe models and scanning methods of various spatial scan statistics and introduce examples of applications. In addition, we make mention of a new development for cluster detection using a pattern enumeration algorithm based on a frequent itemset mining.

Keywords: spatial scan statistic; echelon analysis; cluster detection

Fumio Ishioka
Okayama University e-mail: fishioka@okayama-u.ac.jp

Koji Kurihara
Okayama University e-mail: kurihara@ems.okayama-u.ac.jp

Health management using digital signage and activity meter

Kenichiro Ito, Yusuke Kurita, and Tetsuro Ogi

Abstract Health management is an important topic not only for people that have issues, but it is important for also healthy people. Managing health before having issues is actually the most important health management rather than trying to be healthy again after having issues. However, it is hard to notify or gain attention of health management to people who are currently healthy, since healthy people think themselves to be healthy with no need of additional health information. To solve the situation of health information being difficult to inform healthy people, recently, technique of pushing information to users is being proposed using a digital signage. Health digital signage that displays health information first needs to identify the proximate user. In order to push information to users, the digital signage will have to identify the proximate user or the user to push information. Several technical solutions exist for identification which is being proposed mostly by using wireless communication methods like Wi-Fi and Bluetooth. Especially, recent smart phone devices are typically known to have both communication methods equipped, though recent privacy policy has made the Wi-Fi address or Bluetooth address for smart phone difficult to be used for identification.
Therefore, in this paper, we propose the use of the "Activity Meter" itself to work as an identifier. In details, by using the Bluetooth Low Energy method of communication between the digital signage and activity meter, we propose a health management digital signage that pushes information to users to notify or gain attention on health. By using the prototype of the digital signage and prototype of the activity meter, we conducted an experiment to observe how the proposed health management system affect the user's interest to health management, in particular, affects the amount of activity measured by the activity meter.

Keywords: health management; digital signage; Activity meter;

Kenichiro Ito
Graduate School of System Design and Management, Keio University e-mail: kenichiro.ito@sdm.keio.ac.jp

Yusuke Kurita
System Design and Management Research Institute, Keio University e-mail: yusuke.kurita@sdm.keio.ac.jp

Tetsuro Ogi
Graduate School of System Design and Management, Keio University e-mail: ogi@sdm.keio.ac.jp

The influence of household assets on choice to work

Shinsuke Ito

Abstract This research uses individual data from Japan's National Survey of Family Income and Expenditure to examine the impact of residential area and real estate prices on employment i.e. individuals' choice to work. The results shows a significant negative impact of real assets on employment. This result indicates a theoretical possibility that an accumulation of household assets induces non-working. This research also finds that the influence of real assets on employment is different depending on the area. In the Kanto, Kitakyushu and Fukuoka metropolitan areas there is a negative influence of real assets on employment. This result reflects the differences in real estate and land prices in these areas, and suggests that if there is the a negative effect of real assets on employment in metropolitan areas, tax benefits on land and housing could have an adverse effect on the labor market in these areas. There are two qualifications for the results of this analysis. First, land and housing prices are estimates calculated using official microdata that reflect the official land price in three closest locations. Therefore, there is a possibility that the price of real assets used in this research fails to reflect the specific characteristics of owned land and housings. Second, there are gender-based differences when it comes to housing. Male heads of households tend to purchase real estate via housing loans, while in cases where the head of household is female there is a tendency for them to own real estate without debt.

Keywords: Microdata Analysis

Shinsuke Ito
Faculty of Economics, Chuo Univerasity e-mail: ssitoh@tamacc.chuo-u.ac.jp

Tail risk measurement and its application in finance

Krzysztof Jajuga

Abstract Tail risk is a risk related to the so called extreme events, that is events having very low probability of occurrence and very large impact (for example financial loss). The literature proposed different approaches to analyze tail risk and different measures of tail risk, both in univariate and multivariate case.
The paper provides the systematic review of these measures, including the measures of tail dependence, suitable in multivariate case. In addition, the performance of the measures is evaluated by simulation studies, as well as by applying real data coming from financial markets.

Krzysztof Jajuga
Wroclaw University of Economics e-mail: krzysztof.jajuga@ue.wroc.pl

RNAseq data clustering: comaparative study

Smail Jamail, Abderrahmane Sbihi, and Ahmed Moussa

Abstract Recent advances in high-throughput cDNA sequencing (RNA-seq) have revolutionized gene expression profiling. This analysis aims to compare the expression levels of multiple genes between two or more samples, under specific circumstances or in a specific cell to give more knowledge of cellular function. One of the fundamental data analysis tasks for gene expression studies involves data-mining techniques such as clustering and classification. There are many clustering algorithms that can be applied in gene expression experiments, the most widely used are hierarchical clustering and model-based clustering that rely on a model-selection criterion to select the number of clusters. Depending on the data structure, a fitting clustering method has to be used. In this study, we present clustering algorithms and statistical approaches for grouping similar gene expression profiles that can be applied to RNA-seq data analysis. In addition, we discuss challenges in cluster analysis and describe the strength and weakness of each method by comparing their performance. Besides the clustering algorithms, the adoption of a distance metric to measure the similarity between genes is an important parameter for the clustering procedure, we also describe in this paper different distance measure that can be used to cluster RNA-seq gene expression data.

Bibliography

1- Emanuel Weitschek, Giulia Fiscon, Valentina Fustaino, Giovanni Felici and Paola Bertolazzi (2015) Pattern Recognition in Computational Molecular Biology: Techniques and Approaches, pp 347-370
2- Peng Liu, Yaqing Si (2014) Statistical Analysis of Next Generation Sequencing Data, pp 191-217
3- Si Y, Liu P, Li P, Brutnell TP (2014), Model-based clustering for RNA-seq data. Bioinformatics 30(2):197-205
4- Wang N, Wang Y, Hao H, Wang L, Wang Z, Wang J, Wu R (2013), A bi-Poisson model for clustering gene expression profiles by RNA-seq. Brief Bioinform. 15(4):534-41

Keywords: RNA-seq; Cluster analysis; Gene expression

Smail Jamail
Systems and Data Engineering Team, ENSA, Abdelmalek Essaadi University e-mail: jamail@ensat.ac.ma

Abderrahmane Sbihi
Systems and Data Engineering Team, ENSA, Abdelmalek Essaadi University e-mail: sbihi@ensat.ac.ma

AHMED MOUSSA
Systems and Data Engineering Team, ENSA, Abdelmalek Essaadi University e-mail: amoussa@uae.ac.ma

Kernel methods for symbolic data

Woncheol Jang

Abstract Symbolic data have become increasingly popular in the era of big data. In the paper, we consider density estimator and regression for interval-type data, a special case of symbolic data, common in economics and goverment statistics. We propose kernel estimators with adaptive bandwidth to account for variability of each interval. We derive cross-validation and generalized cross-validation bandwidth selectors. The proposed methods are applied to simulation study and real applications to show the performance of the estimator in comparison with some other existing methods.

Keywords: Adaptive bandwidth; Cross-validation; Kernel density estimation; Symbolic data

Woncheol Jang
Seoul National University e-mail: wcjang@snu.ac.kr

Copula selection in random coefficient degradation models

Shuen-Lin Jeng

Abstract The main goal of this study is to select a proper Copula model for the random effect coefficients in a degradation path model and to estimate the failure proportion at the specific time. The major benefit of the proposed approach is to release the restricted assumption of particular joint distribution forms for the random coefficients, so that the model will has a better goodness-of-fit with data and the estimation of the failure proportion will be more accurate. When the number of coefficients of a degradation path model is more than one and the coefficients are correlated, a typical assumption is the multivariate normal distribution of the coefficients. However, this assumption is restrictive and may be violated in the encountered real cases. The consequence of the improper assumption of distribution may lead to a biased estimation. We use the contour plots of fitted density and a formal goodness-of-fit test to select the proper copula. Several popular copula models are explored, such as, Frank's, Plackett's, and Clayton's models. The marginal distributions considered are normal, lognormal, Weibull and gamma. The study is demonstrated with a degradation data set of the roughness of highway pavement. The failure proportions are estimated based on the selected copula and their confidence intervals are constructed through a bootstrap approach.

Keywords: Copula; Degradation Model; Random Coefficient; goodness-of-fit.

Shuen-Lin Jeng
National Cheng Kung University Department of Statistics e-mail: sljeng@mail.ncku.edu.tw

Applied feasible generalized ridge regression estimation to linear basis function models

Masayuki Jimichi

Abstract The linear basis function models (e.g. Bishop (2006)) have the mean structure with a linear combination of suitable basis functions of some independent variables. We are able to flexibly fit them to the data. The most popular estimation method of the weights in the models is least squares. However, some regularized versions of it are also applied to the over-fitting problem in the fields found in statistical learning and machine learning. We treat the generalized ridge regression (GRR) estimator (cf. Hoerl and Kennard (1970)) in this work. If we use the GRR estimator, the ridge coefficients must be determined. We choose them which minimize a mean square error (MSE) criterion. By reason that the ridge coefficients depend on unknown parameters, we must substitute these estimators to the parameters. If the estimated ridge coefficients are plugged in the appropriate parts of the GRR estimator, we call a feasible generalized ridge regression (FGRR) estimator. In general, it is difficult to obtain the exact moments of the feasible version of the regularized least squares estimators because the estimated ridge coefficients depend on some random variables. We will try to give the exact moments of the FGRR estimator by using the results of previous works (e.g. Jimichi (2016)). They are needed to get the MSE criteria of the FGRR estimator. It is important that they are available to evaluate the accuracy of the FGRR estimator.

References

[1] Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*, Springer-Verlag.
[2] Hoerl, A. E., and R. W. Kennard (1970) Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, Vol. 12, pp. 55-67.
[3] Jimichi, M. (2016) *Shrinkage Regression Estimators and Their Feasibilities*, Kwansei Gakuin University Press.

Keywords: Linear Basis Function Models, Feasible Generalized Ridge Regression Estimator,

Masayuki Jimichi
Kwansei Gakuin University, School of Business Administration e-mail: jimichi@kwansei.ac.jp

Credit scorecard development with support vector machines

Seohoon Jin and Marcel Eduard MBARGA

Abstract With the growing concern of credit risk, lending institutions are focused on developing effective systems that can evaluate and manage credit risk of customers. There is a strong need to measure, monitor, manage and control the financial risk associated with loans. In order to be proactive in preventing delinquency or default by customers (individual or institution), lending institutions need to collect, analyze and classify the different elements comprising a customer's credit. Credit scoring is a technique by which financial institutions develop a numerical score for each applicant so as to reduce the probability of delinquency or payment default by customers. The classic and still widely used approach for evaluating credit worthiness (borrowers' willingness and ability to repay) and risk (potential loss due to any real or perceived change in borrowers' credit worthiness) is based on the building of "scorecards", involving one or a combination of the following techniques: discriminant analysis, linear regression, logistic regression, probit analysis, support vector machines and neural networks. This study discusses procedure on how to use Support Vector Machines (SVM) model to determine the consumer credit score and establish the credit scorecard. In this aim, we first fit a sigmoid function that maps SVM outputs to posterior probabilities of class membership p(class|input) .We then derive marginal probabilities through a log-linear model analysis of a multi-way contingency tables of characteristics' attributes (coarse classes). We further calculate these coarse classes' scores and aggregate them through a simple linear function of scores associated with each characteristic coarse class' value, to produce the final credit score (credit risk) and thus the desired scorecard. A features ranking is further provided based on SVM probabilistic outputs.

Keywords: Credit Score, Support Vector Machines (SVM)

Seohoon Jin
Korea University Sejong Campus, College of Public Policy, Division of Economics and Statistics, Major of Big Data Science e-mail: seohoon@korea.ac.kr

Marcel Eduard MBARGA
Graduate School of Korea University Sejong Campus, Department of Economics and Statistics e-mail: juniormbarga@yahoo.fr

Image processing methods for gaze pattern analysis in marketing research

Jasmin Kajopoulos, Hajime Murakami, Keita Kawasugi, Marin Aikawa, and Kazuhisa Takemura

Abstract In recent years, eye gaze analysis for shopping behavior in real life situations has become increasingly important for marketing research. Thus, numerous techniques have been tried to overcome the complex challenge of linking gaze patterns to behavior. In this paper a novel approach to gaze behavior analysis was proposed. By utilizing already existing image processing tools, we were able to take advantage of the image properties of gaze patterns for our analysis. For that purpose, gaze patterns generated during real life scenarios were first converted into still images, compressing all gaze points over time into one image for each participant. These images were then transformed using image processing methods, such as, 1) Fourier transform, 2) singular value decomposition,3) wavelet analysis and 4) texture analysis methods. By means of these transformations unique features of the composite images could be extracted and further, discriminant analysis could be applied to classify the images into various categories. The goal of this study was to find the most meaningful approach of applying image processing tools to gaze patterns and thus improve the informative value of marketing research. An empirical example was shown and discussed from the perspective of consumer decision theory.

Keywords: Eye Movement; Gaze Behavior; Decision Making; Marketing Research

Kajopoulos
Department of Psychology, Waseda University Ludwig-Maximilians-Universitaet Muenchen Technical University of Munich e-mail: Jasmin.Kajopoulos@tum.de

Hajime Murakami
Department of Psychology, Waseda University e-mail: whassjeidmae@gmail.com

Keita Kawasugi
Department of Psychology, Waseda University e-mail: keita.kawasugi@gmail.com

Marin Aikawa
Department of Psychology, Waseda University e-mail: aikawamarin5630@gmail.com

Kazuhisa Takemura
Department of Psychology, Waseda University e-mail: kazupsy@waseda.jp

Classification and visualization by self-organizing maps and hierarchical cluster analysis

Yo Kameoka and Yoshiro Yamamoto

Abstract In this study, we looked at some classification techniques of compacted data by Self-Organizing Maps (SOM). Candidates of classification are units generated by algorithm of SOM. If we classify appropriately units to some clusters by using hierarchical cluster analysis then we can classify input vectors (i.e. raw data), because each input vector is contained in some one unit. For example, we use a determination method about number of clusters by JD methods (Jain and Dubes, 1988) for determining number of clusters. Even if analysts do not use some determination method about number of clusters, they voluntarily select number of clusters and can examine result of the classification in with reference to the dendrogram. On the other hand, it is difficult to decide how section units on maps of SOM. However, we can discuss results of classification from a different angle by a combination of SOM and hierarchical cluster analysis. In addition, we suggest some classification techniques added considerations of adjacency relationship of units and units without input vector. If two units are unadjacent, then them are not connected even them are similarity. These suggestions lead to results of classification in keeping with characteristic of SOM. Finally, we consider data visualization.

Keywords: SOM; hierarchical cluster analysis; Data Visualization; Two-Staged classification

Yo Kameoka
KSO Corporation e-mail: kameokayo@gmail.com

Yoshiro Yamamoto
School of Science Tokai University e-mail: yama@tokai-u.jp

On robustness to outliers for two abstract fuzzy clustering optimization problems

Yuchi Kanzawa

Abstract Fuzzy c-means is the representative fuzzy clustering algorithm, and various variants have been proposed including the Shannon entropy-regularized fuzzy clustering, the power-regularized fuzzy clustering, and Tsallis entropy-based fuzzy clustering. Robustness to outliers is an important property in clustering tasks, and it is characterized by the membership function value for an extreme object.If the membership function value for an extreme object is crisp: zero or one, such clustering methods is affected by outliers and is not robust.If the membership function value for an extreme object is fuzzy: zero or one, such clustering methods is not so affected by outliers and is robust. Among the above four fuzzy clustering algorithms, the fuzzy c-means and Tsallis entropy-based fuzzy clustering are robust and the other is not robust.For these conventional algorithms, their robustness were analyzed individually.This study investigates two abstract optimization problems for fuzzy clustering in terms of robustness to outliers.Both optimization problems consist of the k-means objective function and an abstract regularizer, each regularizer is different.Specifying such regularizers reduced to that for various fuzzy clustering algorithms.One regularizer is weighted with object-cluster affinity and the other is not.The optimization problem with object-cluster affinity weighted regularizer includes the standard fuzzy c-means and the Tsallis entropy-based fuzzy clustering.The optimization problem without object-cluster affinity weighted regularizer includes the Shannon entropy-regularized fuzzy clustering and the power-regularized fuzzy clustering.It is proved in this study that the object-cluster affinity weighted regularizers produce robust clustering algorithms and the the object-cluster affinity unweighted regularizers not.Therefore, we can see the robustness to outliers of various clustering algorithms only from their corresponding optimization problem.

Keywords: Fuzzy Clustering; Robustness

Yuchi Kanzawa
Shibaura Institute of Technology e-mail: kanzawa@sic.shibaura-it.ac.jp

A hierarchical topic model for the e-commerce purchase behavior

Sotaro Katsumata, Eiji Motohashi, and Akihiro Nishimoto

Abstract This study proposes a model to classify products based on consumer purchase record of an EC (E-commerce) mall. EC malls enable consumers to purchase products in their home and visit many stores without walking. In major EC malls, millions of products are available from thousands of stores and consumers can freely purchase them. Some of products, such as complementary goods, are purchased simultaneously. It is important for EC firms to know complementary and substitute relationship between products to develop customer management programs. In recent years, firms have been trying to find knowledge from purchase record database and applying it to business strategy. However, these purchase record database is a kind of "big data", so it is difficult to obtain some knowledge using traditional quantitative analysis methods. Among many practical models, Latent Dirichlet allocation (LDA) is widely applied because the model can handle high dimensional dataset, and some high speed algorisms are available. LDA assumes latent topics and classify entities such as documents, words, consumers, and products into topics like factor analysis. We can evaluate the fitness of a model and obtain inference of topics. However, the number of topics of optimal model is often too much to interpret for practitioners. Therefore, in this study, we propose a hierarchical topic model that assumes interpretable hyper-topic. We expand the Dirichlet-multinomial regression model by incorporating linear combination term on the prior of the topic parameter and develop the high speed estimation algorism to apply the model to the purchase record database.

Keywords: Latent Dirichlet Allocation; Clustering; E-Commerce

Sotaro Katsumata
Osaka University e-mail: katsumata@econ.osaka-u.ac.jp

Eiji Motohashi
Yokohama National University e-mail: motohashi@ynu.ac.jp

Akihiro Nishimoto
Kwansei Gakuin University e-mail: anishimoto@kwansei.ac.jp

Enumeration algorithms for political districting

Jun Kawahara, Takashi Horiyama, Keisuke Hotta, and Shin-ichi Minato

Abstract In the electoral system of the House of Representatives in Japan, political districts are created by partitioning prefectures. The most important thing for formulating political districts is to reduce the vote-value disparity, which is defined as the ratio of the maximum population of districts to the minimum one. This research formulates the political districting as the graph partitioning problem, which asks to divide a graph into the specified number of connected components. In the previous work (Kawahara et al. WALCOM2017), an algorithm has been proposed, which not only computes the partition with the smallest disparity but also enumerates partition swith disparities smaller than a specified value. The algorithm uses a compact data structure, called zero-suppressed binary decision diagram(ZDD), to represent a huge number of partitions as a compressed expression, and directly constructs a ZDD representing partitions with small disparities. In this research, new experimental results for various graph instances are shown to confirm the performance of the algorithm. The data used in the experiments reflect the actual latest number of political districts to be created and the latest population.By improving the algorithm, ZDDs for partitions can be constructed for graphs larger than ones used in the previous work.

Keywords: Enumeration Algorithms; Zero-suppressed Binary Decision Diagram; Frontier-based Search; Graph partitioning; Political Districting

Jun Kawahara
Nara Institute of Science and Technology e-mail: jkawahara@is.naist.jp

Takashi Horiyama
Saitama University e-mail: horiyama@al.ics.saitama-u.ac.jp

Keisuke Hotta
Bunkyo University e-mail: khotta@shonan.bunkyo.ac.jp

Shin-ichi Minato
Hokkaido University e-mail: minato@ist.hokudai.ac.jp

Using the NSGA optimization algorithm for variable selection in spectral data classification: some lessons learnt

Martin Kidd and Martin Philip Kidd

Abstract The Non-dominated Sorting Genetic Algorithm (NSGA) is a multi-criteria optimization algorithm well suited for applying to variable selection in classification. The NSGA algorithm was applied using Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA) and Support Vector Machines (SVM) as underlying classification methods. In the case of LDA, the number of variables and prediction error were the two criteria optimized by the NSGA. In the case of PLS-DA, three criteria were optimized namely prediction error, number of variables, and number of PLS components. For SVM, prediction error, number of variables and number of support vectors were selected for optimization. The NSGA with PLS-DA method was further extended by selecting blocks of variables using NSGA, instead of selecting individual variables. In this case another criterion for optimization was added namely the block size. Due to the large number of variables (usually in excess of 1000), and the resulting large search space for the NSGA, the possibility of reducing the search space was investigated by doing cluster analysis on the variables using fixed number of clusters and then selecting representative variables from each cluster for input to the NSGA algorithm. The performance of above mentioned techniques were compared with standard PSL-DA (on all the variables) as well as lasso with binomial as underlying distribution. A number of real data sets were used to test the performance by repeatedly selecting random training and test sets and reporting prediction accuracies on the test sets. Results showed that the NSGA based methods in some cases gave similar prediction accuracies to PLS-DA (but with fewer variables in the model). In other cases it outperform the PLS-DA. These results indicated that the NSGA methods could be useful when fitting spectral calibration models.

Keywords: NSGA;Genetic Algorithms;Partial Least Squares;Discriminant Analysis;Lasso;Support Vector Machines

Martin Kidd
Director, Centre for Statistical Consultation, University of Stellenbosch, Centre for Statistical Consultation, University of Stellenbosch, South Africa e-mail: mkidd@sun.ac.za

Martin Philip Kidd
Assistant professor, Dept of Management Engineering, Dept of Management Engineering, Technical University of Denmark,Copenhagen e-mail: martin.philip.kidd@gmail.com

Acquiescence or deep processing? latent class regression for modeling response styles

Kunihiro Kimura

Abstract Latent variable models, such as structural equation models and latent class factor models, have contributed to the modeling of response styles in surveys. A latent class regression model is also promising. In particular, simultaneous estimation of latent classes and their association with covariates helps us to examine whether the greater number of the options endorsed is a manifestation of acquiescence or a result of deeper processing. Analyses of perceived unfairness revealed that respondents with no less than upper secondary education is more likely to belong to the latent class characterized by higher probabilities of endorsement than respondents with lower secondary education when the question is presented in a check-all-that-apply format. This suggests a support for the deeper processing hypothesis rather than the satisfication hypothesis.

Keywords: Latent Class Analysis; Response Strategy; Satisfication

Kunihiro Kimura
Department of Behavioral Science, School of Arts and Letters, Tohoku University e-mail: kkimura@m.tohoku.ac.jp

A note on fuzzified even-sized clustering based on optimization

Kei Kitajima and Yasunori Endo

Abstract Clustering is a method of data analysis without any supervised data. Dividing a dataset into even-sized clusters can be considered in various situations such as K-anonymization and task distribution problem. K-member clustering was proposed as a method focusing on cluster size. The method classifies a dataset into some clusters of which cluster sizes are at least K, and many algorithms was proposed. However, because they are not based on optimization, they sometimes give unnatural results. Therefore, Even-sized Clustering Based on Optimization (ECBO) was proposed as an algorithm based on optimization. ECBO has constraints that each cluster size is K or K+1, and is a method which alternately optimizes the objective function by the simplex method. Some numerical experiments shows that ECBO has higher classification accuracy than other K-member clustering algorithms. However, due to the severe constraint, ECBO is inconvenient in case that a certain margin of cluster size is arrowed. We can consider two ways to solve this problem. The first is to loosen the constraints for cluster size. On this idea, COntrolled-sized Clustering based on optimization (COCBO) has been proposed. COCBO is an algorithm that allows a parameter K to take a certain width of cluster size. Second, we can consider to introduce the concept of fuzzy to membership of each object to clusters. By doing so, the membership can be represented more flexibly and as a result, the constraints become ambiguous.
In this paper, we will propose a new clustering algorithm in the latter way. Concretely, it is an algorithm that minimizes an objective function by alternative optimization for the degrees of membership and cluster centers.Since the objective function is second order to each membership, minimization of the function is performed by the primal-dual path following algorithm.

Keywords: clustering; optimization; even-sized cluster; fuzzification

Kei Kitajima
Student of Graduate School of Systems and Information Engineering, University of Tsukuba., Graduate School of Systems & Information Engineering, University of Tsukuba e-mail: s1720624@s.tsukuba.ac.jp

Yasunori Endo
University of Tsukuba e-mail: endo@risk.tsukuba.ac.jp

Automated speech scoring: A text classification approach

Yuichiro Kobayashi

Abstract The present study aims to automatically evaluate second language (L2) spoken English using automated scoring techniques. Automated scoring is the ability of computer technology to evaluate and score written or spoken productions. It aims to classify a large set of data into a small number of discrete proficiency levels. In automated scoring, objectively measurable features are used as "exploratory variables" to predict scores defined as a "criterion variable." Thus, I have chosen the NICT JLE Corpus, a corpus of 1,281 Japanese EFL (English as Foreign Language) learners, which is coded into nine oral proficiency levels, for the analysis. The nine levels were used as a criterion variable and linguistic features commonly used in the field of corpus linguistics as explanatory variables. Random forests, one of the algorithms for automated scoring, was employed to predict oral proficiency. It is a powerful method for text classification and feature extraction. As a result of random forests with the out-of-bag error estimate, 61.28% of L2 spoken productions were correctly classified. Compared to the baseline accuracy of the simplest possible algorithm of always choosing the most frequent level (37.63%), my random forests model improved prediction by 23.65 points. Predictors that can clearly discriminate oral proficiency levels were prepositions, first person pronouns, adverbs, and contractions in the order of strength. The results of this study can be applied to creating assessments that are more appropriate for scaling oral performances of EFL learners.

Keywords: automated scoring; text classification; random forests

Yuichiro Kobayashi
College of industrial technology, assistant professor, Nihon University e-mail: kobayashi0721@gmail.com

Double robust asymmetric logistic regression model for estimation of global marine stock abundance

Osamu Komori

Abstract The long time-series data on population assessments are essential for global ecosystem assessment because the temporal change of biomass in such a database reflects the status of global ecosystem properly. However, the available assessment data usually have limited sample sizes and the ratio of populations with low abundance of biomass (collapsed) to those with high abundance (non-collapsed) is highly imbalanced. To allow for the imbalance and uncertainty involved in the ecological data, we propose a binary regression model with mixed effects for inferring ecosystem status through an asymmetric logistic model. In the estimation equation, we observe that the weights for the non-collapsed populations are relatively reduced, which in turn puts more importance on the small number of observations of collapsed populations. Moreover, we extend the asymmetric logistic regression model using propensity score to allow for the sample biases observed in the labeled and unlabeled datasets. It robustified the estimation procedure and improved the model fitting.

Keywords: double robust estimation, ecological binary data, mixed effect logistic regression model, propensity score

Osamu Komori
Department of Electrical, Electronic and Computer Engineering, University of Fukui e-mail: komori0926@gmail.com

Proposal of efficient defensive formation and classification the batters it valid

Kazuki Konda and Yoshiro Yamamoto

Abstract In this paper, we propose effective defensive formation of Nippon Professional Baseball League (NPB). Effective defensive formation is the formation that can put the batter out efficiently. In recent years, we began to be able to see the formation for to put the particular batters out efficiently in NPB. However, such a kind of formation is not familiar in NPB, they can be seen on a daily basis in Major League Baseball (MLB). On the contrary, there are many statistically estimated and interesting formation like we are never seen in NPB. First of all, I have studied the formation, which can put grounder batted balls outs efficiently by the in fielders in NPB (Konda and Yamamoto, 2017). As a result, it was found that the formation such as in MLB is valid for NPB. The formation can be estimated from 60 samples of grounder ball that were hit by the batter in the season. It is effective in more than 75 percent of the batters who hit 60 grounder balls, and it could be inferred the total distribution of batted ball of the batter in the season correlation 0.8 or more. Next, I studied how to categorize the less than 25 percent batters not be able to estimate the season of the batter. In other words, I studied about the batters who is not effective to estimate special formation for them. I looked at batting stats of the batters, for example, direction of the batted balls, coordinate of the batted balls caught, and so on. Finally, I studied which batter is effective to estimate the special defensive formation using cluster analysis.

Keywords: cluster analysis; baseball

Kazuki Konda
Graduate school of Science, Tokai University e-mail: k.konda0626@gmail.com

Yoshiro Yamamoto
e-mail: yama@tokai-u.jp

A measuring and modelling way of multicollinearity with Petres' Red indicator

Peter Kovacs and Tibor Petres

Abstract Multicollinearity can be a problem in multiple statistical models. This phenomenon has more than 20 indicators and detection methods. One approach is the Red indicator, which is based on the eigenvalues of the correlation matrix, as eigenvalues' normalized coefficient of variation. The Red indicator can be expressed as a quadratic mean of correlation coefficients between the explanatory variables. We can prove with the help of normalized Herfindahl-index, that the multicollinearity can be expressed as the concentration of eigenvalues. Higher concentration indicates higher level of multicollinearity. The question arises how the multicollinearity can be visualized. A basis method is the examination of the orthogonality of the explanatory variables. As a new approach, the elliptical model of multicollinearity can be formulated on the basis of Red indicator with an m-1 dimensional ellipsoid in the case of m explanatory variables. This presentation shows and compares properties of Red with another indicators, multicollinearity's modeling possibilities.

Keywords: multicollinearity; linear regression, concentration

Peter Kovacs
associate professor, Deputy Dean of General Affairs, Faculty of Economics and Business Administration Chair, Department of Statistics and Demography, Faculty of Economics and Business Administration University of Szeged
e-mail: kovacs.peter@eco.u-szeged.hu

Tibor Petres
associate professor, University of Szeged e-mail: petres@juris.u-szeged.hu

The tale of Cochran's rule

Pieter M. Kroonenberg

Abstract The presentation concerns the numerical evaluation of **Cochran's Rule** about the minimum expected value in $r \times c$ contingency tables with fixed margins when testing independence with Pearson's X^2 statistic using the χ^2 distribution. This rule is quoted, - often imprecisely - in virtually every elementary statistics book.
If we define a *Cochran table* as a contingency table or a pair of margins with $df \geq 2$, all expected values ≥ 1, and 80% or more of all expected values ≥ 5, then *Cochran's Rule* can be formulated as: **For Cochran tables .04 $<\mathbf{p_{X^2.05}}<$ .06** and **.005 $<\mathbf{p_{X^2.01}}<$ .015**. The presentation will provide evidence about the accuracy of his rule, which will shed light on the appropriateness of the use of the continuous χ^2 distribution as an approximation for the discrete distribution of Pearson's X^2.

Pieter M. Kroonenberg
Leiden University & The Three-Mode Company, Leiden, The Netherlands
e-mail: p.m.kroonenberg@fsw.leidenuniv.nl

Visualization and spatial statistical analysis for Vietnam household living standard survey

Takafumi Kubota

Abstract In this study, the data of Vietnam Household Living Standard Survey in 2006 (VHLSS2006) were used as microdata level data to find out some relations among characteristics, medical cares and work styles in Vietnam. VHLSS2006 were aggregated by province level to a spatial data set to find out spatial characteristics and to apply spatial statistical models. The R package shiny dashboard was applied to present the data interactively as dashboards for representative values, tables and maps. In spatial statistical analysis, it was focus on working variables such as working rate, employment rate, self-employed in agriculture or non-agriculture rate in each province. To detect spatial dependence the Moran's I Statistics were applies these working related objective variables and maps of Vietnam. The shape files of Vietnam in GADM database of Global Administrative Areas were used to calculate neighborhood information of provinces in Vietnam and to draw Choropleth map. Conditional Autoregressive model was applied to explain spatial dependences by provinces and characteristics of the relation between working status and characteristics such as gender, age, education and medical cares.

Keywords: Visualization; Spatial Statistics; Micro data

Takafumi Kubota
Tama University e-mail: kubota@tama.ac.jp

Tensor based author name disambiguation as a way of identifying authors

Kei Kurakawa snd Yasumasa Baba

Abstract Author identification of items in a scholarly digital library is a well-recognized necessary function in the proprietary market. Nevertheless, its functionality is not sufficiently provided yet. In digital library research, the issue has been dealt for over fifty years and still being challenged along with the massively changing information environment. A variety of author name disambiguation methods have been proposed for journal publications in a probabilistic approach (Ferreira 2012), on the other hand in recent years identifier assignment for authors, i.e. ORCID has started among a variety of academic stakeholders including journal publishers, academic societies and researchers. These two approaches are complementary to each other, but both of them are not sufficient to provide the production level of the functionality.We tried to take an author name disambiguation approach, in particular with tensor factorization, which has been attracting attention as a technique for data mining, data analysis, and data science (Kolda, 2009;Nickel, 2011). In our method, we represent author similarity matrices of bibliographic citation attributes on tensor slices, and apply tensor factorizations such as CP (CANDECOMP/PARAFAC) decomposition and Tucker decomposition to extract latent feature vectors as of author. Then, we cluster the authors represented by the feature vectors with a clustering technique, k-means to identify the group of citations for each author. We conducted experiment of our approach with bibliographic citation data that we prepared for two author names. To process the data, we used a tensor library "scikit-tensor" and a machine learning library "scikit-learn" and compared effectiveness of our tensor-based model to another model with latent variables LDA (Latent Dirichlet Allocation). As a result, the tensor-based model and LDA showed a better performance for author identification rather than a baseline, i.e. random labeling, but they were similar at the performance level over purity and inverse-purity score of clustering performance measure.

Keywords: Author Identification; Set Theory; Mapping; Author Name Disambiguation; Latent Variables; Tensor Decomposition

Kei Kurakawa
Project Associate Professor, National Institute of Informatics e-mail: kurakawa@nii.ac.jp

Yasumasa Baba
Professor Emeritus, The Institute of Statistical Mathematics e-mail: baba@ism.ac.jp

A robust model selection criterion family and its application for the causal model

Sumito Kurata and Etsuo Hamada

Abstract Since we can not prevent outliers occurring in most practical cases, the robust method is needed when we evaluate a statistical model of a phenomenon. The statistical divergence is often used to select a proper model, by measuring the farness (it is not the statistical distance) between the true distribution and the model. For example, the AIC (Akaike (1974)) is a representative criterion that has the KL-divergence (Kullback and Leibler (1951)) as the origin. However, most information criteria do not suppose that there are some extreme outliers. Kurata and Hamada (2017) generalized the AIC to a family of criteria. The proposed criteria are based on the robust divergence defined by Basu, et al. (1998) and modified by Ghosh and Basu (2013). As with the original divergence family, an advantage of the criteria is the robustness. In this presentation, we report the features of the family of model selection criteria. We also show some applications about the selection of the directed acyclic graphs (DAGs) that express some causal structures. In many cases, conventional methods (e.g. AIC, BIC; Schwarz (1978), and DST; Shipley (2013)) can not perform well when the true causal structure is contaminated by some outliers. While, on the contrary, the proposed criteria keep higher ability to choose the proper structure.

Keywords: model selection; robustness; divergence; causal inference; directed acyclic graph

Sumito Kurata
Graduate School of Engineering Science, Osaka University e-mail: kurata@sigmath.es.osaka-u.ac.jp

Etsuo Hamada
Graduate School of Engineering Science, Osaka University e-mail: hamada@sigmath.es.osaka-u.ac.jp

Statistical evaluation for spatial complexity based on echelon trees

Koji Kurihara, Shoji Kajinishi, and Fumio Ishioka

Abstract The echelons (Myers et al., 1997) are useful techniques to prospect the areas of interest in regional monitoring of a surface variable. The echelons are derived from the changes in topological connectivity with decreasing surface level, based on the areas of relative high and low values of response variables. The echelon approach aggregates the areas in which the values have the same topological structure and makes hierarchically related structure of these areas. An echelon tree is a family tree of echelons with peaks as terminals and the lowest level as root. An echelon tree thus provides a dendrogram representation of surface topology which enables graph theoretic analysis and comparison of surface structures.In order to evaluate the complexity and simplicity of the spatial structure, we investigate the structure of the visualized echelon tree. We consider two approaches. In the case of small-scale spatial data, all patterns are directly calculated and the complexity of the obtained data is evaluated. In the case of large-scale spatial data, most echelon trees are much too complicated for visual study. It is also not possible to calculate all patterns, we evaluate the complexity of the data based on the limbs and bough obtained by the processes such as pruning. A pruning process on an echelon tree recursively divide the tree into an inner set of limbs and an outer set of boughs. In this paper, we propose a method to evaluate the complexity of small and large spatial data. Furthermore, we verify the effectiveness of the method along with some examples.

Keywords: spatial cluster; echelon tree; complexity

Koji Kurihara
Okayama University e-mail: kurihara@ems.okayama-u.ac.jp

Shoji Kajinishi
Okayama University e-mail: pd0q9i51@s.okayama-u.ac.jp

Fumio Ishioka
Okayama University e-mail: fishioka@okayama-u.ac.jp

Initial value selection for the alternating least squares algorithm

Masahiro Kuroda, Yuichi Mori, and Masaya Iizuka

Abstract The alternating least squares (ALS) algorithm is a popular computational algorithm for solving various matrix optimization problems in nonlinear multivariate analysis and nonnegative matrix factorization problems. The algorithm is a simple computational procedure and a stable convergence property, while the algorithm only guarantees local convergence. Then, in order to obtain an optimal solution, initial value selection is deeply related. In this talk, we provide a simple method for finding effectively an initial value of the ALS algorithm. The initial value selection is based on Biernacki et al. (2003) and has three steps consisting of random initialization, short ALS runs and selection. Furthermore, we try to reduce the computational time in the initial value selection using the Aitken accelerator. We present numerical experiments to examine the performance of the initial value selection with/without Aitken accelerator.

Keywords: Alternating least squares algorithm; initial value selection, acceleration

Masahiro Kuroda
Okayama University of Science e-mail: kuroda@mgt.ous.ac.jp

Yuichi Mori
Okayama University of Science e-mail: mori@mgt.ous.ac.jp

Masaya Iizuka
Okayama University e-mail: iizuka@okayama-u.ac.jp

A convex analysis interpretation of fuzzy c-means

Yoshifumi Kusunoki

Abstract Fuzzy c -means is one of the most popular techniques for cluster analysis. Methods of the fuzzy c -means are interpreted as optimization problems, which determine cluster centers (centroids) and the memberships of objects (data points) to minimize the error of the partition defined by the sum of squared distances from objects to the centroids to which they belong. The problems are solved by alternative minimization, where the positions of centroids and the memberships of objects are alternatively and iteratively determined to minimize the error. In this study, we reformulate the optimization problems to problems maximizing max-like objective functions using the convex conjugate (Legendre-Fenchel transformation). They are classified into DC (Difference of Convex functions) programmings, which are optimization problems related to DC functions. Here, a DC function is a function defined by the difference of two convex functions. The DC algorithm developed by Pham Dinh Tao is a method to solve DC programmings. It iteratively minimizes a modified DC programming problem whose concave part is approximated by an affine function at the current solution. We show that the alternative minimization of fuzzy c-means coincides with the DC algorithm by clarifying the equivalence between the memberships of objects and subgradients of max-like objective functions. The differences of fuzzy c-means methods are explained by the differences of max-like objective functions. The effect of parameters of fuzzy c-means methods are also explained by the functions.

Keywords: fuzzy c-means; clustering; convex analysis

Yoshifumi Kusunoki
Osaka University e-mail: kusunoki@eei.eng.osaka-u.ac.jp

Exhaustive relabeling experiments for biomarker selection

Ludwig Lausser, Alexander Groß, and Hans A. Kestler

Abstract Real life objects reflect a whole set of abstract concepts. This eclectic nature can affect the process of learning abstractions such as categorizations. Unaware of the common characteristics of a concept, a learner might be distracted by the patterns of a different one. The quality of a set of training instances is therefore not only determined by the possibility of learning a chosen classification. It is also characterized by its potential for learning distracting categorizations.In this work, we present an exhaustive relabeling test that evaluates leave one out cross-validation experiments for all dichotomies of a dataset. The test setup is designed for the properties of the nearest neighbor type classifiers and utilizes their graph struc- ture for fast and efficient evaluation. We evaluate the method in biomarker selection experiments and try to identify marker combinations that are specific for the proposed dichotomy of a dataset.

Ludwig Lausser
Ulm University e-mail: ludwig.lausser@uni-ulm.de

Alexander Groß
Ulm University e-mail: alexander.gross@uni-ulm.de

Hans A. Kestler
Ulm University e-mail: hans.kestler@uni-ulm.deD

A visualization technique to identify critical variables in multivariate process monitoring

Niel Le Roux

Abstract Biplots can be effectively used in multivariate process monitoring for visualising real-time process behaviour in two or three dimensions by furnishing various principal component related biplots with monitoring ellipses. The early identification of variables affecting the process adversely is of critical importance. Such variables are not necessarily identified in the optimal monitoring biplot with the best overall fit. Therefore, measures are needed to (a) determine the number of suboptimal biplots to consider in addition to the optimal biplot; (b) find the dimensions to use as biplot scaffolding for the various biplots to be displayed; (c) quantify the approximation of each individual variable in every dimension and (d) make a decision in identifying variables which are critical when the process moves towards being out-of-control. Measures considered include the use of permutation testing procedures and in particular measures based on the mean standard prediction error (MSPE), a distance-based criterion and measures based on ratios of fitted sums of squares to total sum of squares i.e. sample predictivity and axis predictivity in pursuing the above challenges. The properties, use and relative advantages in using these different measures are illustrated with data coming from a complicated industrial application where monitoring biplots are used to visualise process behaviour providing in real time information on which variables need to be adjusted when the process enters an out-of-control state.

Keywords: Biplot; Mean standard prediction error; Multivariate process monitoring; Sample and axis predictivity.

Niel Le Roux
Emeritus-Professor, Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch University, South Africa e-mail: njlr@sun.ac.za https://sites.google.com/site/homepageofnieljleroux/

False assignment error rate in discrete latent variable models

Donghwan Lee and Youngjo Lee

Abstract In clustering problem, to model the intrinsic structure of unlabeled data, the latent variable models with discrete random effects are frequently used. These model-based clustering methods often provide the estimation of total misclassification error rate, and corresponding optimal clustering rule. However, in many clustering applications such as disease diagnosis, it is desirable to treat misclassification errors for certain groups differently. In this study, we define the false assignment error rate, a generalized concept of false discovery rate for multiple group clustering and estimate it by using the extended likelihood approach. Some real data examples like Alzheimer's disease PET data and smartphone addiction data illustrate the usage of false assignment error rate estimation and numerical study shows that error controls are consistent as the sample size increases.

Keywords: Clustering; Latent variable model; Mixture

Donghwan Lee
Ewha Womans University e-mail: donghwan.lee@ewha.ac.kr

Youngjo Lee
Seoul National University e-mail: youngjo@stat.snu.ac.kr

Decision tree approaches for interval valued symbolic data response

Seong-Keon Lee

Abstract There have been many methodologies developed about interval valued symbolic data in the field of Statistics, such as regression analysis, principal component analysis, decision trees and so on. However, there is little literature on decision trees for interval valued symbolic response. Most researches are interested in the interval values as independents. In this study, we propose a decision tree for interval valued response data, using a maximum of bivariate normal-lognormal likelihood as the split criterion. The lower and upper bound of intervals can be changed to the lower bound (location) and the range which is greater than zero. Therefore, one of the choice of the distributions of two variables can be bivariate normal-lognormal distribution having correlations of the location and range. To compare the performance of the split criteria, simulation studies and real data applications are presented. As a result, the suggested method would be more efficient than previous ones when the analysts are interested in not only the mean of interval response, but also the correlation of locations and ranges of response.

Seong-Keon Lee
Sungshin Women's University e-mail: sklee@sungshin.ac.kr

Flexible mixtures of factor models using the skew normal distribution

Sharon Lee

Abstract A flexible mixture of skew normal factor analyzers (FMSNFA) is proposed for the modelling and clustering of high-dimensional data that exhibit non-normal distributional features. The approach provides a robust generalization of the traditional mixtures of factor analyzers, where the assumption of normality for the latent factors is relaxed to cater for skewness in the observed data. By adopting a very flexible form of the skew normal distribution as the density for the component latent factors, the FMSNFA model can capture various types of skewness and asymmetry in the data. In particular, its ability in accommodating multiple arbitrary directions of skewness provides a distinct advantage over existing skew factor models, which are typically limited to modelling skewness concentrated in a single direction. As such, it encompasses a number of commonly used models as special cases, including the mixture of restricted skew normal factor analyzers. An ECM algorithm is derived for maximum likelihood estimation of the parameters of the FMSNFA model. The usefulness and potential of the proposed model is demonstrated using both real and simulated datasets.

Keywords: factor analysis, mixture model, clustering, skew distribution

Sharon Lee
Australian Research Council DECRA Fellow at the School of Mathematics and Physics at the University of Queensland, Australia., University of Queensland e-mail: s.lee11@uq.edu.au

A study on predictors of internet and smartphone addiction in adolescents via structural equation modeling

Sohyun Lee, Eunmin Park, and Donghwan Lee

Abstract With the extensive usages and side effects in psychological functioning of internet and smartphone, their addictions in adolescents are now important social issue. This study aims to identify the relevant psychological factors to affect the internet and smartphone addiction in adolescents via structural equation modeling. Using self-diagnosis questionnaires that investigate Young's Internet Addiction Test (Y-IAT) and Smartphone Addiction Scale (SAS), and various clinical and psychological scales such as depression, anxiety, aggression, impulsiveness, behavioral activation system and so on, the data was obtained from 714 middle school students in South Korea. Due to the non-normality and missingness of the addiction variables, data was analyzed based on a robust maximum likelihood method using R package lavaan. After a proper model modification with reliability and validity criterion, the current results of the structural equation modeling suggest that clinical features including aggression, anger expression and ADHD symptom have positive and significant effects on both internet and smartphone addiction. Furthermore, smartphone addiction also positively influences to internet addiction. We discuss limitations and future research scope.

Keywords: Structural equation modeling; Smartphone addiction

Sohyun Lee
Ewha Womans University e-mail: gusthgus@naver.com

Eunmin Park
Ewha Womans University e-mail: hailie.eunmin@gmail.com

Donghwan Lee
Ewha Womans University e-mail: donghwan.lee@ewha.ac.kr

Clinical Decision Support System for HCC using Classification Models

Taerim Lee

Abstract Objectives: The purpose of this paper is to construct and optimize performance of a classification model to aid management of Clinical Decision Support System and to evaluate performance implementation effectiveness and barriers to adoption

Methods: We used 8 classification models including RF, SVM, k-nearest neighbor, linear discriminant analysis, logistic regression, boosting and ANN were trained and their performance were compared for predicting patient's prognosis. Variable selection was performed and only clinical variables relevant to outcomes were utilized for predicting the outcome to avoid over fitting the model and to detect the surveillance clinical exam's significance. After that we plan to compare the performance of RF, ANN and SVM to other classification results using ROC curves. Using several packages of Random forest(package for Random Forest), Support vector machine(package e1071) Shrunken centroid(package pamr) LDA(package sml), KNN(package class), and logistic regression(package stats), Boosting(package boost), ANN(Neural networks package) we can get the results of the classification for clinical support system.

Results: We find that several clinical variables and SNP data were selected as important factor for clinical decision of HCC patient prognosis and manage surveillance clinical exams using 8 classification models. For the comparison of models and methods all models were run evaluation step and validation step using 10 fold cross validation and the accuracy, sensitivity, specificity, positive predictive value and negative predictive value, ROC curves and area under ROC(AUC) with Mann –Whitney test.

Conclusions: Random Forest seems to be overall the best performing model in terms of accuracy and balance of sensitivity and specificity. Using clinical decision support system Physicians can improve their diagnosis and decision making assisted and it could be efficient and cost effective management of medical resources and surveillance clinical exam.

Keywords: clinical decision support system, ROC, surveillance clinical exam

Taerim Lee
Dept. of Information & Statistics, KNOU e-mail: trlee@mail.knou.ac.kr

A Cluster Analysis Using Gridded Temperatures and Precipitation Data in Korea

Yung-Seop Lee, Hee-Kyung Kim, Youngho Lee, Myungjin Hyun, and Jae-Won Lee

Abstract The climatological data of South Korea is observed from about 100 ASOS(Automated Synoptic Observing System) and about 500 AWS(Automatic Weather Station). In order to produce the high quality data, quality control is needed to the observed data. Especially, clustering techniques should be used for the spatial differentiation of the observational stations. By the way, the meteorological data might not be reflected uniformly in the climate characteristic of South Korea because of density difference. The grid data of numerical model, on the other hand, is reflected uniformly because it spaced at 5km×5km apart is distributed evenly. In this study, the temperatures and precipitation data of South Korea are analyzed using K-means clustering method with long-term grid data. Based on the result of gridded data clustering, the automated QC techniques by clusters on meteorological data, which are ASOS and AWS data, can be developed.

Keywords: cluster analysis, gridded meteorological data, automated quality control, K-means algorithm.

Yung-Seop Lee
Department of Statistics, Dongguk University e-mail: yung@dongguk.edu

Hee-Kyung Kim
Department of Statistics, Dongguk University e-mail: khk0228@empas.com

Youngho Lee
KMA National Climate Data Center e-mail: youngkma@korea.kr

Myungjin Hyun
KMA National Climate Data Center e-mail: hyunmj@kma.go.kr

Jae-Won Lee
KMA National Climate Data Center e-mail: jlee99@korea.kr

Privacy preserving em learning for model-based clustering

Kaleb Leemaqz and Sharon Lee

Abstract Privacy is becoming increasingly important in collaborative data analysis, especially those involving personal or sensitive information commonly arising from health and commercial settings. The aim of privacy preserving statistical algorithms is to allow inference to be drawn on the joint data without disclosing private data held by each party. We propose a privacy-preserving EM (PPEM) algorithm, a novel scheme for training mixture models for clustering in a privacy-preserving manner. Here we focus on the case of horizontally distributed data among multiple parties and that cooperative learning is required. More specifically, each party wants to learn the global parameters of the mixture model while preventing the leakage of party-specific information, including any intermediate results that may potentially be traced to an individual party. Another advantage of PPEM is that it does not involve a trusted third party, unlike most existing schemes that implement a master/slave hierarchy. This helps prevent information leakage in the case of a corrupted party. For illustration, PPEM is applied to the widely popular Gaussian mixture model (GMM) and its effectiveness is analysed through a security analysis.

Keywords: EM algorithm, clustering, privacy, mixtue model

Kaleb Leemaqz
University of Queensland e-mail: k.leemaqz@uq.edu.au

Sharon Lee
University of Queensland e-mail: s.lee11@uq.edu.au

Recursive sparse Path analysis via lasso type penalty

Ji Yao Li

Abstract Path analysis (PA) is a commonly used multivariate statistical method for examining the causal processes among observed variables, since the patterns of causation are expressed in terms of path coefficients and diagrams. Path analysis requires explicit instructions before parameter estimation in order to obtain a clear and solvable model, consequently one needs to pre-identify how one variable is related with the others, which is not an easy task in most of the applications. On the other hand, the model itself is extremely sensitive to alternations in casual links. In this presentation I introduce a modified recursive path analysis model for finding necessary links via convex Lasso type penalty, which can effectively avoid user-specifications as well as identify the least necessary paths automatically. A general iterative algorithm is proposed for the maximum likelihood estimation of the parameter matrices. Simulation studies and real data examples show that the proposed model can reliably guarantee a data driven causal structure detection.

Keywords: path analysis; Lasso; Regularization

JI YAO LI
Graduate school of human sciences, Osaka university e-mail: elenli1103@gmail.com

A flexible approach to identify interaction effects between moderators in meta-analysis

Xinru Li, Elise Dusseldorp, and Jaqueline J. Meulman

Abstract Background: Meta-analysis is a valuable tool to quantitatively synthesize findings from multiple studies in a systematic way. It can be used to evaluate the overall outcome (i.e., effect size), and estimate the relationship between study-level covariates (i.e., moderators) and the effect sizes. In many areas, there are often multiple moderators available (e.g., patient characteristics). In such cases, traditional meta-analysis methods often lack sufficient power to investigate interaction effects between moderators, especially high-order interactions. To solve this problem, meta-CART (Dusseldorp et al. 2014) was proposed by integrating Classification and Regression Trees (CART) into meta-analysis. The aim of this study is to improve meta-CART upon two aspects: 1) to integrate the two steps of the approach into one; 2) to consistently take into account the fixed-effect or random-effects assumption in both the splitting and the interaction detection process.
Method: For fixed effect meta-CART, weights were applied and subgroup analysis was adpated. For random effect meta-CART, a new algorithm was developed. The performance of the improved meta-CART was investigated via an extensive simulation study on different types of moderator variables (i.e., dichotomous, ordinal, and multinomial variables), and via an application study.
Results: The simulation results show that the new methods can achieve good control of Type I error (< 0.05) and power (> 0.80) in general. To achieve good recovery rates of moderators (> 0.80), the number of studies needs to be larger than 40 to identify simple interaction effect, and large than 80 to identify complex interaction effects.
Discussion: The improved version of meta-CART applies the fixed- or random-effects assumption consistently in both detection and test procedure. Researchers may choose between fixed- or random-effects model based on their research question and the assumption of residual heterogeneity. The application example shows that meta-CART is able to identify interaction between moderators and provide interpretable results. Knowledge about such interaction effects can be useful for evaluating existing treatments and designing new treatments.

Keywords: meta-anlysis; classification and regression tree; interaction; moderator analysis; fixed effect model; random effects model

Xinru Li
Mathematical Institute, Leiden University e-mail: x.li@math.leidenuniv.nl

Elise Dusseldorp
Institute of Psychology, Leiden University e-mail: elise.dusseldorp@fsw.leidenuniv.nl http://www.elisedusseldorp.nl

Jaqueline J. Meulman
Mathematical Institute, Leiden University e-mail: jmeulman@math.leidenuniv.nl http://www.math.leidenuniv.nl/ jmeulman/

Flexible clustering via mixtures of skew-t factor analysis models

Tsung-I Lin

Abstract The mixture of common t factor analyzers (MCtFA) has been shown its effectiveness in robustifing the mixture of common factor analyzers (MCFA) when handling model based clustering of high-imensional data with heavy tails. However, the MCtFA model may still suffer from a lack of robustness against observations whose distributions are highly asymmetric. In this paper, we present a further robust extension of the MCFA and MCtFA models, called the mixture of common skew-t factor analyzers (MCstFA),by assuming a restricted multivariate skew-t distribution for the common factors. The MCstFA can be used to accommodate severe non-normal (skewed and leptokurtic)random phenomena while preserving its parsimony in factor-analytic representation and perform graphical visualization in low-dimensional plots. A computationally feasibleExpectation Conditional Maximization Either (ECME) algorithm is developed to carryout maximum likelihood estimation. The numbers of factors and mixture components are simultaneously determined based on common likelihood penalized criteria. The usefulness of our proposed model is illustrated with simulated and real datasets, and results signify its better performance over existing MCFA and MCtFA methods.

Keywords: clustering; common factor loadings; data reduction; outliers

Tsung-I Lin
Ph.D. Professor, Department of Applied Mathematics and Institute of Statistics National Chung Hsing University
e-mail: tilin@nchu.edu.tw http://www.amath.nchu.edu.tw/tilin/

Cluster correspondence analysis and reduced K-means: A two-step approach to cluster low back pain patients

Fengmei Liu, Suchara Gupta, and Cristina Tortora

Abstract n response to the IFCS 2017 data challenge regarding low back pain (LBP) patients clustering, we used a two-step approach. Beforehand we preprocessed the data applying feature selection techniques and missing value imputation. On the resulting dataset we applied the two-step approach; in the first step, we grouped the variables into 6 different domains, and performed the domain clustering using cluster correspondence analysis (clusCA). In the second step, we used the Reduced K-means clustering on the combined output variables from each domain. In the result part, we showed the final clustering result of this two-step approach and a profile plot of the clusters. Every cluster is highly interpretable and evaluated well with regard to the first 10 variables in the data.

Keywords: Multiple Correspondence Analysis; Reduced k-Means, Cluster Correspondence Analysis

Fengmei Liu
Master student, Department of Mathematics and Statistics, San Jose State University e-mail: lfm.ustb@gmail.com

Suchara Gupta
Master student, Department of Mathematics and Statistics, San Jose State University e-mail: sucharu7115@gmail.com

Cristina Tortora
Department of Mathematics and Statistics, San Jose State University e-mail: cristina.tortora@sjsu.edu http://www.cristinatortora.com

The changes over time in Kōji Uno's writing style

Xueqin Liu

Abstract This study examined the changes over time in the works of Kōji Uno using statistical methods to stylistically analyze a number of standard features. Kōji Uno, a well-known Japanese litterateur, suffered a mental illness in 1927, and stopped writing for approximately six years while undergoing treatment and recovery. Additionally, at the end of World War II, Uno again suspended his writing for approximately three years owing to the instability of society. It was reported that his literary style had changed when he resumed writing after his recovery from illness. However, there are pros and cons regarding classifying Uno's works into two groups according to before and after the interruption of his writing by illness. In order to understand the changes in Uno's works, a quantitative analysis was therefore employed. The style markers used in this study were the frequency of Chinese characters, nouns, pronouns, proper nouns, and commas. By analyzing these features, we found that Uno's writing style did vary in different periods. Both the frequency of proper nouns and commas increased over time whereas pronoun usage showed a permanent declining trend. Moreover, the frequency of Chinese characters changed during his literary career with the frequency of their use in works written after his illness significantly higher than others. After further consideration, we found that the increase of nouns had a positive impact on the increase of Chinese characters. The enumeration of proper nouns and increase of kana may also have influenced the frequency of commas. These changes are considered to be the factors that have generated different stylistic impressions for readers, resulting in the proposal that Uno' works should be distributed into three groups.

Keywords: Kōji Uno; writing style; quantitative analysis; mental illness

XUEQIN LIU
Graduate School of Culture and Information Science, Doshisha University, Graduate School of Culture and Information Science, Doshisha University e-mail: liuxueqin1024@yahoo.com

Hybrid decomposition for three-way correspondence analysis with nominal-ordinal variables

Rosaria Lombardo, Eric J Beh, and Pieter M Kroonenberg

Abstract Three-way correspondence analysis has been used for the purpose of analysing bivariate and trivariate associations in three-way contingency tables. Classically, this has been done without regard to the ordinal structure of the variables (Carlier & Kroonenberg, 1996; Kroonenberg, 2008, chap. 17, Beh and Lombardo, 2014, chap 11). Recently, Lombardo et al. (2016) proposed a method of decomposition called *trivariate moment decomposition* that is based on the orthogonal basis of Emerson's polynomials for analysing interactions in three-way contingency tables (see Beh and Davy, 1998).
Here, we propose hybrid decompositions for modelling situations where not all variables are ordered. Therefore the association in such three-way contingency tables will be modelled using correspondence analysis based on the components from a three-mode component analysis, such as Tucker3 or PARAFAC (Kroonenberg, 2008), and the polynomial components from the trivariate moment decomposition.
The symmetric association among the three variables is visually portrayed using these different components via an interactive, or nested-mode, polynomial biplot. Polynomial biplots explicitly use the orthogonal polynomials to portray trends among ordinal and nominal variables in contingency tables. The specific characteristics of these polynomials are then used to enhance the interpretation of the structure of the association.

References

Beh, E. J., & Davy, P. J. (1998). Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *The Australian and New Zealand Journal of Statistics*, *40*, 465-477.
Beh, E. J., & Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. Chichester, UK: Wiley. Carlier, A., & Kroonenberg, P. M. (1996). Biplots and decompositions in two-way and three-way correspondence analysis. *Psychometrika*, *61*, 355 – 373.
Lombardo, R., Kroonenberg, P.M. & Beh E.J. (2016). Modelling Trends in Ordered Three-Way Non-Symmetrical Correspondence Analysis. PROCEEDINGS of the 48th scientific meeting of the Italian Statistical Society. (Eds: Monica Pratesi and Cira Perna) ISBN: 9788861970618 Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis.* Hoboken, NJ: Wiley.

Keywords: Three-way Correspondence Analysis; Ordinal and Nominal Categorical Variables; Tucker3 Decomposition; Trivariate Moment Decomposition; Hybrid Decomposition; Polynomial Biplot

Rosaria Lombardo
University of Campania "Luigi Vanvitelli" e-mail: rosaria.lombardo@unicampania.it

Eric J Beh
e-mail: eric.beh@newcastle.edu.au

Pieter M Kroonenberg
e-mail: p.m.kroonenberg@fsw.leidenuniv.nl

Random forests followed by abc analysis as a feature selection procedure for machine-learning

Jorn Lötsch, and Alfred Ultsch

Abstract Data acquisition for biomedical research becomes increasingly complex due to the increasing molecular and clinical knowledge of pathomechanisms of diseases. Data from DNA and other biomolecular measurements are typically high dimensional with variable numbers in the range of 10^2 to 10^6. If an understanding of the biological mechanisms or processes is one of the goals of data analysis, then a rational selection of most informative variables is a necessity. Random forest machine learning employs a multitude of decision trees to learn a highly irregular combination of features [1]. It is usually employed for classifier creation. Common technical implementations such as R-libraries (e.g. [2]) output quantitative measures of the importance of each feature for the overall classification performance. These measures are given as the mean decrease in classification accuracy or in the Gini impurity when the respective variable was excluded from random forest building. An optimal selection of the most informative variables can be achieved by the calculated ABC analysis [3] of the importance measures. ABC analysis is a categorization technique for (positively) skewed distributions. It identifies the most important subset among a larger set of items, aiming at dividing a set of data into three disjoint subsets called “A”, “B” and “C”. Subset “A” comprises the profitable values, i.e., ”the important few” [4], i.e., the features selected for classifier building.

References

1. Breiman, L.: Random Forests. Mach. Learn. 45,**5-32** (2001)
2. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R News 2,**18-22** (2002)
3. Ultsch, A., Lötsch, J.: Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data. PLoS One 10,**e0129767** (2015)
4. Juran, J.M.: The non-Pareto principle; Mea culpa. Quality Progress 8,**8-9** (1975)

Keywords: Feature selection, random forests, biochemical data, biomedcial data

Jorn Lötsch
Professor, Goethe - University, Institute of Clinical Pharmacology; Professor, Fraunhofer Institute for Molecular Biology and Applied Ecology IME, Project Group Translational Medicine and Pharmacology TMP, Germany
e-mail: j.loetsch@em.uni-frankfurt.de

Alfred Ultsch
Professor Data Bionics Research Group, University of Marburg, Germany
e-mail: ultsch@Mathematik.Uni-Marburg.de

A comparison of different applications of functional linear discriminant analysis

Sugnet Lubbe

Abstract Functional linear discriminant analysis is extended to multivariate longitudinal data. Instead of optimally separating two groups of functions, a set of p functions are used to form a linear combination of functions such that these canonical functions are optimally separated between the two groups. Plotting the canonical functions gives the researcher a visual assessment of the degree of separation or overlap between the groups as well as identifying individuals involved in some overlap. The discriminant coefficients can be interpreted in the usual linear discriminant analysis sense, indicating individual variables' contribution to the optimal separation.
A naïve approach to perform separate discriminations at each time point will be compared to methods based on maximising the longitudinal ratio of between group variance to within group variance. A single expression is obtained for the within group variance, but different options can be used for the between group variance: averaging the between group variance over the time points will be compared to using a specific single time point or computing discrimination coefficients at each time point and smoothing the final solution.

Keywords: Discriminant analysis; Functional data analysis

Sugnet Lubbe
Professor in Department of Statistics and Act Sci, Stellenbosch University e-mail: slubbe@sun.ac.za

Benchmarking classification of stock performance by corporate performance measures - Insights from different modelling techniques

Karsten Luebke, Roland Wolf, and Sebastian Sauer

Abstract Numerous corporate performance measures exist as well as numerous studies of the association of these measures with e. g. the corporate stock performance. The results of these studies are very heterogeneous, partly due to the multidimensionality of the problem but also due to the stochastic, random aspects of the measures and their association. But the empirical results also depend on the data preprocessing like e. g. transformations, binning and outlier handling and on the statistical modelling method used. We focus on the different insights gained by different modelling techniques. The methods based on stochastic models are compared to the more algorithmically ones. Additionally to the performance of these methods also the subject matter conclusion that can be drawn from the results are analysed. This is done by the analysis of 12 different accounting and financial market indicators and their distribution and the comparison of the results for 30 companies from Germany, Japan and Poland each for five different years.

Keywords: Benchmark; Classification; Corporate Performance

Karsten Luebke
FOM University of Applied Sciences e-mail: karsten.luebke@fom.de

Roland Wolf
FOM University of Applied Sciences e-mail: roland.wolf@fom.de

Sebastian Sauer
FOM University of Applied Sciences e-mail: sebastian.sauer@fom.de

Analysis of visit record in interviewer mediated surveys: A case study using the survey on the Japanese national character and others

Tadahiko Maeda

Abstract Interviewer-mediated surveys are still most common survey mode in Japan. Here interviewer mediated-survey includes both face-to-face interviewing and self-administered questionnaire delivered by interviewers. In such surveys, in addition to interviewers' observations on the respondent's and household's characteristics etc., interviewer's activity records, which mainly consists of the time of visit and nature of interaction with the household, are obtained. These are examples of survey paradata. Analysis on such data may contribute to the survey quality control, for example, for improving visit schedule. The present study demonstrates examples of such analysis, using data from survey on the Japanese National Character (Nakamura et al., 2017), and others. These survey are conducted in the year from 2013 to 2015 in Japan, and mode includes face-to-face interviewing, self-administered questionnaire, and CAPI (Computer-assisted personal interview). For example, we analyzed the interviewer's visit record on the respondents' residence by type of housing unit, defining some indices for summarizing the outcome of the visit. In this study we classified the visit outcome simply as "Cooperation (interview completed)", "refusal", "continuing visits" "quit". Some of the findings are as follows: 1) Proportion of gaining successful contact with the household member slightly declines after fourth visits to the residence. 2) Proportion of gaining successful cooperation scores maximum at the second visit. 3) Gaining cooperation from residents in an apartment with lockable (security) gate are not so difficult as it is believed among survey takers. Once they are contacted, cooperation rate are not much lower than the residents in an apartment without lockable gate. Implications of such findings on field survey operation will be discussed.

References

Nakamura, T., Yoshino, R., Maeda, T., Inagaki, Y. & Shibai, K. (2017). A Study of the Japanese National Character: The Thirteenth Nationwide Survey (2013) – English Edition –. ISM Survey Research Report No.119.

Keywords: visit record; survey paradata; Japanese National Character

Tadahiko Maeda
The Institute of Statistical Mathematics e-mail: maeda@ism.ac.jp

The effect of an immersive analytical tool on the exploratory analysis of big data

Iwane Maida, Kenichiro Ito, So Sato, Shunichi Nomura, Tetsuya Toma, and Tetsuro Ogi

Abstract In recent years, the use of big data analysis in the discovery of new information has attracted a great deal of attention. In particular, as the condition of one's health is affected in an accumulative manner over the course of many days, big data analysis is well suited to predicting future health conditions by discovering relationships between variables that would be difficult to discern using linear thinking.
However, as big data sets are made up of a large number of variables, it is difficult to grasp the causal relationship between behavior and results. In such a situation, a tool that an analyst can use to intuitively investigate combinations of factors while experimenting with a variety of combinations of variables using a trial and error process is needed.
To assist in resolving this problem, we have developed an immersive visual analytics tool for use in biological log data analysis. This tool combines a statistical package with an immersive, stereoscopic display system, and allows its user to perform interactive analysis. In this study, a statistical specialist used a walking data set as an example to examine the effectiveness of this visualization tool in big data analysis.
The results of this study demonstrated the effects of using a large screen and the stereoscopic function to achieve separation effects in the display of a large number of plots. They also showed the effects that the tool had, using the excellent colors of the video display device, on discovering trends between combinations of variables in color-coded plot groups. Finally, this study confirmed that, by operating the visualization tool that features a built-in statistical tool, it was easier for the statistical analyst to classify "walk pattern" based on healthcare data.

Keywords: Classification; Exploratory analysis; Intuitive analysis; Walk pattern; Big Data

Iwane Maida
Keio University e-mail: iwane2050@gmail.com

Kenichiro Ito
Keio University e-mail: kenichiro.ito@sdm.keio.ac.jp

So Sato
Keio University e-mail: so@sa.to

Shunichi Nomura
Tokyo Institute of Technology e-mail: nomura@is.titech.ac.jp

Tetsuya Toma
Keio University e-mail: t.toma@sdm.keio.ac.jp

Tetsuro Ogi
Keio University e-mail: ogi@sdm.keio.ac.jp

Selection of variables and decision boundaries in functional logistic regression model

Hidetoshi Matsui

Abstract Sparse regularization is one of the most useful tools for variable selection and they are also effective for regression settings where the data are functions. We consider the problem of selecting variables and decision boundaries in logistic regression models for functional data, using the sparse regularization. The functional logistic regression model is estimated by the framework of the penalized likelihood method with the group lasso-type penalty, and then tuning parameters are selected using the model selection criterion. The effectiveness of the proposed method is investigated through real data analysis

Keywords: Functional Data; Logistic Regression; Group Lasso

Hidetoshi Matsui
Faculty of Data Science, Shiga University e-mail: hmatsui@biwako.shiga-u.ac.jp

Classification of tree-valued data and its application to cancer evolutionary trees.

Yusuke Matsui, Satoru Miyano, and Teppei Shimamura

Abstract Multi-regional sequencing provides new opportunities to investigate genetic heterogeneity within or between common tumors from an evolutionary perspective. Several state-of-the-art methods have been proposed for reconstructing cancer evolutionary trees based on multi-regional sequencing data to develop models of cancer evolution. However, there have been few studies on comparisons of a set of cancer evolutionary trees. We propose a clustering method for cancer evolutionary trees, in which sub-groups of the trees are identified based on topology and edge length attributes.Comparison of phylogenetic trees has long been discussed in the context of the evolution of species, and several comparative analytical methods have been developed, however, they are not suitable for cancer evolutions since: (1)The parental node has one child or more than two children (2)The number of nodes varies among patients (3)The label sets of nodes that indicate somatic mutations differ among the patients.To overcome the issues raised in (1)–(3), we develop a method, called tree registration, for transforming tree objects by mapping tree topologies and their attributes to make the trees comparable. The registered trees are embedded in Euclidean space, which enables defining the distance between the cancer evolutionary trees. Based on this distance, we divide a set of the trees into several sub-groups with a clustering method.Simulation showed that the proposed method can detect true clusters with sufficient accuracy. Application of the method to actual multi-regional sequencing data of clear cell renal carcinoma and non-small cell lung cancer allowed for the detection of clusters related to cancer type or phenotype.

Keywords: tree-valued data;clustering;cancer evolutionary trees;bioinformatics

Yusuke Matsui
, Graduate School of Medicine, Nagoya University e-mail: ymatsui@med.nagoya-u.ac.jp

Satoru Miyano
Institute of Medical Science, The University of Tokyo e-mail: miyano@hgc.jp

Teppei Shimamura
Graduate School of Medicine, Nagoya University e-mail: shimamura@med.nagoya-u.ac.jp

A utility model in intertemporal choice that is yielded by introducing psychological time duration

Yutaka Matsushita

Abstract In intertemporal choice, it has been found that if the receipt time is closer to the present, then people tend to grow increasingly or decreasingly impatient. The former or latter situation is recognized as increasing impatience or decreasing impatience, respectively. Nevertheless, most of normative utility models cannot reflect such nonconstant impatience because they are a certain type of the exponential discount function. On the other hand, it was reported that perceiving time according to a logarithmic scale and constantly discounting in terms of this perceived time yields discount impatience. This study, introducing psychological time duration, constructs a weighted additive model in the context of axiomatic measurement theory to reflect nonconstant impatience. Our utility model is expressed as the multiplication of an additive utility of commodities by a weight function of (subjective) durations. This model is constructed as follows. First, we introduce a right action of durations on an extensive structure consisting of commodities. The right action means an advanced operator with an increment in a subjective duration; that is, a commodity multiplied by an increment in a subjective duration on the right is defined as a commodity received at the present time that is equivalent to the commodity received advanced toward the increment. Second, the weighted additive model is obtained as an additive representation of the generalized extensive structure equipped with the right action. Moreover, we show that the utility model is transformed into a different discount function (e.g., the generalized hyperbolic discount function) depending on a scale used in assessing the subjective duration.

Keywords: Extensive structure; Right action; Time preference; Impatience; Stationarity

Yutaka Matsushita
Department of Media Informatics, Professor, Kanazawa Institute of Technology e-mail: yutaka@neptune.kanazawa-it.ac.jp

On clustering longitudinal data

Paul D. McNicholas

Abstract Different approaches for clustering longitudinal data are discussed. While each one is based on a mixture model, they differ in several respects. Some methods are designed to cluster a single variable over time, and are based on multivariate finite mixture models. These approaches are useful, for example, in gene expression time course studies. Methods based on mixtures of matrix variate distributions are also discussed. Such approaches facilitate cluster analyses that consider multiple random variables over time, and are useful for data from a wide range of studies. The option to allow for skewness or heavy tails in the analyses, as well as the use of latent variables, are also discussed.

Keywords: Clustering; longitudinal data; matrix variate; mixture models

Paul D. McNicholas
Professor & University Scholar,Canada Research Chair in Computational Statistics,Department of Mathematics and Statistics., McMaster University e-mail: paulmc@mcmaster.ca http://www.paulmcnicholas.info

On model-based clustering under measurement uncertainty

Volodymyr Melnykov

Abstract Finite mixtures present a powerful tool for modeling complex heterogeneous data. One of their most important applications is model-based clustering. It assumes that each data group can be reasonably described by one of mixture model components. This establishes a one-to-one relationship between mixture components and clusters. In some cases, however, this relationship can be broken due to the presence of observations from the same class that are recorded in various ways. This effect can occur because of recording inconsistencies due to the use of different scales, operator errors, or simply various recording styles. The idea presented in this paper aims to alleviate this issue through modifications incorporated into mixture models. While the proposed methodology is applicable to a broad class of mixture models, we illustrate it on Gaussian mixtures. Several simulation studies and an application to a real-life dataset are considered, yielding promising results.

Keywords: clustering; transformation; measurement uncertainty

Volodymyr Melnykov
University of Alabama e-mail: vmelnykov@cba.ua.edu

Combining individual and aggregated data to investigate the role of socio-economic disparities on cancer burden in Italy

Maura Mezzetti and Francesca Dominici

Abstract An integrated MCMC model was developed to complement detailed exposure information from population surveys in 20 Italian regions with corresponding aggregate-level cancer incidence figures to estimate the relationship between income, as an indicator of socio-economic status, and the risk of breast and lung cancers in Italy. Among western countries, Italy provides an excellent opportunity to investigate variation in site-specific cancer occurrence over different areas. Italian regions show substantial differences in lifestyle and environmental factors, as well as heterogeneous cancer incidence rates across regions. However, the association between socio-economic differences on cancer occurrence in Italy has not been investigated in depth, mainly due to lack of high-quality data in population-based investigations. To take up this challenge we developed a three-stage model to improve ecological level estimates with individual-level exposure estimates, through the combination of diverse sources of data. Cancer registry data provided aggregate-level information on disease incidence figures. Survey data provided information on the distribution of several different risk factors from a sample of individuals over different areas. Evidence from individual-level data from the European Prospective Investigation into Cancer and Nutrition (EPIC) study was also included in the first stage to build the predictive score for the imputation of individual-level information on disease status. At each iteration of the multiple imputation scheme, a complete dataset with individual disease status, income and other exposures was used to estimate the individual level effect. Finally, an ecological level measure was defined by averaging the individual income odds ratio over the population of each region. The main advantage of the proposed model is the possibility to estimate ecological level association while controlling for individual-level confounding and exploiting the heterogeneity in exposure distribution. Results of our study confirm socio-economic disparities for breast and lung cancer incidence in 2005 and in 2013.

Keywords: ecological inference; cancer epidemiology; aggregate data

Maura Mezzetti
Associate Professor, Department of Economics and Finance University "Tor Vergata", Roma, Italy.
e-mail: maura.mezzetti@uniroma2.it

Francesca Dominici
Department of Biostatistics, Harvard T.H Chan School of Public Health, Boston, MA, USA
e-mail: fdominic@hsph.harvard.edu

Modelling sea surface temperature in time effects perspective

M Miftahuddin, Asma Gul, Zardad Khan, Benjamin Hofner, Andreas Mayr, and Berthold Lausen

Abstract One of the most significant components in the model fitting of SST data is time effect, which is investigated in detail of three buoys in the Indian Ocean. Therefore, in order to get a model we incorporated this into the fitting of additive models. Initially, we used three continuous covariates prior to the model fitting, i.e., air temperature, rainfall and humidity. We considered two time covariates, i.e. days of the year (Doy) and the number of days (Nrdays) before and after the gap, and with and without continuous covariates in the model fitting. In this study, we applied linear model, Generalized Additive Models (GAM) and GAM by Boosting (GAMBoost) models with P-splines basis to fit the SST data. We use these model to obtain smoothness on the SST data fitting. For seasonal effects, we get stability of GAM and GAMBoost model fitting with and without continuous covariates compared to annual effect. Pattern and trend for time effects more smoothing than used to GAM model fitting compare with the linear model fitting. Moreover, by using GAM model fitting can be obtained the best model with stability indicator of its time effect. Furthermore, the GAMBoost model fitting is used to optimal explore pattern and trend of continuous covariates.

Keywords: SST; Buoys, GAM and GAMBoost Models; Time Effects; Seasonal and Annual Patterns; Stability

M Miftahuddin
Department of Statistics Faculty of Mathematics and Science, Syiah Kuala University e-mail: miftah@unsyiah.ac.id

Asma Gul
Department of Statistics, Shaheed Benazir Bhutto Women University e-mail: gulasma24@gmail.com

Zardad Khan
Department of Statistics, Abdul Wali Khan University e-mail: zardadkhan@awkum.edu.pk

Benjamin Hofner
Department of Biometry and Epidemiology, University of Erlangen-Nuremberg e-mail: benjamin.hofner@fau.de

Andreas Mayr
Departmemt of Statistics, Ludwig-Maximilians University Munich e-mail: andreas.mayr@stat.uni-muenchen.de

Berthold Lausen
Department of Mathematical Sciences, Faculty of Science and Health, University of Essex e-mail: blausen@essex.ac.uk

Decision diagram-based enumeration techniques and applications for statistical data analysis

Shin-ichi Minato

Abstract Decision diagrams are one of the classical methods for representing and processing many kinds of discrete structures. A Binary Decision Diagram(BDD) is a representation of a Boolean function, one of the most basic models of discrete structures. It was originally developed for efficientBoolean function manipulation required in VLSI logic design. A Zero-suppressed BDD (ZDD) is a variant of the BDD, customized for enumerating and indexing a set of combinations. Recently, ZDDs have been successfully applied not only to VLSI design, but also for solving various combinatorial problems, such as frequent pattern mining, graph enumeration, and statistical data analysis. In this talk, we first explain basic techniques of the decision diagrams used for various combinatorial problems. We also present a brief history of the research activity related to BDDs and ZDDs. We then show an overview of state-of-the-art techniques for efficiently enumerating and indexing the solutions of combinatorial problems. We also present several topics on various applications of those techniques for statistical data analysis and testing.

Keywords: decision diagram; graph enumeration; statistical data analysis

Shin-ichi Minato
Proffesor, Graduate School of Information Science and Technology, Hokkaido University, Hokkaido University e-mail: minato@ist.hokudai.ac.jp

Agglomerative hierarchical clustering with automatic selection of the number of clusters using AIC and BIC

Sadaaki Miyamoto

Abstract Two linkage methods of agglomerative hierarchical clustering having the capability of automatic selection of the number of clusters are studied. These methods determines the number of clusters using the reversals in dendrograms, i.e., the number of clusters when a reversal starts is selected as optimal number of clusters. The two linkage methods use two different Gaussian mixture models. One is a standard Gaussian mixture while the other is a multiplicative model used in X-means. As a result the former leads to a completely new linkage algorithm while the latter is reduced to a variation of the Ward method. Model selection methods using AIC and BIC are considered in these model in order to have reversals. A number of test examples are used to compare effectiveness of these algorithms.

Keywords: agglomerative hierarchical clustering; number of clusters; AIC; BIC

Sadaaki Miyamoto
Professor Emetritus, University of Tsukuba, University of Tsukuba e-mail: sdk.miyamoto@gmail.com

A study on distance between fields and fusion of different fields by using co-authored information of article

Yuji Mizukami

Abstract Innovation is the act of creating new value by using ”new connection”, ”new point of view”, ”new section”, ”new way of thinking”, ”new usage method” and it is regarded as one of the factors driving the development of the world (Schumpeter 1912). In recent years, the promotion of the Innovation has been strongly promoted. In the field of research, attempts are also being made to create new value through connection between those fields. For example, in 2016, the Ministry of Education, Culture, Sports, Science and Technology conducts a survey to encourage integrated research with other fields based on mathematical theory. Moreover, along with the move to promote integration among these research fields, research is being conducted to grasp and promote the degree of them. However, the current situation is that there is no general method for quantitatively indicating the degree of fusion in different fields and the distance between the fields. In this research, for the purpose of providing indices for measuring the degree of them, we show indices quantitatively indicating the degree of fusion in different fields and the distance between the fields. As an example of its application, we showed the degree of integration of different fields and the distance between them at national universities in Japan based on coauthorship information of articles by using the dissertation database Web of Science.

Keywords: Innovation; Institutional Research; Co-authored Information; Network analysis

Yuji Mizukami
College of Industrial Technology, Associate professor, Nihon University e-mail: mizukami.yuuji@nihon-u.ac.jp

Time series classification and clustering using dissimilarities between lagged distributions

Pablo Montero-Manso

Abstract A new dissimilarity measure for classification and clustering of time series is introduced. The proposed dissimilarity considers lagged observations of each series as samples from a multivariate distribution, and then computes the distances between empirical estimates of these distributions. Automatic methods for selecting suitable lags maximizing the discriminatory power are provided. These methods often select a small amount of lags. Compared to other subsequence-based approaches, our procedure is not only able to improve accuracy but also interpretability of results, which is especially interesting in the case of clustering, but also useful for generalizing training results. Reversely, users can also manually choose lags which are meaningful in the specific context of application and then proceed to clustering. The principle of managing a few specific lags aligns more naturally than using whole subsequences with many popular time series models, including stationarity, seasonality or autoregressiveness, but at the same time is more versatile and robust than the model-based approaches. Preliminary results from simulated data and specific study cases show the good performance of the dissimilarity measure. In classification, nearest neighbour approaches based on the proposed dissimilarity outperform the so far best reported methods in some simulated scenarios. Despite its generality, the measure also exhibits a very satisfactory behavior in clustering compared to other more restrictive measures (e.g. based on linear AR models). Specifically, the obtained results are competitive when the regularity assumptions are met and superior when they are not, thus showing a nice property of robustness to the underlying dynamic structure. Furthermore, the low-dimensional output of the presented method allows for the easy introduction/removal of assumptions commonly used in time series such as phase, rotation or shift invariances, and even use data driven approaches for testing their validity

Keywords: classification; time series; distance; dissimilarity; shapelet; subsequence

Pablo Montero-Manso
PHd. Student,Department of Mathematics,Universidade de Coruña, Spain, Universidade da Coruña e-mail: p.montero.manso@udc.es

Effects of retargeting ads in the upper and lower purchase funnel

Takeshi Moriguchi, Guiyang Xiong, and Xueming Luo

Abstract Online marketers have widely adopted retargeting ads to convert customers who had previously browsed the websites or abandoned shopping carts. Yet, the effectiveness of such retargeting remains unclear. This study exploits several randomized field experiments collaborating with an online retailer to test how the effects of retargeting ads vary depending on the ad copy content and purchase funnel stages. The experimental results from the model-free analysis and the probit model suggest that compared to the hold out without retargeting, the retargeting ads in lower funnel based on shopping cart abandonment history can engender significant incremental purchase responses. The effects are driven by the ad content that highlights product return information rather than product reminder or shipping information. Due to the lack of touch, feel, and product trial of online orders, the ad copy with product returns can nudge customers to try the products with reduced online shopping risks and, thus, increase the purchase conversation rates. These results implicate that retargeting ads with customer services of product return information can effectively recover the shopping carts and boost sales revenues.

Keywords: Marketing; Consumer Behavior; E-commerce

Takeshi Moriguchi
Marketing, Waseda University e-mail: moriguchi@waseda.jp

Guiyang Xiong
Marketing, University of Massachusetts at Boston e-mail: gy.xiong@gmail.com

Xueming Luo
Marketing, Temple University e-mail: xueming.luo@temple.edu

Classification and visualization of eye movement data in judgment and decision making studies

Masahiro Morii

Abstract In this study, we would provide example of application of classification and visualization of eye movement data in judgment and decision making studies. We conducted two experiments using eye tracking methodology. One of the experiments was concerning preference judgments, where two Fourier descriptors were represented on a computer screen and participants were asked to choose more attractive shape by key pressing. The other experiment was simulating Web survey, where the questions about their own happiness were presented and participants were asked to answer from Likert-like items by using a computer mouse. Participants' eye movements were recorded during the experiments. We analyzed judgment and decision making processes by using eye movement data as time-series data. Time-series analysis is important in order to examine judgment or decision making processes. Although data acquired in eye tracking are commonly used in judgment and decision making researches, there are several problems or limitation. For examples, eye movement data is mainly categorized as "fixations" and "saccades." Fixations mean pausing or stopping of eye movement, and saccades mean jumping from one fixation point to other. Fixations are defined by duration, velocity, or distance of eye movements. However, there was no standardized definition of fixation these indexes, so this lack makes it difficult to compare results among studies. Fixation data is analyzed by Area Of Interests (AOIs) which indicate certain objects or certain regions as visual targets. The heatmap is one of the visualizing tools used for the distributions of fixations. Although the heatmap represents which area participants focused on with different colors about frequencies of visiting the area and its powerful graphics is easy to understand intuitively, it does not contain any temporal information. In order to solve these problems, we would introduce the techniques of analyzing and visualizing eye movement with serial information.

Keywords: eye movements; judgments; decision-making; fixations

Masahiro Morii
Keio University e-mail: masa.morii@gmail.com

Identification and semiparametric adaptive estimation with nonignorable nonresponse data

Kosuke Morikawa and Jae-kwang Kim

Abstract When the response mechanism is believed to be not missing at random (NMAR), a valid analysis requires stronger assumptions about the response mechanism than do standard statistical methods. Semiparametric estimators have been developed under the correct model specification assumption for the response mechanism and the instrumental variable assumption. In this talk, a new necessary and sufficient condition is proposed to guarantee the model identifiability without using any instrumental variable. Furthermore, we consider a scheme for obtaining the optimal semiparametric estimation for the parameters such as the population mean. Specifically, we propose two semiparametric adaptive estimators that do not require any

Kosuke Morikawa
Graduate School of Engineering Science, Osaka University e-mail: morikawa@sigmath.es.osaka-u.ac.jp

Jae-kwang Kim
e-mail: jkim@iastate.edu

An overview of hybrid latent variable models and how they can be used to address outliers, flexible distributions of latent variables and issues of data dimensionality

Irini Moustaki

Abstract The talk will discuss different specifications and uses of the hybrid latent variable models for categorical observed variables and mixed latent variables (discrete and continuous). Hybrid models have been successfully used to model more than one response strategy. They have also been used to accommodate more flexible distributions for the latent variables and also to detect the correct number of factors in classical factor analysis. Real examples from attitudinal surveys and psychiatric data will be used to illustrate the special features of hybrid latent variable models in research.

Irini Moustaki
Department of Statistics, London School of Economics & Political Science e-mail: I.Moustaki@lse.ac.uk

Validation of sum of squared error clustering: Bootstrapping versus subsampling in view of the influence of multiple points.

Hans-Joachim Mucha

Abstract In simulation studies based on many synthetic and real datasets, we found out that subsampling has a weaker behaviour in finding of the true number of clusters than bootstrapping (Mucha and Bartel 2015). Based on further investigations, here especially concerning the K-means clustering, we were able to find out the most likely reason why bootstrapping outperforms subsampling (mucha and Bartel 2017). Obviously, it is because multiple points give essential additional information in bootstrapping for the investigation of stability of K-means clustering as well as for the determination of the true number of clusters. For the purpose of making fair direct comparisons of the performance of bootstrapping and subsampling, exactly the same subsets of drawn observations were investigated. Concretely, we took a bootstrap sample but díscarded multiple points, and called this special subsampling scheme "Boot2Sub". As a result, the cardinality H of the drawn subsample "Boot2Sub" will vary around 63% of the total sample size, and the serious problem of a necessary specification of H no longer exists.In this paper, we look at the influence of multiple points in more detail when the stability of sum of squared error clustering such as the partitional K-means method and the hierarchical Ward's method is investigated.

References:

Mucha, H.-J. and Bartel, H.-G. (2015): Resampling techniques in cluster analysis: Is subsampling better than bootstrapping? In: B. Lausen, S. Krolak-Schwerdt, and M. Böhmer (Eds.): DataScience, Learning by Latent Structures, and Knowledge Discovery. Springer, Heidelberg, 113-122.

Mucha, H.-J. and Bartel, H.-G. (2017): Validation of K-means clustering: Why ist bootstrapping better than subsampling? Archives of Data Science, Series A, inprint.

Keywords: sum of squared error clustering, validation, bootstrapping, subsampling

Hans-Joachim Mucha
WIAS Berlin e-mail: mucha@wias-berlin.de

Eye movement analysis using image processing methods: for gaze data during decision making between a pair of product images

Hajime Murakami, Keita Kawasugi, Jasmin Kajopoulos, Marin Aikawa, Tokihiro Ogawa, and Kazuhisa Takemura

Abstract The purpose of the present study was to find gaze patterns linked to choice behavior by means of image processing tools and to show an empirical example of these methods. Twenty participants' gaze data were used. The participants were presented with thirty pairs of product images and were asked to choose their preferred product for each pair. One third of stimuli were set in a neutral context (gray background), the other two thirds were set in a meaningful context (background associated with one of the product image pairs). To utilize image processing methods for analysis, gaze movements collected during each trial were first combined into one gaze pattern image. Decomposing these images into features was done via 1) Fourier transform, 2) singular value decomposition, 3) wavelet analysis and 4) texture analysis methods. The extracted features may be used to find a connection between gaze patterns and choice behavior. In the presentation examples will be shown and discussed in terms of consumer research.

Keywords: Eye movement, Gaze behavoir, decision making, consumer research

Hajime Murakami
Department of Psychology, Waseda University e-mail: whassjeidmae@gmail.com

Keita Kawasugi
Department of Psychology, Waseda University e-mail: keita.kawasugi@gmail.com

Jasmin Kajopoulos
Department of Psychology, Waseda University, Technical University of MunichLudwig-Maximilians-Universität, München e-mail: Jasmin.Kajopoulos@tum.de

Marin Aikawa
Department of Psychology, Waseda University e-mail: aikawamarin5630@gmail.com

Tokihiro Ogawa
National Research Institute of Police Science e-mail: t-ogawa@nrips.go.jp

Kazuhisa Takemura
Department of Psychology, Waseda University e-mail: kazupsy@waseda.jp

Generalized orthonormal polynomial principal component analysis

Takashi Murakami

Abstract Orthonormal polynomial principal component analysis (OPPCA; Murakami, 2016; 2017) is simple classical PCA with Harris-Kaiser rotation applied to a set of Likert-type items response categories of which are quantified by slightly revised orthonormal polynomials in advance. The procedure yields the same individual scores as given by standard multiple correspondence analysis, and the solution explains exactly the same proportion of variances. The method is easily modified for analyzing common ordered categorical variables including dummy variables usually used in social survey research.

The modified OPPCA is expected to make it possible to facilitate the use of a higher dimensional solution by axes-based interpretations rather than visual representations on a plane. We will demonstrate a numerical example based on a real data set, and some inferences by Bootstrap method are examined.

Keywords: Principal Component Analysis; Multiple Correspondence Analysis; Orthonormal Polynomials

Takashi Murakami
Department of Sociology, Chukyo University e-mail: tandem06@sass.chukyo-u.ac.jp

Linear time visualization and search in big data using pixellated factor space mapping

Fionn Murtagh

Abstract It is demonstrated how linear computational time and storage efficient approaches can be adopted when analyzing very large data sets. More importantly, interpretation is aided and furthermore, basic processing is easily supported. Such basic processing can be the use of supplementary, i.e. contextual, elements, or particular associations. Furthermore pixellated grid cell contents can be utilized as a basic form of imposed clustering. For a given resolution level, here related to an associated m-adic (m here is a non-prime integer) or p-adic (p is prime) number system encoding, such pixellated mapping results in partitioning. The association of a range of m-adic and p-adic representations leads naturally to an imposed hierarchial clustering, with partition levels corresponding to the m-adic-based and p-adic-based representations and displays. In these clustering embedding and imposed cluster structures, some analytical visualization and search applications are described.

References.

F. Murtagh, "Semantic Mapping: Towards Contextual and Trend Analysis of Behaviours and Practices", in K. Balog, L. Cappellato, N. Ferro, C. MacDonald, Eds., Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, Évora, Portugal, 5–8 September, 2016, pp. 1207–1225, 2016.http://ceur-ws.org/Vol-1609/16091207.pdf

F. Murtagh, Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics, Chapman and Hall, CRC Press, 2017

Keywords: Cluster analysis; Data Science; Big Data analytics; visualization

Fionn Murtagh
Professor of Computer ScienceSecretary (former President), British Classification Society Board member (former President), Classification Society Member, Gesellschaft für Klassifikation Member, Council, IASC; Member, ISIChair, Committee on CS-DM&KD, Computational Statistics and Data Mining for Knowledge Discovery, IASCFellow, IAPR - International Association for Pattern Recognition Fellow, Royal Statistical SocietyFellow, British Computer Society Member, Gesellschaft für Informatik, ACM, IEEEMember of Council, IFCS, Department of Computing Goldsmiths, University of London Big Data Lab, Department of Electronics, Computing and Mathematics, University of Derby
e-mail: fmurtagh@acm.org http://www.fmurtagh.info

Causal analysis with cyclic structural equation models

Mario Nagase

Abstract Causal analysis is primary issue in any disciplines. They might be divided into two categories, one that aimed to measure quantity of causal effect under presupposed causal directions and one that aimed at detecting causal directions themselves only using an observed dataset. I have been focused on the later category and introduced a new idea that takes advantage of cyclic Structural Equation Models (SEMs). Even though cyclic SEMs have problems with difficulties in achieving identification, once models are identified, they can work well not only for analyzing cyclic and reciprocal structure but for identifying causal direction and one can know it by comparing values of coefficients associated with each causal directions. I came up with a new type of method that allows us to incorporate an error covariance into the model, whereas almost all existing methods seem to evade introducing it. An error covariance is considered to be resulted form unobserved confounders, and it often get in the way of measuring true causal effect, because it obviously forms one way of so-called "Back Door". In other words, introducing error covariance can block the back door. However, it changes for worse in light of identification, so that I use Non-Linear cyclic SEMs. Under a certain condition, I will show that the model can be identified and, even with the condition, one can deal with wide range of relationships. I will also give some simulation studies as examples, and will present a way to apply to binary data, if time permits. Detail will be delivered in our presentation.

Keywords: Causal analysis; Non-linear SEMs; Identification; Estimation

Mario Nagase
Graduate School of Engineering Science Osaka University e-mail: nagase@sigmath.es.osaka-u.ac.jp

Spherization of multiway tables - A linear algebraic method for multiple testing problems using maximum statistics and concept of distance

Kazuhisa Nagashima, Tapati Basak, Satoshi Kajimoto, and Ryo Yamada

Abstract Contingency table is a basic tool for testing association of categorical data set. In genetic epidemiology, multiple testing is a fundamental issue and the implicit assumption is the distribution of nominal p-values are correct. This assumption is not often valid for genomics data where p-values are obtained by asymptotic theory. So, the nominal p-values must be corrected to control type I error. There are different methods for this association tests and most them are based on fixed marginal counts. We propose a linear algebraic method called Spherization for this purpose. It is based on vectorization of table arrays using the Kronecker product of simplex-based rotation matrices and an eigenvalue decomposition. Our method defines multiple tests in a higher dimensional space. Coverage of spherized space with random unit vectors rather than random points is the superiority of our method over its conventional counterpart. The transformed space is symmetric with respect to distance and direction and be able to handle the tables that produces very small p-values. This method enables us to evaluate arbitrary combinations of df=1 tests in an integrated way. We compare the suitability of our method to its conventional counterparts. Simulated data sets are used for this evaluation. Our method shows superiority in estimation in terms of variation and stability by overcoming the shortcomings of conventional methods.

Keywords: Spherization; Multiple testing; p-values; Type I error

Kazuhisa Nagashima
Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: kazuhisa.nagashima@genome.med.kyoto-u.ac.jp

Tapati Basak
Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: tapati@genome.med.kyoto-u.ac.jp

Satoshi Kajimoto
Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: kajimoto.satoshi@gmail.com

Ryo Yamada
Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: ryamada@genome.med.kyoto-u.ac.jp

Changes in the gendered division of labor and women's economic contributions within Japanese couples.

Miki Nakai

Abstract The trend towards a dual-earner family can be detected in recent years in many industrialized countries. However, despite the continuing rise in Japanese women's rates of participation in the economy over the period of industrialization and beyond, gender division of labor have been accepted as "natural" and still strong in Japanese society. The aim of this paper is to examine whether the determinants of married women's labor force participation has changed. Based upon social survey of national sample in Japan conducted in 1985, 1995, 2005, and 2015, we analyze change/stability of the factors that differentiate dual-income couples from male-breadwinner households and the factors that differentiate couples where the husband provides the majority of the couple's income from equal providers. Results show that dual-income couples have increased but it is not at a constant rate: it increased at a slow pace until around 2005, and dual-income couples increased dramatically recently. However, the influence of a woman's own human resources on her employment has not been significantly positive until now. Husband's low income have a significant positive effect on labor force participation of married women, suggesting that high occupational resources of the husband drive wife out of the labor market, which has been found in conservative and Mediterranean welfare regimes.

Keywords: Gender division of labor; Male breadwinner; Wives' human resources

Miki Nakai
Ritsumeikan University e-mail: mnakai@ss.ritsumei.ac.jp

Association analysis among myocardial infarction, cardiovascular disease-related single nucleotide polymorphisms, and DNA methylation sites utilizing the cluster analysis

Masahiro Nakatochi, Sahoko Ichihara, Ken Yamamoto, Tatsuaki Matsubara, and Mitsuhiro Yokota

Abstract The development of cardiovascular disease (CVD) depends on environmental and genetic factors. Although many single nucleotide polymorphisms (SNPs) associated with CVD susceptibility have been identified to date, the mechanisms through which these polymorphisms contribute to disease development remain unclear. Recently we reported that DNA methylation (DNAm) play an important mechanism of the development for myocardial infarction (MI), which is a type of CVD, with the use of the genetic epidemiological approach. Some of DNAm in blood cells has also been found to be heritable, and SNPs have been associated with differences in DNAm level. Furthermore, recent findings suggest that a disease might be influenced by disease-associated SNPs via changes in DNAm near the SNPs. Thus, the combination analysis among SNPs, DNAm levels, and disease phenotype is needed to reveal the mechanism for the development of MI. To assess the associations among MI, DNAm status at DNAm sites and CVD-associated SNPs, we applied the association analysis utilizing the cluster analysis to the data in elderly Japanese individuals. This study is based on the genetic epidemiological data. A total of 192 patients with MI and 192 controls were recruited from hospital attendees and the general population, respectively. Genome-wide DNAm profiles in whole blood were obtained by analysis with an Infinium HumanMethylation450 BeadChip. The CVD-associated SNPs were genotyped using an SNP array and were imputed using 1000 genomes phase 3 as a reference. The analysis strategy is composed of three steps. First, association analyses of DNAm level at each DNAm site with each CVD-associated SNP are performed with the used of the general linear regression analysis. Next, the cluster analysis are applied based on the results of the association analyses. DNAm sites with the similar association pattern with SNPs are clustered. Finally, the association of MI with SNP via clustered DNAm pattern are assessed. In this presentation, we introduce the analysis strategy and suggest the insight that MI is influenced by CVD-associated SNPs via changes in DNAm in blood samples.

Keywords: DNA methylation;Single nucleotide polymorphism; Association analysis; Cluster analysis

Masahiro Nakatochi
Statistical Analysis Section, Center for Advanced Medicine and Clinical Research, Nagoya University Hospital e-mail: mnakatochi@med.nagoya-u.ac.jp

Sahoko Ichihara
Department of Environmental and Preventive Medicine, School of Medicine, Jichi Medical University e-mail: saho@gene.mie-u.ac.jp

Ken Yamamoto
Department of Medical Biochemistry, School of Medicine, Kurume University e-mail: yamamoto_ken@med.kurume-u.ac.jp

Tatsuaki Matsubara
Department of Internal Medicine, School of Dentistry, Aichi Gakuin University e-mail: matt@dpc.aichi-gakuin.ac.jp

Mitsuhiro Yokota
Department of Genome Science, School of Dentistry, Aichi Gakuin University e-mail: myokota@dpc.aichi-gakuin.ac.jp

Non-hierarchical clustering for large data without recalculating cluster center

Atsuho Nakayama and Shinji Deguchi

Abstract The present study introduces an efficient technique for segmentation when the large data is analyzed by K-means clustering that is non-hierarchical clustering. The K-means clustering is a widely used clustering technique that seeks to minimize the average squared distance between objects in the same cluster by a simple iterative calculation. It is very attractive that the calculation process to obtain the clusters is simple and convenient. However, it is known that these iterative calculation are especially sensitive to initial starting conditions. It is necessary to duplicate the analysis by using a different initial values. To find the optimum solution for each pair of initial values, it is necessary to repeatedly calculate the cluster centers minimizing the average squared distance between objects in the same cluster. The computing time and the resource for the analysis increase when data is very huge. Therefore, the present study proposes the method that doesn't recalculate the cluster center. The process of the proposed technique is as follows. Firstly, some small sampling data as hierarchical clustering and interpretation of result can be easily done is made from original data. The some sampling data is analyze by the K-means clustering specifying numbers of clusters that are more than the assumed number of clusters. The cluster centers obtained from each analysis is analyzed by the Ward's method that is the hierarchical clustering. Initial values to analyze former data based on the classification result at the cluster center is made. Original data analyze by K-means clustering using the initial values. Each object is classified into the cluster which the cluster center is the nearest. The recalculation at the cluster center is not done.

Keywords: Non-hierarchical clustering; K-means clustering; Initial values

Atsuho Nakayama
Department of Business Administration, Graduate School of Social Sciences, Tokyo Metropolitan University e-mail: atsuho@tmu.ac.jp

Shinji Deguchi
Dataexploring e-mail: shinji.deguchi@dataexploring.com

An exploratory study on the clumpiness measure of inter-transaction times: how is it useful for customer relationship management?

Yuji Nakayama and Nagateru Araki

Abstract Customer relationship management (CRM) is crucial for retailers, because large costs are necessary for acquiring new customers, and maintaining existing loyal customers with high lifetime value is a key to excel other retailers in the competitive economic environment.
In the field of marketing science, CRM is also an important area of research. One of essential topics is to investigate the measure of customers' behavior based on the purchase history in order to predict their future behavior and find valuable customers. The recency/ frequency/ monetary value (RFM) framework that has been used since 1960s is well-known in the industry, and still valuable for summarizing customers' purchase history.
Recently, Zhang, Bradlow, and Small (2015) *Marketing Science* proposed the clumpiness (C) measure of inter-event times defined as the degree of nonconformity of equal spacing. They applied this measure to customers' purchase (or visit) intervals at online/offline stores, and showed that adding C to RFM framework enhanced the predictive power of customers' future behavior. However, there remains much to be investigated on this measure.
In this study, we explore how effective the clumpiness measure is for market segmentation, in particular whether it is useful to find (or at least not to overlook) profitable customers in the future. For this purpose, we use ISMS Durables Goods Dataset 1, a panel data set of about 20 thousand households' purchase history at a major U.S. consumer electronics retailer for 6 years, provided by Ni, Neslin, and Sun (2012) *Marketing Science*.

Keywords: Customer Relationship Management; Market Segmentation; Buyer Behavior

Yuji Nakayama
Osaka Prefecture University e-mail: nakayama@eco.osakafu-u.ac.jp

Nagateru Araki
Osaka Prefecture University e-mail: araki@eco.osakafu-u.ac.jp

Simulation and analysis by growth model reflecting non-uniformity

Kazuhide Namba

Abstract In this paper we introduced product growth model with non-uniformity of purchase. And simulation by growth model was done. The non-uniformity of the information network and consumer behavior and other purchase factor were reflected to model in the growth process. As a result of simulation, we can get growth phenomenon. The growth phenomenon includes normal and chasm and other complicated phenomenon. The phenomenon varies depending on the parameters of non-uniformity. It was found that the growth phenomenon changes with the degree of influence of consumer behavior, information network and other factors of affected parameters. Complicated growth phenomena are different depending on the timing of impact with early adaptor in the growth process. By this research, the growth structure could be analyzed in more detail.

Kazuhide Namba
Kochi University of Technology e-mail: nan@hop.ocn.ne.jp

Data quality management of chain stores based on outlier detection

Linh Nguyen and Tsukasa Ishigaki

Abstract For successfully analysing data in business of chain stores, quality of data recorded in their shops or factories is a key factor. Data quality management is an important practical issue, because data qualities widely vary depending on managers or workers of many stores in chain. In this paper, we present a data quality evaluation method for shops in chain businesses based on outlier detection and then, we apply this method to a dataset observed in real chain stores, which provide tire maintenance for vehicles. To evaluate the data quality of each shop, we use data about truck's tire information such as tread depth, tread pattern and distance which was recorded by shop at maintenance time to calculate obviously low quality data by using outlier detection methods with reliable experimental data and practical knowledge. Some of outlier detection methods such as Isolation Forest and one-class Support Vector Machine are applied to detect anormalous tire information, which uses to calculate data's anomal rate in each shop. Our result showed that with this kind of data, Isolation Forest is outstanding than other methods because Isolation Forest is designed to detect "few and different" outliers. The proposed method can support better maintenance services for customer as well as be able to get more correct data from these shops, which will be useful for next research.

Keywords: Abnormal Data; Maintenance Service; Isolation Forest

Linh Nguyen
Master Student at Graduate School of Economics and Management, Tohoku University, Tohoku University e-mail: linh.nguyen.1992@gmail.com

Tsukasa Ishigaki
Tohoku University e-mail: isgk@econ.tohoku.ac.jp

Supervised nested algorithm for classification based on k-means

Luciano Nieddu and Donatella Vicari

Abstract The aim of this paper is to present an extension of the k-means algorithm based on the idea of recursive partitioning, that can be used as a classification algorithm in the case of supervised classification.
The problem of supervised classification has been gathering lot of interest since the seminal work of Sir R. Fisher.
In this paper we will be dealing with recognition with perfect supervisor and we show how the proposed methodology can be extended to handle the case of imperfect supervisor.
Clearly no algorithm is able to give the best performance without any prior assumptions on the data. This has been mathematically clarified in 1996 with the proof of the "no free lunch theorem".
Some of the most robust techniques for supervised classification are those based on classification trees that make no assumption on the parametric distribution of the data and are based on recursively partitioning the feature space into homogeneous subsets of units according to the class the entities belong to. This approach has been recently developed incorporating simple parametric models into the terminal nodes of the tree. Research in this direction was motivated by the fact that a constant value in the terminal node tends to produce large and thus hard to interpret trees.
One of the shortcomings of these approaches is that the recursive partitioning of the data, i.e. the growing of the tree, is achieved considering only one variable at a time and, although it makes the trees pretty simple in terms of determining rules to obtain a classification, it also makes them hard to interpret.
Building on these ideas, we carry the integration of parametric models into trees one step further and propose a supervised classification algorithm based on the k-means routine, that sequentially splits the dataset according to the whole feature vector

Keywords: Classification; Supervised; k-means

Luciano Nieddu
Assistant Professor in Statistics, Faculty of Economics, UNINT University e-mail: l.nieddu@gmail.com http://www.unint.eu

Donatella Vicari
Associate Professor of Statistics, Department of Statistics, Sapienza - University of Rome e-mail: donatella.vicari@uniroma1.it

Challenges in visualising and imputing missing categorical data

Johané Nienkemper-Swanepoel, Sugnet Lubbe, and Niël le Roux

Abstract The application of the GPAbin methodology to real incomplete categorical data sets will be presented. The GPAbin approach is an amalgamation of Rubin's rules for multiple imputed datasets and generalised orthogonal Procrustes analysis (GOPA) to obtain a final visual representation of multiple imputations. Multiple imputation is a favoured unbiased approach to handling missing values, in which multiple completed data sets are available for further individual inspection after completion of the imputation procedure. Standard statistical analyses can be applied to each individual imputed data set and subsequently the descriptive statistics can be combined for final inference using Rubin's rules. Exploratory analysis of a large number of multiple imputations can rapidly become unmanageable and inhibit a concise final conclusion from the multiple visualisations. Multiple correspondence analysis (MCA) biplots are used for the visualisation of the categorical data sets. The multiple MCA biplots are then combined using the GPAbin approach. A simulation study considering various data scenarios have shown promising measures of fit within in the Procrustes framework, by comparing the final combined visualisations with the MCA biplot of the simulated complete data. A brief overview of the simulation study will be presented with the focus of the presentation being aimed at the real application.

Keywords: multiple imputation; multiple correspondence analysis; Procrustes analysis

Johané Nienkemper-Swanepoel
Lecturer, Department of GeneticsPhD Student, Department of Statistics and Actuarial Science, Stellenbosch University
e-mail: jnienkie@gmail.com

Sugnet Lubbe
Professor, Department of Statistics and Actuarial Science, Stellenbosch University e-mail: slubbe@sun.ac.za

Niël le Roux
Professor, Department of Statistics and Actuarial Science, Stellenbosch University e-mail: njlr@sun.ac.za

Bi-modal clustering and quantification theory: Flexible-filter clustering of contingency and response-pattern tables. Numerical demonstration of a proposed framework

Shizuhiko Nishisato and Jose Garcia Clavel

Abstract Nishisato and Clavel (2010) proposed a new approach to investigate the relations between rows and columns of the contingency table (C) in expanded space, where the key information used consists of both within-set and between-set distances. To deal with doubled multidimensional space, their view was to give up multidimensional joint graphical display and adopt cluster analysis instead. In the meantime, Nishisato (2012, 2014) proposed a simple method of cluster analysis with a p-percentile (flexible) filter. Nishisato and Clavel (2017) investigated merits and demerits of three types of clustering for the analysis, namely, hierarchical clustering (CH), partitioning clustering (CP) and clustering with flexible filters (CF). Their work identified CF as the most promising method of cluster analysis, and recommended that the between-set distance matrix is a preferred input for cluster analysis. Nishisato (2017) further verified that the contingency table could be reformatted into the response-pattern table (F) in which the rows consist of combinations of rows and columns of C and the columns consist of rows and columns of C, and that F typically requires more than twice the dimensional space than C. The extra dimensions for F suggest that analysis of C may not be exhaustive of information in C. Based on those previous studies, the current work will explore clustering with flexible filters using F as the input for dual scaling: Subject F to dual scaling; using the entire set of coordinates for rows, calculate the between-row distance matrix; subject the portion corresponding to the row-by-column distance matrix of C to clustering with flexible filters, with the focus on how to use clustering with flexible filters for identifying clusters.

Keywords: Dual Scaling; Flexible Filtering

Shizuhiko Nishisato
Professor Emeritus, University of Toronto, Canada, University of Toronto e-mail: shizuhiko.nishisato@utoronto.ca

Jose Garcia Clavel
Universidad de Murcia e-mail: jjgarvel@gmail.com

Classification of in-week and -day patterns in ambulatory activity and body composition change

Shunichi Nomura, Michiko Watanabe, and Yuko Oguma

Abstract In this study, we extract several in-week and -day patterns in ambulatory activity and body composition change from a big data of activity monitor and body composition meter records. As for ambulatory activity, a hierarchical topic model with two layers of in-week patterns as super-topics and in-day patterns as sub-topics is applied to aggregated records of step counts. As a result, four diurnal activity patterns named commuting, morning, daytime and night activities are extracted as sub-topics and several combinations of these patterns within a week are extracted as super-topics. In super-topics, diurnal activity patterns are different between Monday through Friday and weekend. As for body composition change, we apply a state space model with two kinds of periodic variations within a day and within a week to records of body composition meter. State space models are suitable to body composition records because they are not recorded every day and include measurement errors. We also classify the estimated periodic variations in body composition. It is found that many people have their highest weight at night and their lowest weight in the afternoon within a day. On the other hand, some people have highest peak of their weights in weekends and other people vice versa. Using in-week and -day patterns extracted in the analysis, we discuss dynamical relation between ambulatory activity and body composition change.

Keywords: Healthcare Data; Topic Model; State Space Model

Shunichi Nomura
Assistant Professor, The Institute of Statistical Mathematics e-mail: shun1982jcomhome@gmail.com

Michiko Watanabe
Professor, Graduate School of Health Management, Keio University e-mail: watanabe_michiko@nifty.com

Yuko Oguma
Associate Professor, Graduate School of Health Management, Keio University; Sports Medicine Research Center, Keio University e-mail: yoguma@a7.keio.jp

Extensions of Pearson's inequality between skewness and kurtosis to multivariate cases

Haruhiko Ogasawara

Abstract An extension of Pearson's inequality between squared skewness and kurtosis to the case with three possibly distinct variables is obtained. A similar extension to the multivariate analogue of skewness defined by Mardia (1970) is also derived. Upper and lower bounds for kurtosis are shown with inequalities between the multivariate fourth cumulants for standardized variables. Improved inequalities for the multivariate versions of squared skewness when a vector variable is infinitely divisible are obtained.

Haruhiko Ogasawara
Otaru University of Commerce e-mail: hogasa@res.otaru-uc.ac.jp

Attempt to set a step count target by applying latent class analysis

Kotaro Ohashi, Michiko Watanabe, and Yuko Oguma

Abstract The purpose of this research was to use the concept of latent class to refer to the correlation between step count and a measure of body composition, and to set a feasible and reasonable step count goal for health promotion, considering age and sex.
It is well known that regular physical activity (PA) is important to maintain and improve health as well as to prevent non-communicable diseases, such as diabetes status, cardiovascular disease, and cancer. Visceral fat accumulation is said to be present in the **upstream** of a variety of disorders including cardiovascular disease. Step count by pedometer is used to monitor daily physical activity conveniently. Therefore, in this research, we conducted the following three analyses:

1. Latent clustering of pace of increase of participants
2. Based on the result of 1, setting a target step count for safe and easy-to-use health promotion
3. Confirmation of the difference in visceral fat level caused by a difference in the degree of achievement of target steps

As a result, regarding the data we used first, we decided that four groups were appropriate for all ages and sexes. The determination of these four groups is a judgment in consideration of safety and convenience. The group having the smallest number of target steps was the first group, and as the target step number increased, it determined the second, third, and fourth groups. However, the value of the target step count in each group was different depending on sex and age. The participants were subdivided into four groups using the target step number created, and covariance analysis was performed among these groups. As a result, after controlling the physique-related variables, significant differences between the visceral fat levels were found.

Keywords: Latent Class; Pedometer; Health Science

Kotaro Ohashi
Rikkyo University e-mail: kotaro-0084@rikkyo.ac.jp

Michiko Watanabe
Graduate School of Health Management, Keio University e-mail: watanabe_michiko@nifty.com

Yuko Oguma
Graduate School of Health Management, Keio University; Sports Medicine Reserch Center, Keio University.
e-mail: yoguma@a7.keio.jp

Dissimilarity based on manner of dissimilarity/similarity to/from the others

Akinori Okada

Abstract While the definition of a cluster differs from one researcher to another, there is the common notion; a cluster is a set of objects which are similar, and objects in different clusters are dissimilar (e.g. Anderberg, 1971). The cluster relies on the dissimilarity/similarity among objects. The dissimilarity/similarity between two objects is obtained by e.g., the intimacy among people obtained by rating scale, the frequency of brand switching among brands, the distance or correlation based on a set of variables, …which depend on a relationship between two objects. The dissimilarity/similarity depending on a relationship between two objects is not necessarily reasonable when we consider objects belong to different clusters, because objects belong to different clusters should be dissimilar and simultaneously each object should be similar to objects in a cluster the object belongs. One practicable method to cope with this is the dissimilarity based on the difference of the manner of dissimilarity/similarity. Objects in a same cluster should have the similar manner of dissimilarities/similarities between the objects and those in the other clusters. Objects in different clusters should have the dissimilar manner of dissimilarities/similarities between objects in the other clusters. An example of such dissimilarity is the square root of the sum of squared differences of corresponding elements in two columns (rows) of a brand switching matrix. The dissimilarity is defined not by a relationship of two objects directly but by relationships of an object with the other objects indirectly, and accords with the etymology of the Japanese (Han) character "classification" which emphases distinction among clusters rather than cohesion within a cluster. The conception of the dissimilarity can be extended to deal with asymmetric relationships and to represent the local or the global structures of dissimilarity/similarity relationships.

Reference

Anderberg, M. R. (1971). Cluster analysis of applications. Academic Press, New York.

Keywords: cluster, dissimilarity, manner of dissimilarity/similarity, Similarity

Akinori Okada
Tama University e-mail: okada@rikkyo.ac.jp

Extension of Sinkhorn method: Optimal movement estimation of agents under law of inertia

Daigo Okada, Naotoshi Nakamura, Yosuke Fujii, Takuya Wada, Ayako Iwasaki, and Ryo Yamada

Abstract
OBJECTIVE: In this research, we propose a new method to estimate the movement of agents following the law of inertia in arbitrary space. Such a method is considered to be useful in various fields including image data analysis with time series information
METHOD: First, we observe the position coordinates of agent at three consecutive times. The observed space is discretized into d cubicles and we create a d^3l cost array which weighs acceleration, which is the expansion of the optimal transport estimation the based on Sinkhorn distances (Cuturi, Adv Neural Inf Process Syst, 2013) from two dimensions to three dimensions.
RESULT: We applied our methods to the simulation 2d microscopic images of mobile cells. As a result, when the number of cells was 100 or less, high performance was obtained, and accuracy decreased as the number of cells increased further. The method was also applied to the estimation of cellular surface flow with production of reasonable vector field.
CONCLUSION: This method is considered to be effective as a method of tracking change of agents from data with time series information where one of the fundamental law in physics, the law of inertia, is assumed, and it can be applied to multiple settings.

Keywords: Biomedical Data Analysis and Imaging; Optimal Transport; Movement

Daigo Okada
Master course student, Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: okada.daigo.47z@st.kyoto-u.ac.jp

Naotoshi Nakamura
Postdoctoral researcher, Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: nnakamura@genome.med.kyoto-u.ac.jp

Yosuke Fujii
Doctor course student, Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: fujii@genome.med.kyoto-u.ac.jp

Takuya Wada
Undergraduate student, Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: peacefield.taku3@gmail.com

Ayako Iwasaki
Undergraduate student, Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University e-mail: iwasaki.ayako.38n@st.kyoto-u.ac.jp

Ryo Yamada
Graduate School of Medicine, Kyoto University e-mail: ryamada@genome.med.kyoto-u.ac.jp

Generalized extended procrustes post-processing of MCMC samples in bayesian multidimensional scaling

Kensuke Okada and Shin-ichi Mayekawa

Abstract Multidimensional scaling model provides a spatial representation of the observed dissimilarity data in low dimensions. The extracted dimensions may be eligible for a meaningful classification of observed variables. However, it has been known that the multidimensional scaling model exhibits an indeterminacy of rotation, reflection, and translation of the target configuration matrix. This indeterminacy issue is especially relevant in Bayesian estimation that employs Markov chain Monte Carlo (MCMC) estimation technique, because in that case every random sample from the posterior distribution has its unique indeterminacy. In this study, we first review the existing methods to deal with this indeterminacy issue. The existing method can be classified as either parameter-fixation approach or post-processing approach. Then, we propose a new post-processing method that is based on the generalized extended Procrustes rotation. The idea behind the proposed method is to jointly optimize the origin and rotation of the the set of MCMC samples by using iterative least-squares type algorithm. We conducted a Monte Carlo simulation study and found that the proposed method worked at least as good as the best existing method. Then, the proposed method was applied to the real dataset to illustrate its usefulness and interpretability. The convergence property of the proposed as well as existing method is also discussed.

Keywords: Bayesian Approach; MDS; MCMC

Kensuke Okada
Department of Psychology, Associate Professor, Senshu University e-mail: ken@psy.senshu-u.ac.jp

Shin-ichi Mayekawa
Professor, Tokyo Institute of Technology e-mail: mayekawa@nifty.com

Online ensemble learning using hierarchical bayesian model averaging

Makoto Okada, Keisuke Yano, and Fumiyasu Komaki

Abstract We consider online prediction using multiple statistical models. For such a prediction, Raftery et al. proposed a Bayesian model averaging method using the Expectation-Maximization algorithm and a sliding window. In their method, each model has a weight that is updated according to a predictive loss in the sliding window. These weights are the posterior probability of models and the predictors of models are aggregated by the weighted average based on the weights of models. However, since in their algorithm the weights tend to overfit to the loss in the sliding window, the weights change drastically, which causes a decrease of predictive accuracy. We solve this problem by a hierarchical Bayesian model averaging method that introduces a prior distribution for weights to smoothe them. We use a variational Bayesian method to derive explicit update equations. Through numerical experiments, we show that our method has better predictive performance even when a sliding window is short and that the Expectation-Maximization algorithm in our method converges earlier. In addition, we apply our method to the online classification using a weighted majority method.

Keywords: EM algorithm; Emsemble learning; Expert model; Hierarchical model; Model averaging method; Online learning; Variational Bayesian method

Makoto Okada
Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo e-mail: makoto_okada@mist.i.u-tokyo.ac.jp

Keisuke Yano
Department of Mathematical Informatics, Graduate Schoool of Information Science and Technology, The University of Tokyo e-mail: yano@mist.i.u-tokyo.ac.jp

Fumiyasu Komaki
Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo RIKEN Brain Science Institute e-mail: komaki@mist.i.u-tokyo.ac.jp

Leveraging local data structure for multi-view correlation analysis with graph-structured associations

Akifumi Okuno and Hidetoshi Shimodaira

Abstract Various types of data, such as texts, images, and sounds, become easily obtainable these days. Different types of data are referred to as "views" and approaches to integrate the multi-view data into a unified representation have attracted much attention.These multi-view datasets may have many-to-many associations across views while one of the most popular multi-view learning methods called Canonical Correlation Analysis (CCA) assumes one-to-one associations. Shimodaira (2016) generalizes CCA as Cross-Domain Matching Correlation Analysis (CDMCA) to handle many-to-many associations across views. The input of CDMCA can be represented as a weighted graph where the associations are weighted links.These graph-structured associations are generally specified by external knowledge, i.e. annotation by human effort. In practice, each of these obtained associations is given as binary value which only tells us whether two vectors have an association, but not the weight of the association. Since CDMCA can consider the non-binary weights of associations, binary associations do not fully utilize the method. In this study, we aim at giving weights to the associations before applying CDMCA by leveraging local data structure. It corresponds to taking a local weighted average of associations in the vicinity, and particularly important associations may be emphasized. Moreover, we can complement unobserved associations beforehand; we show that our method improves CDMCA's performance in numerical experiments.

Akifumi Okuno
Doctoral student of Shimodaira Lab., Department of Systems Science, Graduate School of Informatics, Kyoto University, and RIKEN Center for Advanced Intelligence Project, Department of Systems Science, Graduate School of Informatics, Kyoto University, and RIKEN Center for Advanced Intelligence Project e-mail: okuno@sys.i.kyoto-u.ac.jp

Hidetoshi Shimodaira
Professor, Department of Systems Science, Graduate School of Informatics, Kyoto University, Department of Systems Science, Graduate School of Informatics, Kyoto University, and RIKEN Center for Advanced Intelligence Project e-mail: shimo@i.kyoto-u.ac.jp

The stylometry on Japanese political documents

Yohei Ono

Abstract The main objective of this paper is to address the application of stylometry to Japanese political documents. The various researches have contributed to the Japanese stylometry since the monumental paper of Jin and Murakami (1993). Jin and Murakami (1993) and their series of researches have clarified that the use of commas or auxiliary verbs is a good indicator on the authorship problems of Japanese texts. The previous study of the author (Ono, submitted) showed that their methodologies can be applied to not only the authorships of the text but also the inference on how the texts were written. In this paper, the author expands upon this method to the Japanese political documents. Since the researches, led by some historians (e.g. Emeritus Prof. Takashi Itoh [The University of Tokyo]), have provided us large amounts of Japanese political documents including diaries, interviews or reminiscences of famous bureaucrats, commanders and politicians. From the viewpoints of stylometry, the present author calls their invaluable achievement "Big Data for Japanese political documents" and proposes the new development of Japanese stylometry, in which researchers seek how documents (e.g. Japanese prime ministers' diet addresses or diplomatic documents) were written, comparing the characteristic writing styles of those who historian suggests their involvements by philological approach. The author applies this method to some documents, on which various philological views were established, and demonstrates that our approaches can contribute to those discussions. For future research, combining philological and statistical approaches, those complementary studies will matter.

Keywords: Clustering; Japanese Political Documents; Stylometry; Text Classification

Yohei Ono
Graduate student, Division of Arts and Sciences, The Open University of Japan e-mail: metrics.languages@gmail.com

Nonlinear factor analysis for NCT test score linking

Tatsuo Otsu

Abstract A method of nonlinear factor analysis (NFA) that represents nonlinear relationships between a latent variable and observed variables was used fo linking test scores of a nationwide university admission test. Classical factor analysis model represents linear relationships between latent variables and observed variables. Although this is widely used for test analyses, nonlinear relationships need to be represented for linking large scale achievement tests.Here, a NFA model that uses spline transformations of the latent variable is applied for analysing recent NCT (National Center Test) scores. A conditional distribution of an observed variable given by the latent variable is assumed to have a mean (or a location parameter) that is expressed in nonlinear transformation of the latent variable. A discrete approximation of the latent variable enables easy adaptation for the missing values under MAT (missing at random) condition. NFA needs large computational burden for large data. A computational procedure using parallelization (Open MP) is adopted for efficient program implementation. It is suggested that mean score comparisons between marginal distributions are insufficient for monitoring test difficulties. Whole shape of the simultaneous score distribution should be considered for deliberate linking.

Keywords: nonlinear factor analysis; Open MP; test linking

Tatsuo Otsu
Research Division, Professor, National Center for University Entrance Examinations e-mail: otsu@rd.dnc.ac.jp

ABC analysis in corporate bankruptcy prediction

Barbara Pawełek, Józef Pociecha, and Mateusz Baryła

Abstract In the theory and practice of the prediction of company bankruptcy threats, numerous company prediction models have been developed. But on the basis of the results of the empirical research presented in the literature, it is not possible to state clearly what types of prediction models provide the best results. Thus, researchers make attempts to define the sources of errors made in the bankruptcy prediction process. One of the sources of misclassification of items may be the heterogeneity of a research set. Consequently, some researchers point to the need to ensure homogeneity of a set of items in respect to the size of a company, measured by means of the level of assets, liabilities, employment etc. The purpose of the paper is to present the results of the empirical research of the impact of a company's size on the results of bankruptcy prediction. The added value of the paper is the proposal of application ABC Analysis aimed at a division of the set of companies into subsets of entities that are similar to each other in respect to size, measured by means of selected financial features. The analysis was based on the financial data of companies operating in the industrial processing sector in Poland in the years 2005-2008. The examination covered the following methods, among others: classification tree, k-nearest neighbors, support vector machine, bagging, boosting, random forest, neural network and naive Bayes. The models were built on the basis of financial data referring both to all the items in a set under consideration and items classified in one of three subsets by means of ABC Analysis. The predictive capability of the constructed models was assessed on the basis of accuracy rate, sensitivity, specificity, Brier score, AUC and others. The calculations were made by the R language and the main packages used were 'ABCanalysis' and 'rminer'.

Keywords: ABC Analysis; corporate bankruptcy; prediction model; predictive capability

Barbara Pawełek
Cracow University of Economics e-mail: barbara.pawelek@uek.krakow.pl

Józef Pociecha
Cracow University of Economics e-mail: jozef.pociecha@uek.krakow.pl

Mateusz Baryła
Cracow University of Economics e-mail: mateusz.baryla@uek.krakow.pl

Optimal doubling burn-in policy based on tweedie processes with applications to degradation data

Chien-Yu Peng

Abstract In the current competitive marketplace, manufacturers need to screen weak products in a short period of time. It is a challenge for manufacturers to implement a burn-in test that can screen out the weak products quickly and efficiently. When selecting an approach to determine the duration of the burn-in, one could build a criterion aiming to minimize the burn-in cost. In practice, when the optimal burn-in time is unreasonable (e.g., time 0) due to minimizing the cost, this means that the burn-in procedure is unnecessary to perform for manufacturers. In this study,we propose an optimal doubling burn-in policy to improve the predicament without additional experiments. The advantage of the proposed policy is to simultaneously determine the optimal burn-in time and the optimal cutoff point for classifying weak and strong components from the production. In addition, a new degradation model based on a Tweedie mixture process is used fora burn-in test. Several data sets are performed to demonstrate the proposed burn-in procedure.

Chien-Yu Peng
Institute of Statistical Science e-mail: chienyu@stat.sinica.edu.tw

Optimal model-based clustering with multilevel data

Fulvia Pennoni, Francesco Bartolucci, and Silvia Bacci

Abstract We present how to get a classification of the units by means of a maximum a posteriori algorithm which is based on the Viterbi algorithm (Viterbi, 1967, Juang and Rabiner, 1991) when a latent class item response theory model is formulated by considering random effects for data having a hierarchical or multilevel structure. The random effects are used to model the influence of each cluster on the responses provided by the subjects who are included in the cluster. We describe the extended version of the item response multilevel Latent Class (LC) model to deal with covariates collected at both cluster and individual level.
We show how to deal with the Expectation-Maximization algorithm to estimate the model parameters and how to implement the classification algorithm by the results of the fitted models. We show some applications based on real data collected by the Italian INVALSI agency on Literacy and Math and on the Italian and Japanese data on Literacy, Math and Science collected on the large-scale assessment surveys TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study).

Fulvia Pennoni
Department of Statistics and Quantitative Methods University of Milano-Bicocca e-mail: fulvia.pennoni@unimib.it
http://www.statistica.unimib.it/utenti/pennoni/

Francesco Bartolucci
Department of Economics, University of Perugia, e-mail: francesco.bartolucci@unipg.it

Silvia Bacci
Department of Economics, University of Perugia e-mail: silvia.bacci@unipg.it

K-means clustering on multiple correspondence analysis coordinates

Le Phan and Hongzhe Liu

Abstract In response to the International Federation of Classification Societies (IFCS) data challenge, we present our cluster analysis of a dataset consisting of 928 patients with lower back pain. We use a two-pronged approach to clean the data: inferring values missing not at random (based on known relationship between variables) and imputing missing values with the Multiple Imputations by Chained Equations (MICE) method. We also discuss the challenges in clustering mixed data types and the required data transformation prior to applying a clustering algorithm. We call this transformation process "split-then-join". Our k-mean clustering result identified three distinct groups that can be characterized by the intensity of their back pain, the extent to which their functional activities are limited, and their self-reported mood.

Keywords: Adjusted Rand index (ARI); Bayesian information criterion; Calinski-Harabasz criterion; classification; K-means; multiple correspondence analysis; Multiple Imputation by Chained Equations; principal component analysis; variable contribution

Le Phan
Graduate Student, Department of Mathematics and Statistics, San Jose State University
e-mail: lephan1205@gmail.com

Hongzhe Liu
Graduate Student, Department of Mathematics and Statistics, San Jose State University
e-mail: hongzheliu11@gmail.com

Asymptotic theory of the sparse group lasso

Benjamin Poignard

Abstract This paper proposes a general framework for penalized convex empirical criteria and a new version of the Sparse Group Lasso (SGL, Simon and al., 2013), called the adaptive SGL, where both penalties of the SGL are weighted by preliminary random coefficients. We explore extensively its asymptotic properties and prove that this estimator satisfies the so-called oracle property (Fan and Li, 2001), that is the sparsity based estimator recovers the true underlying sparse model and is asymptotically normally distributed. Then we study its asymptotic properties in a double-asymptotic framework, where the number of parameters diverges with the sample size. We show by simulations that the adaptive SGL outperforms other oracle-like methods in terms of estimation precision and variable selection.

Keywords: Asymptotic Normality; Consistency; Oracle Property

Benjamin Poignard
PhD candidate, Finance and Insurance Laboratory at CREST, CREST and Paris Dauphine University
e-mail: bpoignard@gmail.com

A clustering model for decision

Pascal Préa, Célia Châtel, and François Brucker

Abstract A decision tree is a binary tree with a decision criterion associated with each node. From a clustering point of view, a decision tree is a hierarchy, the decision criterion being the membership. The only restriction is that every cluster has exactly two successors. So, seen as binary clustering systems, decision trees can be easily generalized to intervals of linear orders, paired hierarchies (P. Bertrand 2002)...
The aim of this talk is to show that all these generalizations belong to an unique model: *Binary Totally Balanced Hypergraphs*. We will first give a constructive bijection between closed clustering systems and Binary Totally Balanced Hypergraphs. We will also give an efficient algorithms to transform a given closed clustering system into a binarized one, if possible (i.e. if it is totally balanced).
This framework allows us to treat various kinds of data: binary matrices (in this case, decisions are conjonction of attributes), vectors (decisions are centers of mass), dissimilarities (decisions are balls or two-balls)

Keywords: decision tree; totally balanced hypergraphs; clustering models

Pascal Préa
Laboratoire d'Informatique Fondamentale de Marseille, CNRS UMR 7279 Ecole Centrale Marseille
e-mail: pascal.prea@centrale-marseille.fr

Célia Châtel
Laboratoire d'Informatique Fondamentale de Marseille, CNRS UMR 7279, Aix-Marseille Université
e-mail: celia.chatel@lif.univ-mrs.fr

François Brucker
Laboratoire d'Informatique Fondamentale de Marseille, CNRS UMR 7279 Ecole Centrale Marseille
e-mail: francois.brucker@lif.univ-mrs.fr

Evaluating fuzzy clustering results

Elena Rihova

Abstract Nowadays there are many fuzzy clustering algorithms. And all those algorithms have the same question: how evaluate the optimal number of clusters. The optimal number of clusters, which has a deterministic effect on the clustering results. The optimal number of clusters can be determined with the help of cluster validity indices. Cluster validity indices are used for estimating the quality of partitions produced by clustering algorithms and for determining the number of clusters in data. This paper describes a new validity index for fuzzy clustering and proposes a new validity index for fuzzy clustering. The proposed index is tested and validated using several real and generated data sets. The paper also presents experimental results concerning them.

Keywords: Clustering, fuzzy sets, validity, indices

Elena Rihova
University of Business in Prague e-mail: elena.rihova@gmail.com

R shiny implementation of lasagna plot: interactive manipulation and visualization of longitudinal data

Wataru Sakamoto and Marina Kaneda

Abstract Visualization of longitudinal data often requires a complicated task. One reason comes from two different data format. Generally, the long format would be more flexible for analyzing longitudinal data, while the wide format would be apparently more suitable for plotting, with time corresponding to each column in horizontal axis. Another reason is that there are few choices of tools for visualizing longitudinal data.Usually, a connected line is drawn for each subject. However, the overlapping of (often colored) connected lines might confuse or mislead our understanding on the data. Swihart et al. (2010) proposed the lasagna plot for visualizing longitudinal data with large number of subjects, instead of so called the spaghetti plot, a collection of connected lines. The lasagna plot, which looks like a heat map, represents the level of measured values with color grading or shading. It often provides useful interpretation of overall properties behind longitudinal data with sorting and conditioning by row (subject) and/or column (time).We introduce an R Shiny implementation of lasagna plot for interactive manipulation and visualization of longitudinal data, The R package "tidyverse", which is based on the notion of "R for Data Science" advocated by Wickham (2017), actualizes smooth data manipulation, such as conversion between long format and wide format, and easy sorting and conditioning, which assists effective visualization with lasagna plot. An animation of lasagna plot plays a role of an additional dimension and would help our interpretation of complicated longitudinal data. An example of visualizing longitudinal data, such as monthly and hourly PM2.5 concentration at several sites in Okayama Prefecture, is illustrated.

Keywords: interactive plot; longitudinal data; data visualization

Wataru Sakamoto
Division of Human Ecology, Graduate School of Environmental and Life Science, Okayama University
e-mail: w-sakamoto@okayama-u.ac.jp

Marina Kaneda
Fukuyama Myo-o-dai High School e-mail: pkbb3lh9@s.okayama-u.ac.jp

The economic structure of rural areas in the mekong region countries: a comparative analysis of micro official statistics in Cambodia, Thailand, and Viet nam

Daisuke Sakata and Motoi Okamoto

Abstract The Greater Mekong Subregion (GMS), consisting of Cambodia, Lao People's Democratic Republic, Myanmar, Thailand, Viet Nam, and the Yunnan Province and the Guangxi Zhuang Autonomous Region of the People's Republic of China, is attracting attention as an important economic zone. While the intra-regional transportation infrastructure has been improved and the development in this region continues to advance, the intra-regional economic gap is still large. Therefore, comparative analysis of the economic structure of these countries is important, but the current collection of comparative analysis with micro data is inadequate, while comparison with macro data and analysis on a country with micro data abounds. One reason is that each micro data, especially household-level data, differs markedly in survey method, survey item, classification, and reference period.
I conducted a comparative analysis with micro data and a survey slip from three countries in the GMS: Cambodia Socio-Economic Survey (CSES), Thailand's Household Socio-Economic Survey (HSES), and Vietnamese Household Living Standard Survey (VHLSS). These data are stored in the International Official Statistical Micro Database (IOSMDB).
The IOSMDB is built and maintained by the Statistical Information Institute for Consulting and Analysis and the Research Organization of Information and Systems. The data available in the IOSMDB include 2009 CSES, 2007 and 2011 HSES, and 2006 VHLSS. These are 80% resampling data of the original micro data and anonymized data. In the IOSMDB, users' manuals including survey slips and other useful information for understanding data structure also are compiled.
One of the main purposes of these surveys is to assess people's living conditions from many perspectives, including income, expenditures, employment, education, and health. In my presentation, I will discuss the result of a comparison of the economic structures in rural areas, especially regarding the structure of income and expenditures, and the necessary processes to obtain comparable data.

Keywords: CSES; HSES; VHLSS

Daisuke Sakata
Office of Director-General for Policy Planning (Statistical Standards), Ministry of Internal Affairs and Communications
e-mail: d.sakata@soumu.go.jp

Motoi Okamoto
Research Organization of Information and Systems/The Institute of Statistical Mathematics
e-mail: mokamoto@ism.ac.jp

Directional model-based clustering with social-network information for recommendation

Aghiles Salah and Mohamed Nadif

Abstract Recommender systems have become essential in modern web applications so as to guide users towards items that might interest them. Collaborative filtering (CF) is the most widely used recommendation approach in practice. Despite it success, CF-based methods still suffer from the high dimensionality and extreme sparsity of user-item preference data. Recently, social CF approaches, leveraging information from social networks such as friendships and trust relationships, have proven effective in alleviating the sparsity related issues such as the problem of cold start users, i.e., who expressed very few ratings. These approaches build on the assumption that, for making good recommendations, not only the user's expressed preferences are important, but also the user's social interactions. This is natural as people often turn to their friends for advice before choosing a movie, a book, a restaurant, etc. By capturing this real-life behaviour, social CF approaches offer more realistic and better recommendations than traditional CF models, in most cases. Apart from being high dimensional and sparse, CF data are also directional in nature. However, existing social CF models build on popular assumptions, such as Gaussian or Multinomial, which are inadequate for directional data. In order to address this limitation, we propose a new social CF model that builds on the von Mises-Fisher (vMF) assumption, from directional statistics, that turns out to be more adequate to model CF data. More precisely, we propose a regularized vMF mixture model based on the user-user social network. For inference and parameter estimation, we derive a scalable Generalized Expectation-Maximization (GEM) algorithm. Extensive experiments, on several real-wold data sets, illustrate the advantages of our modeling assumption and suggest that, to make more accurate recommendations, not only the social interactions among users should be taken into account, but also the intrinsic directional characteristics of CF data.

Keywords: Recommender systems, Collaborative Filtering, Mixture Models, Von Mises-Fisher distribution, Directional Statistics

Aghiles Salah
University of Paris Descartes e-mail: aghiles.salah@parisdescartes.fr

Mohamed Nadif
University of Paris Descartes e-mail: mohamed.nadif@parisdescartes.fr

Prediction by regression models with missing in covariates

Fumiya Sano, Kaito Takano, Shintaro Tomatsu, and Manabu Iwasaki

Abstract Missing data is an inescapable problem for every research workers in various research fields. There are two types of missingness in regression analysis. One is missing in outcome variable, and the other is missing in covariates. We will discuss the latter situation that missing may occur in covariates. Many of previous researches on missing data have focused on estimation of regression coefficients of regression models with missing in covariates. However it is often the case that the primary purpose of statistical analysis is prediction of outcome variable. We here focus on the prediction with missing in covariates. The purpose of our study is to predict outcome variable by multiple regression using dataset with possible missing in covariates. Even if the regression coefficients are estimated using some conventional methods that incorporates missingness, when new data with missing in covariates is obtained, it is not so obvious to predict the outcome of such new data. Our present study is an attempt to compare the performance of prediction among five methods. Five methods considered here are as follows: (i) multiple regression using observed covariates only; (ii) missing indicator method; (iii) imputation of missing part of covariates based on observed part; (iv) imputation based on principal component analysis and factor analysis; (v) hot deck method based on matching. We conduct Monte Carlo simulations to assess the prediction performance of each method considered under the three missing data mechanisms (MCAR, MAR, MNAR).

Keywords: Missing Data; Regression Analysis; Prediction

Fumiya Sano
Graduate Student, Graduate School of Science and Technology, Seikei University e-mail: us082053@gmail.com

Kaito Takano
Graduate Student, Graduate School of Science and Technology, Seikei University e-mail: dm166208@cc.seikei.ac.jp

Shintaro Tomatsu
Graduate Student, Graduate School of Science and Technology, Seikei University e-mail: dm166210@cc.seikei.ac.jp

Manabu Iwasaki
Professor, Department of Computer and Information Science, Seikei University e-mail: iwasaki@st.seikei.ac.jp

Integrating customer data sets using topic models

Takuya Satomura

Abstract A broad range of customer data is gathered by many companies for the purpose of getting customer insights. These data sets include customer purchase data and customer survey data. The approach which analyzes these data sets simultaneously will be useful to obtain a unified view of the customer. The author proposes latent variable models, which enable a company to create valuable customer insights by integrating customer data sets together. The proposed models are the extension of the Joint Topic Model and can handle different types of data at the same time. In these models, latent topics which are unobserved work to combine multiple data. The concomitant variables are also used to explain the meaning of latent topics. With applying the proposed model to both customer purchase and survey data, these models extract the integrated topics which simultaneously represent the underlying motivation of purchasing and the lifestyle of customers. The proposed models are also applicable to data fusion. By using the estimate results of the proposed models and the new data of customer purchase, the models can predict the lifestyle of the customer whom a company knows only purchase data. In an empirical analysis, the author applied proposed models to the data from an online retailer. The author discusses the predictive accuracy of the customer behavior and the usefulness of customer insights drawn from the analysis.

Keywords: Machine Learning; Bayesian Approach; Data Fusion

Takuya Satomura
Keio University e-mail: satomura@fbc.keio.ac.jp

Calculating neural reliability from EEG recordings of naturalistic stimuli

Pieter C. Schoonees and Niël J Le Roux

Abstract Evidence has emerged from the neuroscience literature that the level of intersubject synchronization between the neural responses of different subjects to, for example, movies is related to population-level measures of movie success, such as box office performance. Measures of such intersubject similarity are also known as neural reliability measures. The assumption is that the more engaging a naturalistic stimuli such as a movie is, the more similar the responses are even when comparing across subjects. Several studies have shown this empirically, using a variety of methods including correlation-based distance measures and component analysis techniques similar to canonical correlation analysis. In this paper, we discuss these approaches and the statistical assumptions they make, and compare the existing methods to new methods for quantifying neural reliability.

Keywords: component analysis; quantification; EEG

Pieter C. Schoonees
Rotterdam School of Management, Erasmus University e-mail: schoonees@rsm.nl

Niël J Le Roux
University of Stellenbosch e-mail: njlr@sun.ac.za

A bayesian "confirmatory" factor analysis applied to WISC data

Kazuo Shigemasu and Masanori Kono

Abstract Factor Analysis methods can be classified into two kinds, namely, exploratory methods and confirmatory methods. In order to confirm a certain structure of factor pattern, some elements of the factor loadings are prefixed, typically to be zero. But it should not be likely that the true factor loadings are zero. Bayesian approach to factor analysis is a good compromise between two kinds of factor analysis. It can provide estimates of all factor loadings based on the hypothesized factor pattern, and moreover, the Bayesian approach makes it possible to compare several models (i.e. hypothesized patterns) and select the best one. In this paper, Bayesian confirmatory factor analysis method is applied to the intelligence test scores, more specifically, WISC data which were gathered for the standardization purpose in Japan. The purpose of the analysis is to answer two research questions, which have been disputed for some time. WISC has 15 scales, which are classified into 4 group factors. Recently, CHC (Cattel-Horn-Carrol) theory has been dominant in the factorial theory of intelligence, and reorganization of WISC scales from the viewpoint of CHC theory has been attempted. We compare two models (i.e. WISC original model and CHC model). The second question in this research is to determine which of the single factor (g-factor) model or the tow factor model is more appropriate at the higher level of intelligence structure. Using a hierarchical modeling of factor pattern, we answer the long-standing question "Does g-factor exist?".

Keywords: Bayesian Hierarchical modeling; Hyper parameters

Kazuo Shigemasu
Visiting Professor, Keio University e-mail: kshigemasu@gmail.com

Masanori Kono
Assistant, Teikyo University e-mail: masa-k4@main.teikyo-u.ac.jp

Dissimilarity by chi-squared statistic for aggregated symbolic data with continuous and categorical variables

Nobuo Shimizu, Junji Nakano, and Yoshikazu Yamamoto

Abstract In these days, huge amount of individual data are frequently available in all fields of science, engineering and social activities. For such "big data", new useful data representation and analyzing methods are required. Symbolic data analysis provides techniques for handling them. Traditional symbolic data analysis uses information only about marginal distribution of each variable. We consider that individual data can be divided into some naturally defined groups and be expressed by information about both marginal distributions and a joint distribution of each group. We also assume that individual data consists of continuous and categorical variables. We use means, variances and covariances for summarizing continuous variables, and contingency tables for summarizing categorical variables. We call them "aggregated symbolic data (ASD)". For clustering such ASD, we use Pearson's chi-squared statistic between ASD as a dissimilarity. Some example data are analyzed by the proposed method.

Nobuo Shimizu
The Institute of Statistical Mathematics e-mail: nobuo@ism.ac.jp

Junji Nakano
The Institute of Statistical Mathematics e-mail: nakanoj@ism.ac.jp

Yoshikazu Yamamoto
Tokushima Bunri University e-mail: yamamoto@fe.bunri-u.ac.jp

Latent probabilistic modeling for mutational signature in cancer genomes

Yuichi Shiraishi

Abstract Thanks to the advances in recent high throughput sequencing technologies, large amounts of somatic mutations from cancer genome have been accumulating all over the world, and it is now possible to extract characteristic patterns of somatic mutations or "mutation signatures," which reflect sources of somatic mutations (e.g., tobacco, ultraviolet and so on), at an unprecedented resolution. Also, it is expected that revealing novel mutation signatures can lead to identification of novel mutagens and prevention of cancer.
For identifying characteristic mutation signatures from large amounts of somatic mutation data, efficient latent variable models are necessary. Currently, the most popular method is a framework based on nonnegative matrix factorization (Alexandrov et al. 2013). However, in this approach, since the number of parameters increases exponentially as we increase contextual factors to take into account, we could not deal with many contextual factors simultaneously due to instability of estimates. Furthermore, interpretation of mutations signatures of huge dimensional vectors is often difficult.
In this presentation, we will first introduce our novel approach based on hierarchical probabilistic modeling (Shiraishi et al., 2015, https://github.com/friend1ws/pmsignature). Assuming the independence on each factor of mutation signatures and reducing the number of parameters, more robust and interpretable estimates can be obtained. Additionally, the proposed model has close relationships with the "mixed-membership models," that have been intensively utilized in statistical machine learning and statistical genetics community. We demonstrate that the proposed approach can identify several novel features of mutation signature such as base frequency at the two base 5' to the mutated sites. Second, we will also introduce a new framework integrating various types of somatic mutations such as insertion, deletion and dinucleotide substitutions.

Keywords: mutation, latent variable mode

Yuichi Shiraishi
Human Genome Center, Institute of Medical Science, The University of Tokyo e-mail: yshira@hgc.jp

Inference for general MANOVA based on ANOVA-type statistic

Łukasz Smaga

Abstract Inference methods for general multivariate analysis of variance (MANOVA) are studied. Homoscedasticity as well as any particular distribution are not assumed in general factorial designs under consideration. In that general framework, the existing methods based on the Wald-type statistic may behave poorly under finite samples, e.g., they are often too liberal under unbalanced designs and skewed distributions. In this paper, the testing procedures and confidence regions based on the ANOVA-type statistic and its standardized version are proposed, which usually perform very satisfactorily in cases where the known tests fail. Different approaches to approximate the null distribution of test statistics are developed. They are based on asymptotic distribution and bootstrap and permutation methods. The consistency of the asymptotic tests under fixed alternatives is proved. In simulation studies, it is shown that some of the new procedures possess good size and power characteristics, and they are competitive to existing procedures.

Keywords: ANOVA-type statistic; bootstrap method; confidence region; general MANOVA; hypothesis testing; permutation method

Łukasz Smaga
Faculty of Mathematics and Computer Science, Adam Mickiewicz University e-mail: ls@amu.edu.pl

Spatial diversification of employment structures vs. educational capital in the European Union

Elżbieta Sobczak and Beata Bal-Domańska

Abstract The purpose of the study is to assess the impact of educational capital quality on the spatial concentration level of employment structure. It was attempted to identify the growth poles and peripheral regions in the European space, taking into account both the spatial dependence and the level of educational capital quality. The study attempts to answer the research question whether high level of educational capital quality constitutes an important determinant of employment spatial concentration in the so-called smart sectors? The research used the spatial statistics (*I Moran, join count*) linear ordering methods and the methods of multivariate statistical analysis. Due to the growing significance of knowledge and innovation the analysis covered employment structures in the economy sectors identified based on the intensity of knowledge (defined in case of manufacturing sector by expenditure on research and development, and in case of services by a level of tertiary educated persons) including high and medium high-technology manufacturing, mid-low and low-technology manufacturing, knowledge-intensive services, less knowledge-intensive services. The time span of the research covered the period 2008-2015.

Keywords: educational capital, employment structures, spatial statistics, inear ordering methods and the methods of multivariate statistical analysis

Elżbieta Sobczak
Dean of Faculty of Economics, Management and Tourism in Jelenia Góra Wrocław University of Economics, Department of Regional Economy, Faculty of Economics, Management and Tourism in Jelenia Góra, Wrocław University of Economics (Poland) e-mail: elzbieta.sobczak@ue.wroc.pl

Beata Bal-Domańska
Department of Regional Economy, Faculty of Economics, Management and Tourism in Jelenia Góra, Wrocław University of Economics (Poland) e-mail: beata.bal-domanska@ue.wroc.pl

The IPUMS approach to harmonizing the world's population census data

Matthew Sobek

Abstract IPUMS is the world's largest collection of population microdata available for research and education. The project integrates census data from 85 countries into one consistent database. The signature feature of IPUMS is to harmonize variables across countries and over a fifty year period, so the same code has the same meaning in all times and places. The anonymized microdata files provided by the partner Statistical Offices come to IPUMS coded into a wide variety of classification schemes utilized at the time the data were originally processed in the different countries. To enable comparative research requires coping with this empirical reality of pre-classified microdata. This paper will describe the general principals IPUMS follows in the harmonization and classification process. We will provide concrete examples of the challenges the project faces in applying international standards to data produced over a long period by dozens of statistical offices. One strategy involves the use of composite codes in which leading digits for a variable apply generally across countries and trailing digits retain details not universally available. Another approach is to provide parallel variables geared to subsets of samples that adhere to particular international classification standards. The web dissemination system is integral to the IPUMS approach to data harmonization. Users need to be able to easily explore the contents of the database and examine the classification structures of the variables to assess their utility while defining their customized data extracts. To aid this process IPUMS provides documentation highlighting comparability issues that may not be self-evident from the codes and links the questionnaire text to the variables. We will close by describing how researchers have used the data series: the topics they address and variables they access.

Keywords: population; census; metadata; harmonization

Matthew Sobek
Research Scientist, PhD Director of Data Integration, University of Minnesota e-mail: sobek@umn.edu http://users.pop.umn.edu/ sobek/

How to cross the river? – new distance measures

Andrzej Sokołowski, Małgorzata Markowska, Sabina Denkowska, and Dominik Rozkrut

Abstract If you are on one side of the river and have to arrive at some point on the other side, the distance you have to cover can be calculated in different ways. You can jump to the other side which is equivalent to euclidean distance, or cross the river perpendiculary and follow your way on the other side to a given point, along Manhattan distance. If there are stones in the river, you can jump from one to another. This is the idea of the first of our propositions for a new distance measure. You can move to the closest point until you arrive at the given one, but you can visit each point only once. The other idea is to reposition some stones and let them lie in a line, but the stones can be taken only from a given distance from this line. This is the idea of so called tunnel distance. The third idea takes also into account the longest distance between consecutive points on the way to a given one. In the paper we present the simulation study of new distances under null hypothesis of multivariate normal distribution and also under a mixture of four normals with different locations.

Andrzej Sokołowski
Department of Statistics Professor, Cracow University of Economics e-mail: sokolows@uek.krakow.pl

Małgorzata Markowska
Wrocław University of Economics e-mail: malgorzata.markowska@ue.wroc.pl

Sabina Denkowska
Cracow University of Economics

Dominik Rozkrut
Central Statistical Office

Granular credit data classification with SVM based approaches

Ralf Werner Stecking

Abstract Forecasting whether a credit applicant is likely to default is an important problem the financial industry poses to statistical modeling. There is a broad spectrum of statistical models, ranging from traditional Logistic Regression to the more recently developed Support Vector Machines (SVM) and especially their nonlinear kernel variants. Granular computing is a recent approach of information processing, consisting of a variety of theories, methodologies, techniques, and tools that make use of granules in the statistical data modeling process. A granule is one of the basic elements (or group of elements) that are the ingredients of larger units, like e.g. cluster, classes, regions or rules. These units may also incorporate a granular structure like e.g. a hierarchy or a network. The representation or description of granules is called granulation. A granular perspective of credit data may be necessary for very large data sets or transaction data in different time resolutions, it may be useful for data privacy concerns, and may occur as a result of data base queries or rule extraction methods. In view of using granular credit data sets, one could ask to what extend is classification model performance affected by aggregating the original case by case or client wise information? However, such compression is expected to preserve enough information from the original data, such that classification based on such granular data will be possible at all. But the consequences for statistical data modeling are quite unclear. To what extent will granular input affect the classification power of credit data models? Furthermore, appropriate descriptions for granular data, which allow a more sophisticated coding of categorical, quantitative or mixed variables, are examined. How does classification performance depend on granulation complexity, i.e. different levels of granularity? Finally, SVM as well as traditional Logistic Regression classification models with different types of granular data are evaluated and compared to models that are trained and validated on the original data.

Keywords: Granular Data; Support Vector Machines; Credit Scoring

Ralf Werner Stecking
University of Oldenburg Department of Economics e-mail: ralf.w.stecking@uol.de

Analysis of expenditure patters of virtual marriage households consisting of working couples synthesized by statistical matching method

Mikio Suga and Yasuo Nakatani

Abstract This paper explores the impact of marriage on income and expenditure by using the micro-data from National Survey of Family Income and Expenditure. Although benefits of marriage are generally recognized, the marriage rate has been declining in Japan. The reasons could attribute to unclear advantages of settling down. In this study we randomly synthesized a single-person male and a single- person woman on the data and created virtual married conditions. The income and expenditure of these virtual couples were compared with those of actual married couples. The results indicate that there are significant differences in expenditure patterns between these groups.

Keywords: Public statistics; Micro data; Statistical matching method

Mikio Suga
Faculty of Economics, Hosei University e-mail: msuga@hosei.ac.jp

Yasuo Nakatani
Faculty of Economics, Hosei University e-mail: ynakatanister@gmail.com

Compare of Taiwan import Japanese fruits and vegetables

Yukiya Suzuki and Yumi Asahi

Abstract According to the report written by Japanese MAFF (Ministry of Agriculture, Forestry and Fisheries), Japan's market of agricultural products is likely to decrease because of low birth rate and longevity. The world's food market, in the meanwhile, is estimated to double from 340 trillion yen in 2009 to 680 trillion yen in 2020. Especially Asian food market is estimated to increase threefold from 82 trillion yen to 229 trillion yen with the growth of population and wealthy people. Japanese government is trying to take advantage of the growth of world and Asian market in order to make the domestic agriculture, forestry and fisheries into the growing industry. Japanese MAFF has set a goal to expand the export figures of agricultural, forestry, fisheries and food to a trillion by 2020.Exporting Japanese agricultural products is important task for Japanese agriculture, and need inquest whether to expansion export future. On the matter of the export of the agricultural products, this paper founded the possibility of expansion of exporting by Time-series Analysis for Taiwan import the agricultural products. Research object is only Taiwan because Taiwan is physical proximity to Japan, and many year's trading partner. In order to explore the potential for the expansion of exporting Japanese agriproducts, this research aimed to analyze Taiwan's import figures of the agriproducts. This analysis founded the import figures of increasing tendency use Time-series Analysis and compared time series fruits to vegetables. Using the results of this analysis, this report revealed the prospect of the export of the agricultural products.

Keywords: Import figures, Taiwan, Time-series Analysis

Yukiya Suzuki
Tokai University, Graduate School of Information and Telecommunication Engineering, Course of Information and Telecommunication Engineering e-mail: 7bjnm015@mail.u-tokai.ac.jp

Yumi Asahi
Tokai University, School of Information and Telecommunication Engineering, Department of Management Systems Engineering e-mail: asahi@tsc.u-tokai.ac.jp

Estimation methods based on weighted cluster analysis

Roland Szilágyi, Beatrix Varga, and Renáta Géczi Papp

Abstract Research based on samples and their conclusions play an increasing role in making business decisions and also in creating information. The study gives an overview about the different estimation methods, which can be used for the description of the socio-economic relations. Consecutively, the statisticians searched for solution variants for handling of mistakes, based on different distributions. Identifying tendencies plays a significant role in eliminating the bias caused by failed of assumption. It is worth examining the differences between tendencies of different groups. In order to achieve a successful procedure the sample should be grouped based on variables which are in stochastic relation with the examined criterion. The tendencies must be examined with the help of simulation and modelled according to this. That was the reason why the researchers created an estimating model of weighted tendencies, in which the estimated values of the above-mentioned tendencies were defined as average estimated values by weighting the explanatory features of function.

Keywords: Estimation; Cluster Analysis; Weighting

Roland Szilágyi
Department of Business Statistics and ForecastingUniversity of Miskolc, Hungarian Statistical Association
e-mail: strolsz@uni-miskolc.hu

Beatrix Varga
Department of Business Statistics and ForecastingUniversity of Miskolc, Hungarian Statistical Association
e-mail: stbea@uni-miskolc.hu

Renáta Géczi Papp
Department of Business Statistics and Forecasting, University of Miskolc e-mail: geczipapp.renata@gmail.com

A new statistical matching methodology using multinomial logistic regression and multivariate analysis

Isao Takabe and Satoshi Yamashita

Abstract Statistical matching techniques aim to build one dataset by combining two or more data sources about the same target population. The most important objective of these techniques is to create useful and informative synthetic microdata without conducting any survey or collecting additional data. In recent years, matching techniques have been employed in various fields including marketing, econometrics, social sciences, etc. However, because of the difficulties in dealing with big datasets and skewed variables, there are only a few applications to firm data. In this study, we proposed a new statistical matching methodology for firm datasets by employing multinomial logistic regression and multivariate analysis. Using some common transformed variables, we calculated the distance between record pairs from the two datasets. Then, the distance values obtained were used to compute the probability of "true match" pairs through multinomial logistic regression. These probability values help classify the record pairs as truly matched or not. It is worth noting that working with large datasets entails a considerable amount of time to calculate the distances of all the possible pairs. To address this problem, we resorted to the principal component analysis method by dividing the donor data sources and making strata to shrink the searching space of the record pairs and search the true matched pairs more efficiently. We applied these techniques to a commercial firm dataset and the official economic census microdata, which is the unbiased true population. The results showed that our method performs better than the methods used in the previous study in terms of speed and misclassification rates.
This work was supported by JSPS KAKENHI Grant Numbers JP16H02013 and 15H03390.

Keywords: Statistical Matching; Multinomial Logistic Regression

Isao TAKABE
The Graduate University for Advanced Studies (Statistics Bureau, Ministry of Internal Affairs and Communications)
e-mail: takabe@ism.ac.jp

Satoshi YAMASHITA
Professor, The Institute of Statistical Mathematics e-mail: yamasita@ism.ac.jp

Clustering methods for preference data in the presence of response styles

Mariko Takagishi, Michel van de Velden, and Hiroshi Yadohisa

Abstract In questionnaire-based survey, preference data such as Likert scale data is often obtained. Also, by clustering individuals based on survey item, latent cluster structure can be discovered. However, the existence of response style, which can be defined as individual's tendency to respond to questionnaire items regardless of contents, may affect the result of cluster analysis for preference data. For example, a cluster of individuals with extreme response style, which is a tendency to choose categories at the end of the scale, can be mistakenly identified as an item-based cluster. Therefore, we propose a new method to cluster individuals while correcting response style bias. Specifically, we identify individual's response styles by the shape of function which corresponds observed responses with unobserved latent responses which are not affected by response styles, and correct response style bias by the estimated function. Parameters in proposed method can be estimated by the algorithm that combines Alternative Least Squares and k-means type algorithms.

Keywords: Clustering; Response Style; Preference Data; Ordinal Categorical Data

Mariko Takagishi
Doshisha University, graduate school of Culture and Information Science e-mail: applesan728@gmail.com

Michel van de Velden
Erasmus School of Economics, Erasmus University Rotterdam e-mail: vandevelden@ese.eur.nl

Hiroshi Yadohisa
Doshisha University, department of Culture and Information Science e-mail: hyadohis@mail.doshisha.ac.jp

The effects of natural disasters on household income and poverty in rural Vietnam : an analysis using Vietnam household living standards survey

Rui Takahashi

Abstract This study clarifies the relationship between natural disasters, household income and poverty using micro econometric methods. The effects of natural disasters on household income and poverty and its relationship with inequality receives international attention from the likes of The United Nations and The World Bank. Even though the risk of natural disasters is higher in developing countries than in developed countries, the ability to cope with natural disasters is poor in developing countries. One of the regions suffering great damage from natural disasters is rural Vietnam, where many people live in poverty. Therefore, I analyzed the effects of natural disasters on household income and poverty in rural Vietnam using VHLSS (Vietnam Household Living Standards Survey) data.
VHLSS contains household and rural commune data. Information on natural disasters is included only in rural commune data. I used household data and commune data in rural areas of Vietnam to estimate the conditional quantile using quantile regression. Through quantile regression I was able to use the Koenker and Bassett estimator to clarify the relationship between natural disasters (the three main types being storms, floods and droughts) and household income and poverty.
As a result, we can confirm the following findings; Firstly, floods cause significant negative effects to middle and high income classes in rural Vietnam. Secondly, storms have a negative effect on the middle income class in rural Vietnam. I can therefore assume that the assets of the middle and high income classes are more likely to be damaged due to floods and storms. Thirdly, droughts have serious effects on all income classes including the poverty group. Therefore, we need to consider countermeasures against droughts to alleviate poverty.

Keywords: Micro data ; Quantile regression ; Vietnam

Rui Takahashi
Associate Professor, Department of Economics, Department of Economics, School of Political Science and Economics, Tokai University e-mail: ruita@keyaki.cc.u-tokai.ac.jp

Bayesian dynamic topic modeling with stable topics over time periods

Keisuke Takahata and Takahiro Hoshino

Abstract Latent Class Modeling (LCA) has been frequently applied in social sciences such as marketing, sociology and psychology. In recent years, instead of LCA, Latent Dirichlet Allocation (LDA), which was originally developed in the area of statistical latent semantic analysis, has been used especially in marketing data analysis, in which the latent structure of high dimensional categorical variables such as purchase behavior are of interest. The distinctive feature of LDA compared to LCA is that LDA can classify both variables and objects simultaneously by using probabilistic function of topics. To treat dynamic structure of the outputs from LDA, Blei et al. (2006) proposed Dynamic Topic Modeling (DTM), in which the time dependence of parameters is expressed as a state space model. This model specification inevitably generates new topics and all the topics in the last period always disappear. We believe this aspect of the DTM is the reason why the DTM is not often used in application studies, although this model is considered very general one. In this talk we propose a new variant of dynamic topic models which enables some topics to remain stable over time periods. We develop a MCMC estimation algorithm for the proposed model, and will be applied to location log data.

Keywords: Latent Dirichlet Allocation;Dyanamic Topic Modeling

Keisuke Takahata
Department of Statistics, Keio University e-mail: bayesian@jasmine.ocn.ne.jp

Takahiro Hoshino
Department of EconomicsProfessor, Keio University e-mail: bayesian@jasmine.ocn.ne.jp

Probability weighting function and time discounting function in decision making: Theory and experimental analysis

Kazuhisa Takemura and Hajime Murakami

Abstract We showed probability weighting functions ($w(p)$) derived from time discounting function (D) and axiomatic basis of the models. Probability weighting function ($w(p)$) is considered to be a nonlinear function of objective probability &(pi n behavioral decision theory. This study proposes a psychometric model of probability weighting functions derived from a generalized hyperbolic time discounting model. Since the expected value of a geometric distributed random variable X is $1/p$, we formulized the probability weighting function of the expected value model for exponential time discounting model. We also derived a model from the generalized exponential time discounting model assuming Fechner's psychophysical law of time. To illustrate the fitness of each model, a psychological experiment was conducted for assessing the probability weighting and value functions at the level of the individual participants. The results of individual analysis indicated that the expected value model of generalized hyperbolic discounting fit better than previous probability weighting decision-making models. We also re-analyzed the time discounting data of previous published data, and found that the generalized hyperbolic time discounting model fit better than other models such as exponential type models. We showed an axiomatic system of the generalized hyperbolic model of probability weighting function and considered psychological basis of the model.

Keywords: Probability weighting function; Time dsiscounting; Decision making; Decision under risk

Kazuhisa Takemura
Department of Psychology, Waseda University e-mail: kazupsy@waseda.jp

Hajime Murakami
Department of Psychology, Waseda University e-mail: whassjeidmae@gmail.com

Bayesian interpretation of the ℓ_0 penalized linear regression estimator

Ryunosuke Tanabe

Abstract A lot of researches have shown that the penalized regression estimator can be interpreted as a Bayes model. The advantage of the Bayesian penalized model is that the credible interval can be calculated without additional efforts in the Bayesian framework. Park and Casella (2008) proposed a Bayesian lasso and showed the efficient Gibbs sampler for the model. As a generalization of the Bayesian lasso, the Bayesian fused lasso, the Bayesian elastic net, and the Bayesian group lasso are proposed. For the normal linear regression model, the generalized information criterion (GIC) (Nishii et al. (1984)) is equivalent to ℓ_0 penalized estimator. The GIC contains Akaike information criteria and the Bayesian information criterion. In our proposed Bayesian ℓ_0 penalized regression model, the posterior mode is equivalent to an estimator of non-Bayesian ℓ_0 penalized regression and the posterior distribution can be obtained by Gibbs sampling. Using the posterior median or the credible interval, we can select an adequate model and estimate its coefficients. Our proposed method can automatically select the tuning parameters using the hierarchical or the empirical Bayes method. Furthermore, we can obtain credible intervals for estimated coefficients of the variable selection. The model selection results in both simulation and real data analysis by the proposed method are very similar to those by the non-Bayesian methods.

Keywords: L0 penalized estimator; sparse estimation; bayes

Ryunosuke Tanabe
Osaka University e-mail: tanabe@sigmath.es.osaka-u.ac.jp

Dimension reduction clustering based on constrained centroids

Kensuke Tanioka

Abstract In these days, multivariate cateogorical data becomes large and complex through the improved information technology. In such a situation, dimension reduction clustering method is useful since it is easy to interpret the features of clusters from the estimated low-dimensions. There are several previous studies for dimension reduction clustering for multivariate categorical data. However, centroids estimated by existing methods are represneted as means on the estimated low-dimensions, although these data are originally represented as categories. Then, new dimensions reduction clustering such that centorids are constrained to coordinates of categories is proposed. From the proposed method, we can interpret the clusters easily through the constrained centroids.

Keywords: k-mode; categorical data

Kensuke Tanioka
Wakayama-medical University e-mail: ktaniok@wakayama-med.ac.jp

Usefulness of factor analyses for validity evaluation of a foot and ankle related quality of life outcome instrument

Shinobu Tatsunami, Takahiko Ueno, Naoki Haraguchi, and Hisateru Niki

Abstract One of the necessary requirements for the validity evaluation of self-reported outcome instruments is the assessment of construct validity. To this end, we used confirmatory factor analysis (CFA) and expiratory factor analysis (EFA) in the development of an outcome instrument for foot and ankle related quality of life. The instrument was finalized based on a total of three field studies. Then, it was used for the purpose of responsiveness analysis from 2013 to 2016. The present instrument consists of 43 questionnaire items and provides five subscale scores, namely "Pain and pain-related (PP)", "Physical functioning and daily living (PF)", "Social functioning (SF)", "Shoe-related (SH)", and "General health and well-being (GH)". In addition, it provides an optional subscale score "Sports activity (SP)". By using CFA, we adjusted the items so that all factor correlation coefficients became smaller than 0.9, and correspondence between extracted factors and expected subscales became more rational compared to the results from the instrument before revisions. Looking at the structure of the correlation matrix from the data of the fourth study, the correlations between subscales were within the expected ranges. For example, the maximum value of the correlation coefficient between PF and SF was observed as 0.84 and the minimum between SF and SH as 0.54 before hallux valgus surgery. The observed values of the elements in the correlation matrix did not show any remarkable changes after the surgery.Construct validity is sometimes evaluated by referring to the results from additional outcome instruments. However, it is sometimes impossible to introduce additional instruments. CFA could provide a measure of divergence and convergence in construct validity. In this context, the use of CFA was very useful in the development of the present instrument.

Keywords: confirmatory factor analysis; expiratory factor analysis

Shinobu Tatsunami
Unit of Medical Informatics, St. Marianna University School of Medicine e-mail: s2tatsu@marianna-u.ac.jp

Takahiko Ueno
Unit of Medical Informatics, St. Marianna University Scool of Medicine e-mail: t2ueno@marianna-u.ac.jp

Naoki Haraguchi
Department of Orthopedic Surgery, Tokyo Metropolitan Police Hospital e-mail: naokihg@aol.com

Hisateru Niki
Department of Orthpedic Surgery, St. Marianna University School of Medicine e-mail: h2niki@marianna-u.ac.jp

Regularized generalized canonical correlation analysis 2.0

Arthur Tenenhaus, Michel Tenenhaus, and Patrick J.F. Groenen

Abstract A new framework for sequential multiblock component methods is presented. This framework relies on a new version of regularized generalized canonical correlation analysis (RGCCA) where various scheme functions and shrinkage constants are considered. Two types of between block connections are considered: blocks are either fully connected or connected to the superblock (concatenation of all blocks). The proposed iterative algorithm is monotone convergent and guarantees obtaining at convergence a stationary point of RGCCA. In some cases, the solution of RGCCA is the first eigenvalue/eigenvector of a certain matrix. For the scheme functions $x, x^2, |x|$ or x^4 and shrinkage constants 0 or 1, many multiblock component methods are recovered.

Keywords: Canonical correlation analysis; Regularization; Partial Least Squares

Arthur Tenenhaus
CentraleSupélec-L2S, UMR CNRS 8506, Gif-sur-Yvette, France and Bioinformatics- Biostatistics Platform IHU-A-ICM, Paris, France e-mail: arthur.tenenhaus@centralesupelec.fr

Michel Tenenhaus
HEC Paris, Jouy-en-Josas, France e-mail: tenenhaus@hec.fr

Patrick J.F. Groenen
President IASC e-mail: groenen@ese.eur.nl

Statistically significant pattern mining with applications to biology

Aika Terada

Abstract Statistical significance (p-value) is an important measure in biology and medical science, while there is few methods that can assess statistical significance of results in in pattern mining including frequent itemset mining and association rule discovery. Enumeration of statistically significant patterns is not only computationally non-trivial but also extremely unlikely due to multiple testing correction. The exponential growth of the number of patterns generally results in a very large p-value correction factor, which causes extremely low sensitivity. In this talk, we introduce a multiple testing method to enumerate statistically significant patterns named Limitless Arity Multiple testing Procedure (LAMP). LAMP counts the exact number of testable patterns and calculates the correction factor to the smallest possible value. Applying it to biological data analyses, such as combinatorial regulation discovery and genome-wide association study, LAMP detected statistically significant patterns that contains three or more items which were overlooked by a traditional multiple testing procedure. Our approach may enable us to easily apply the pattern mining methods to biology and medical science studies.

Keywords:

Aika Terada
PRESTO, JST e-mail: terada@cbms.k.u-tokyo.ac.jp

Semi-supervised learning for functional data

Yoshikazu Terada

Abstract In the usual binary classification problem, the training data consists of labeled positive and negative examples. However, in various practical situations, it is often the case that we have only labeled positive examples, and unlabeled examples, which consists of both positive and negative examples, as the training data. In this semi-supervised learning setting, the aim is to assign labels to the unlabeled examples.In this talk, we have studied binary classification from only positive and unlabeled functional data. Our first contribution is to present a simple classification algorithm for this problem. The key feature of the algorithms is that it is not required an estimation of the unknown class prior. Our second contribution is to prove that, under mild regularity conditions similar to those in a supervised context, the proposed algorithm can achieve perfect asymptotic classification. In fact, we show that the proposed algorithm works well not only in numerical experiments but also for real data examples.

Keywords: Classification; Functional data analysis

Yoshikazu Terada
Division of Mathematical Science for Social Systems, Graduate School of Engineering Science, Osaka University.
e-mail: terada@sigmath.es.osaka-u.ac.jp

Projection based clustering

Michael Christoph Thrun

Abstract Many data mining methods rely on some concept of the dissimilarity between pieces of information encoded in the data of interest. These methods can be used for cluster analysis. However, no generally accepted definition of clusters exists in the literature [Hennig et al., 2015]. Here, consistent with Bouveyron et al., it is assumed that a cluster is a group of similar objects [Bouveyron et al., 2012]. The clusters are called natural because they do not require a dissection; instead, they are clearly separated in the data [Duda et al., 2001, Theodoridis/Koutroumbas, 2009, pp. 579, 600]. These clusters can be identified by distance or density based high-dimensional structures. Dimensionality reduction techniques are able to reduce the dimensions of the input space to facilitate the exploration of structures in high-dimensional data. If they are used for visualization, they are called projection methods. The generalized $\mathbf{U}^*$-matrix technique is applicable for these and can be used to visualize both distance- and density-based structures [Ultsch/Thrun, 2017]. The idea that the abstract $\mathbf{U}^*$-matrix (AU-matrix) can be used for clustering [Ultsch et al., 2016]. The distances required for hierarchical clustering are defined by the AU-matrix [Lötsch/Ultsch, 2014]. Using this distance we propose a clustering approach for every projection method based on the $\mathbf{U}^*$-matrix visualization of a topographic map [Thrun 2017]. The number of clusters and the cluster structure can be estimated by counting the valleys in a topographic map [Thrun et al., 2016]. If the number of clusters and the clustering method are chosen correctly, then the clusters will be well separated by mountains in the visualization. Outliers are represented as volcanoes and can be interactively marked in the visualization after the automated clustering process.

Bibliography

[Bouveyron et al., 2012] Bouveyron, C., Hammer, B., & Villmann, T.: Recent developments in clustering algorithms, Proc. ESANN, Citeseer, 2012.

[Duda et al., 2001] Duda, R. O., Hart, P. E., & Stork, D. G.: Pattern classification, (Second Edition ed.), Ney York, USA, John Wiley Sons, ISBN: 0-471-05669-3, 2001;

[Hennig et al., 2015] Hennig, C., Meila, M., Murtagh, F., & Rocci, R.: Handbook of cluster analysis, New York, USA, CRC Press, ISBN: 9781466551893, 2015.

[Lötsch/Ultsch, 2014] Lötsch, J., & Ultsch, A.: Exploiting the Structures of the U-Matrix, in Villmann, T., Schleif, F.-M., Kaden, M. & Lange, M. (eds.), Proc. Advances in Self-Organizing Maps and Learning Vector Quantization, pp. 249-257, Springer International Publishing, Mittweida, Germany, 2014

[Theodoridis/Koutroumbas, 2009] Theodoridis, S., & Koutroumbas, K.: Pattern Recognition, (Fourth Edition ed.), Canada, Elsevier, ISBN: 978-1-59749-272-0, 2009.

[Thrun, 2017] Thrun, M. C.: A System for Projection Based Clustering through Self-Organization and Swarm Intelligence, (Doctoral dissertation), Philipps-Universität Marburg, Marburg, 2017.

Michael Christoph Thrun
Teaching and Research Assistant, Databionics AG,Mathematics and Computer Science, Databionics Research Group, University of Marburg, Germany e-mail: mthrun@mathematik.uni-marburg.de https://www.uni-marburg.de/fb12/datenbionik/

[Thrun et al., 2016] Thrun, M. C., Lerch, F., Lötsch, J., & Ultsch, A.: Visualization and 3D Printing of Multivariate Data of Biomarkers, in Skala, V. (Ed.), International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), Vol. 24, Plzen, http://wscg.zcu.cz/wscg2016/short/A43-full.pdf, 2016.

[Ultsch et al., 2016] Ultsch, A., Behnisch, M., & Lötsch, J. : ESOM Visualizations for Quality Assessment in Clustering, In Merényi, E., Mendenhall, J. M. & O'Driscoll, P. (Eds.), Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016, (10.1007/978-3-319-28518-4_3pp. 39-48), Cham, Springer International Publishing, 2016.

[Ultsch/Thrun, 2017] Ultsch, A., & Thrun, M. C. : Credible Visualizations for Planar Projections, in Cottrell, M. (Ed.), 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), IEEE Xplore, France, 2017.

Keywords: Cluster Analysis; Projection Methods; Dimensionality Reduction; Visualization

Mixture modelling of multivariate and / or longitudinal data: Arriving at insightful representations

Marieke E. Timmerman

Abstract Mixture modelling has gained great popularity within the behavioural and health sciences. In these applications, though some might have hoped for more, mixture modelling must be seen as an approximating device rather than a tool to identify truly existing groups. In empirical modelling, the challenge is to obtain a model that both adequately describes the joint distribution and enables insight into the distinguishing features of interest. To arrive at such a model, one needs a versatile modelling framework. I will provide an overview of mixture models suitable for multivariate data, including confirmatory and exploratory factor approaches, and for longitudinal data, including polynomial and spline approaches. If desirable, these model building blocks can be combined. I will discuss how the specific model choices may steer the model estimates, and thus the identification of groups, and how to deal with this in empirical modelling. The model selection and usefulness of the models will be illustrated with data from a large scale (n=1,831) longitudinal study towards individual courses of emotion dysregulation problems, stress and ADHD-symptoms across the adolescence.

Marieke E. Timmerman
Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen e-mail: m.e.timmerman@rug.nl

Estimating cross-source relationships from big data using component- and networks-analysis

Pia Tio, Lourens Waldorp, and Katrijn Van Deun

Abstract Network analysis has successfully been applied to many different types of psychological data, including personality, cognitive performance, and clinical symptoms. While investigating these different areas in single domains is useful, a better understanding of their structure requires an integrated analysis with several domains or sources of information. Investigating such cross-source relationships often requires large data sets containing information about individuals from multiple sources (big data). Such data are becoming more and more commonplace.
However, estimating a network using big data is not without its challenges. The dimension of the dataset, often containing more variables than observations, hinders accurate estimation of relations, even when some form of regularisation (e.g. lasso penalty) is used. Reducing the number of variables would be a straightforward way to remove (or at least reduce) this problem, except that we do not yet know which variables are involved in cross-source relationships. An additional challenge is that big data contains data from different sources that inherently may have different characteristics. For example, indicators of cognitive performance are expected to correlate much higher with one another than with indicators of gene expression. Applying network analysis to such data without taking this difference into account again leads to inaccurate estimation of relationships.
We propose the Sparse Network and Component (SNAC) model, which combines regularized simultaneous component analysis with the network framework. Here we present the results of a simulation study demonstrating the benefits of SNAC in estimating cross-source relationships from big data.

Keywords: Component Analysis; Graphical Models; Network Approach; Multi-source data

Pia Tio
Tilburg University & University of Amsterdam e-mail: piatio@gmail.com

Lourens Waldorp
University of Amsterdam e-mail: L.J.Waldorp@uva.nl

Katrijn Van Deun
Tilburg University e-mail: K.VanDeun@uvt.nl

Possibilities in welfare measures

Mónika Galambosné Tiszberger

Abstract Creation and calculation of comparable measures of welfare (or well-being) to cover most of the counties is a very difficult task. Welfare itself has more, very complex definitions, in which the extent of necessary variables (objective and subjective indicators as well) indicates a wide range. The bottleneck is usually the availability of reliable and harmonized datasets, however it is crucial from the point of view of comparability. It is a problem at the countries' level, and would be an even fiercer problem, if we go deeper into the regional level. This study aims to give a "nomenclature" of the potential variables that are used on well-being/welfare measures and to investigate the possibilities of the construction and calculation of a regional level measure, which can be used for most of the countries at least in the EU. Detailed mapping of data sources, methodological backgrounds and reliability analysis would be necessary, before we can reach our aim. Besides the suggestions on regional level measures, the paper will give an overview of the current situation about the existence of harmonised data in diverse fields of official statistics.

Keywords: welfare, well-being, regional level data

Mónika Galambosné Tiszberger
assistant professor e-mail: tiszbergerm@ktk.pte.hu

Clinical classification of cancer subtypes based on gene mutations and expressions

Shuta Tomida

Abstract Successful development of molecular targeting therapy as well as high throughput cost-effective genome-wide sequencing technology have revolutionized cancer research in the past decade. Especially in the field of lung cancer research, development of the first-generation epidermal growth factor receptor tyrosine kinase inhibitors (*EGFR*-TKIs), gefitinib and erlotinib, and also historical discoveries of association between *EGFR* kinase mutations with an enhanced sensitivity to gefitinib and erlotinib, fundamentally changed the conventional classification of non-small-cell lung cancer (NSCLC) subtypes. Profiles of oncogenic mutations with clinically applicable molecular targeting drugs (so-called "actionable" mutation), such as *EGFR*, *BRAF*, *MET* and *ALK*-fusion mutations, are the essential information not only for cancer diagnosis but also for cancer treatment nowadays. Therefore cancer classification based on those mutational information is inevitable from the clinical point of view. However, recent our research showed that gene expression profiles could distinguish subtypes of NSCLC in terms of acquired resistant mechanisms even in the same mutational profiles, suggesting the potential of hybrid classifier based on gene mutation profiles as well as gene expression profiles. In addition, we applied the same hybrid concept into the classification of cancer of unknown primary (CUP), in order to support the diagnosis of potential primary site. We will discuss the potential and also the practical structure of the hybrid classifier in this session.

Keywords: Classification; Cancer; Mutation

Shuta Tomida
Okayama University Graduate School of Medicine Dentistry and Pharmaceutical Sciences e-mail: shuta.tomida@okayama-u.ac.jp

The multiple scaled contaminated Gaussian distribution

Cristina Tortora and Antonio Punzo

Abstract The p-variate contaminated Gaussian distribution (CGD) was proposed to model data sets characterized by the presence of outliers. The CGD is a two-component Gaussian mixture; one of the components, with a large prior probability, represents the good observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the bad observations, or outliers. Compared to the multivariate Gaussian distribution, the CGD has two extra unidimensional parameters: proportion of good observations and inflation parameter. The CGD can be also seen as a Gaussian scale mixture model, with a convenient Bernoulli distribution as a weight distribution.
However, the use of univariate parameters is limiting for real applications because the proportion of outliers and their impact on the inflation parameter may be different in each dimension. Therefore, we propose a multiple scaled contaminated Gaussian distribution (MSCGD), with p-dimensional proportion and inflation parameters. The MSCGD incorporates a multi-dimensional weight within the density of a Gaussian scale mixture via an eigen decomposition of the symmetric positive-definite scale matrix, specifically p unidimensional convenient Bernoulli distributions are used as weight distributions.
The generalized EM-algorithm is used for parameters estimation. The advantages of the MSCGD are shown on real and simulated data sets.

Keywords: Contaminated Gaussian distribution; multiple-scaled distribution; mixture models; heavy tails

Cristina Tortora
Assistant professor, Department of Mathematics and Statistics, San Jose State University e-mail: cristina.tortora@sjsu.edu http://www.cristinatortora.com

Antonio Punzo
Associate professor, Department of Economics and Business, University of Catania e-mail: antonio.punzo@unict.it

An integrative tool for visualization of gene set analysis

Chen-An Tsai

Abstract Gene set enrichment analyses (GSEA) provide a useful and powerful approach to identify differentially expressed gene sets with prior biological knowledge. Several GSEA algorithms have been proposed to perform enrichment analyses on groups of genes. However, little attention has been paid to address the association between gene sets. In this study, we propose a general strategy of gene set correlation analysis (GSCA) to uncover novel gene set associations by integrating prior biological knowledge and gene expression data. A multivariate method is applied to the comparisons of cross-gene sets in gene expression data to identify the common relationships in expression profiles between pairs of gene sets. We also combine between-gene set correlation analysis and GSEA (gene set enrichment analysis) to discover relationships between gene sets significantly associated with phenotypes. In addition, we provide a graphical technique for visualizing and simultaneously exploring the associations of be-tween and within gene sets and their interaction and network. In the real data study, we illustrate how to combine Microarray data with RNA-Seq data through this integrative analysis.

Keywords: Gene set enrichment analyses; gene set correlation analysis; Microarray; RNA-Seq

Chen-An Tsai
Division of Biometrics, Department of Agronomy, National Taiwan University e-mail: catsai@ntu.edu.tw

Regularization method using Gini index for core array of Tucker3 model

Jun Tsuchida

Abstract Tucker3 model is popular method as principal component analysis for three-mode three-way data (object by variable by condition.) However, it is difficult for interpret the parameters of Tucker3 model, because core array represents the interactions between object, variable and condition. To overcome this problem, many researchers have proposed methods, such as L1 regularization, for obtaining sparse core array. Sparsity is important for interpret the parameter. Therefore, our approach is maximization for sparsity measure. In this paper, we propose regularization method using Gini index. Gini index has good property for sparsity. Thus, we obtain a core array, which is easily interpreted, using by maximizing Gini index directly. This method is assessed in numerical example and real data example.

Keywords: Tucker decomposition; dimension reduction; PCA

Jun Tsuchida
Graduate School of Culture and Information Science, Doshisha University e-mail: eip1003@mail4.doshisha.ac.jp

Quantitative storyline structure of ”the tale of utsuho”

Gen Tsuchiyama

Abstract This study involves quantitative research of “The Tale of Utsuho,” which is a classical literature of Japan, using a stylometric method. Stylometry is the application of the analysis of the style of a written language. The analysis of linguistic characteristics is applied to identify authorship, creation period, and creation order. By using different approaches, many studies have discussed authorship or the creation periods of historical manuscripts, such as the classical literature, by focusing on descriptive contents or investigating historical facts. However, stylometry collects quantitative data obtained from sentences, analyzes the data to understand literary styles, and then draws conclusions based on the analysis.“The Tale of Utsuho” is the first long story in Japanese literature. It comprises 20 volumes and is one of the most famous and greatest accomplishments in Japanese classical literature from the Heian period (between 794 and 1185). Although literary scholars have long debated the authorship and formation of this work, such issues have remained unresolved. Therefore, in this study, we statistically analyze the sentences of the 20 volumes using word frequency and part-of-speech component ratio. Then, the word frequency and component ratio are analyzed using principal component analysis.As a result, it is suggested that the 20 volumes can be classified into two groups with different quantitative trends. In the word frequency analysis, we observed that the groups have different appearance trends relative to postpositional particles, which are a type of Japanese function words. Moreover, in the part-of-speech component ratio analysis, we observed that one group has a high component ratio of verb and nounand another group has a high ratio of adjective and adverb. Therefore, we conclude that the writing order and the order of the chapters as they are presented are not same.T

Keywords: Stylometry, The Tale of Utsuho, function words

Gen Tsuchiyama
Organization for Research Initiatives and Development, Doshisha University e-mail: gtsuchiy@mail.doshisha.ac.jp

A study on individual small-area differences clustering of residential characteristics by any selection from differences scaling

Mitsuhiro Tsuji, Mayumi Ueda, and Toshio Shimokawa

Abstract We discuss some variable-selection approaches to realize the small-area geographical statistics of HigashiNada ward and Nagata ward in Kobe City. Higashinada ward includes the modern residential area, and Nagata ward includes the old town. We try to consider some geographical characteristics of damages after the great Hanshin-Awaji Earthquake.We can get some interesting solutions of the individual small-area differences clustering from helpful solutions of the individual differences scaling results.

Keywords: Clustering; Small Area; Geographical model; Individual model; Scaling

Mitsuhiro Tsuji
Faculty of Informatics, Kansai University e-mail: tsuji@kansai-u.ac.jp

Mayumi Ueda
University of Marketing and Distribution Sciences e-mail: mayumi_ueda@red.umds.ac.jp

Toshio Shimokawa
Wakayama Medical University e-mail: shimokaw@wakayama-med.ac.jp

Penalized multidimensional unfolding of asymmetric data with self-distances constrained to be short

Gaku Tsujii and Kohei Adachi

Abstract Multidimensional scaling (MDS) refers to procedures for analyzing a data matrix of the dissimilarities among objects to graphically display the structure underlying the data. A variety of asymmetric MDS procedures have been proposed for considering cases where dissimilarity data matrices are asymmetric. In this paper, we also consider such asymmetric cases to propose a procedure based on unfolding, which is a variant of MDS. Unfolding allows asymmetric relationships to be spatially represented simply by expressing the row and column occupied by the same object as two different points in a space. However, the unfolding approach has a drawback that it tends to provide the configurations which are not easy to interpret due to long self-distances. Here, the self-distance refers to the points for the row and column corresponding to the same object. To deal with the above drawback, we propose a penalized unfolding procedure, whose loss function is formed by combining the least squares unfolding function and a penalty one. Here, the latter is the squared sum of self-distances, which penalizes the distances for being long. For minimizing the loss function, an alternate least squares algorithm is developed, in which majorization and monotone regression steps are alternately iterated. A simulation study is performed for assessing the behaviors of the proposed procedure and its usefulness is demonstrated with real data examples.

Keywords: Multidimensional scaling; Unfolding; Asymmetric data; Majorization; Monotone regression

Gaku Tsujii
Osaka University Graduate School of Human Sciences e-mail: gagogago1220@gmail.com

Kohei Adachi
Osaka University Graduate School of Human Sciences e-mail: adachi@hus.osaka-u.ac.jp

Generalization of relationship between tweets and products sales in case of new beverage products

Hiroyuki Tsurumi, Junya Masuda, and Atsuho Nakayama

Abstract In our previous study, we analyzed the data of new beverage product A and measured the effect of TV advertising on tweets and the effect of tweets on product sales. In this study, we try to generalize these effects. Today, many marketing researchers analyze large amounts of online word-of-mouth (WOM) data were collected from SNS (Social Networking Service). However, if we continue to analyze online WOM data, we should confirm whether there is a relationship between "online WOM about the product" and "sales of the product" which is one of the objectives of marketing. It is impossible for marketer to directly control online WOM on SNS. However, if online WOM on SNS are related to product sales, we should analyze online WOM even if we cannot control it. Conversely, if it is confirmed that the WOM on SNS and product sales are irrelevant, we should change our current approach on online WOM data analysis. Therefore, we have studied the relation between online WOM on SNS and product sales. Tsurumi et al. (2013) analyzed the relation between the sales of product A (a new product of beer like-beverage), gross rating point of product A, and the number of tweets about product A using path analysis. The analysis result shows the indirect effect that TV advertisement gives to sales of product A via tweets. In this study, we analyze data of 4 new products of beverages using multiple group structural equation modeling for generalization of this indirect effect. The analysis result shows that this effect is general in the case of new products of beverages.

References

Hiroyuki Tsurumi, Junya Masuda, and Atsuho Nakayama (2013) "Analysis on Relationship between Tweets and Product Sales", "Operations Research" 58 (8), pp. 436-441 (in Japanese).

Keywords: marketing science; online word-of-mouth; social networking service

Hiroyuki Tsurumi
Yokohama National University e-mail: tsurumi@ynu.ac.jp

Junya Masuda
INTAGE Inc. e-mail: masuda-j@intage.co.jp

Atsuho Nakayama
Tokyo Metropolitan University e-mail: atsuho@tmu.ac.jp

A comparative study on rough set-based k-means clustering

Seiki Ubukata, Keisuke Umado, Hiroki Kato, Akira Notsu, and Katsuhiro Honda

Abstract Hard C-means (HCM; k-means) is one of the most famous non-hierarchical clustering method and has been utilized in many areas. Fuzzy C-Means (FCM) is proposed by Bezdek as a fuzzy extension of HCM in order to deal with vague cluster memberships and has been widely used as a flexible and robust method. Rough set theory as well as fuzzy theory is a promising approach to represent vagueness of clusters. In rough set theory, vague cluster is represented by the lower and the upper approximations considering the positive membership and the possible membership of clusters, respectively. Lingras and West firstly established a Rough k-Means (RKM) clustering scheme by introducing a viewpoint of rough set theory. RKM executes clustering detecting the positive, the possible, and the boundary regions of clusters. A variety of extensions of RKM such as Peters Rough k-Means (RKMP) has been proposed, however, RKM-based methods have a problem that they are not based on the original definitions of rough set theory. Ubukata et al. have proposed Rough Set k-Means (RSKM) method by using the original definitions of rough set theory. RSKM is regarded as a clustering model in a granulated space by a binary relation. In this research, the performances of boundary detection of these methods are compared by using the Breast Cancer Wisconsin (BCW) dataset. Rough set approaches detect uncertain objects as the boundary region and attempt to reduce misclassification in the positive regions. Especially for medical data, reduction of misclassification risk is an important issue in order to realize safe and secure clustering. Consequently, we confirmed that the proposed RSCM method provides better performance of boundary detection than conventional methods in the BCW dataset.

Keywords: Rough Set Theory; Rough k-Means; Rough Set k-Means

Seiki Ubukata
Osaka Prefecture University e-mail: subukata@cs.osakafu-u.ac.jp

Keisuke Umado
Osaka Prefecture University e-mail: umado@hi.cs.osakafu-u.ac.jp

Hiroki Kato
Osaka Prefecture University e-mail: kato@hi.cs.osakafu-u.ac.jp

Akira Notsu
Osaka Prefecture University e-mail: notsu@cs.osakafu-u.ac.jp

Katsuhiro Honda
Osaka Prefecture University e-mail: honda@cs.osakafu-u.ac.jp

Proposal of the method to diagnose brand image and ability of a respondent by using reaction time and hierarchy model

Masao Ueda

Abstract Since most brand managers are required to understand what kind of brand association consumers have and which consumers have affection to their corporate brand, this research is conducted to propose a method to diagnose brand association and the degree of consumers' affection to the corporate brand. In order to achieve this objective, the data of the strength between the corporate brand and that brand's associations were collected by internet survey and analyzed by hierarchical bayesian model. 24 brand associations, which is consisted of the product brand and corporate image, were used in the questionnaire. The strength of the connection between the corporate brand and its brand associations were measured by reaction time. The internet survey is designed to present a brand association on a smartphone to a respondent and only to ask whether that association is suited for a corporate brand or not. Reaction time indicates the difficulty to respond the questionnaire and can be measured more correctly than the previous method as like ordered alternatives. To account for the difficulty of consumers' reactions to each brand association, the model allows for heterogeneity among each brand association. Additionally, it is assumed that this model has heterogeneity among each respondent to clarify the degree of affection for the corporate brand. The result of the analysis showed that estimated parameters can discriminate which brand association is important to consumers. At the same time, the brand association that should be strengthened for the brand was also revealed. Proposed method makes it possible to assess the favorability of brand association for the corporate brand by using reaction time data and statistical model and can carry out efficient brand management. That is a contribution of this research.

Keywords: Reaction Time; Hierarchical Bayesian Model; Brand Mangement

Masao Ueda
Waseda University Graduate School of Creative Science and Enginering e-mail: m_ueda@aoni.waseda.jp

Detecting genetic association through shortest paths in a bidirected graph

Masao Ueki

Abstract Genome-wide association studies (GWASs) commonly use marginal association tests for each single-nucleotide polymorphism (SNP). Because these tests treat SNPs as independent, their power will be suboptimal for detecting SNPs hidden by linkage disequilibrium (LD). One way to improve power is to use a multiple regression model. However, the large number of SNPs preclude simultaneous fitting with multiple regression, and subset regression is infeasible because of an exorbitant number of candidate subsets. We therefore propose a new method for detecting hidden SNPs having significant yet weak marginal association in a multiple regression model. Our method begins by constructing a bidirected graph locally around each SNP that demonstrates a moderately sized marginal association signal, the focal SNPs. Vertexes correspond to SNPs, and adjacency between vertexes is defined by an LD measure. Subsequently, the method collects from each graph all shortest paths to the focal SNP. Finally, for each shortest path the method fits a multiple regression model to all the SNPs lying in the path and tests the significance of the regression coefficient corresponding to the terminal SNP in the path. Simulation studies show that the proposed method can detect susceptibility SNPs hidden by LD that go undetected with marginal association testing or with existing multivariate methods, including lasso. When applied to real GWAS data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), our method detected two groups of SNPs: one in a region containing the apolipoprotein E (APOE) gene, and another in a region close to the semaphorin 5A (SEMA5A) gene.

Keywords: bidirected graph; conservative multiple test; hidden association; linkage disequilibrium; shortest path

MASAO UEKI
Kurume University e-mail: uekimrsd@nifty.com

Generative learning with emergent self-organizing neuronal networks

Alfred Ultsch

Abstract The goal of standard classification learning is to find a conditional distribution $p(c|x)$, i.e., to identify the class c of a particular case given a data set x. By contrast, the goal of generative learning is to find the joint distribution $p(x, c)$ and to subsequently use this joint distribution to produce so far unseen values of x, which is followed by making class predictions c for these x [1]. This is typically achieved using a two-step approach comprising (A) a discrimination step aimed at construction of a model, typically by using multivariate density estimation, and (B) a data generation step that uses the model from step A to generate new data including their labels. A key issue in generative learning is an optimal estimation of probability density in multivariate data [2]. Major current implementations of generative learning comprise restricted Boltzmann machines [3] and deep generative models [4]. In the present work, a novel method is proposed that performs generative learning based on emergent self-organizing maps (ESOM), i.e., Kohonen SOMs [5] enhanced with the U-matrix and P-matrix [6].

Keywords: Generative Learning, Deap Learning, Neuronal Networks

Alfred Ultsch
University of Marburg e-mail: ultsch@informatik.uni-marburg.de

Selective inference for high dimensional classification

Yuta Umezu

Abstract Post-selection inference is a statistical inference for discovering truly important features after model or feature selection, and is being rapidly developed in statistics and machine learning community. Roughly speaking, there are two approaches, which are so-called simultaneous inference and selective inference, to conduct inference after model selection. The former is done by considering all possible candidates of the model selected by model selection procedure although the latter is done only for a selected model. As a recent advance in selective inference, in a linear regression model with Gaussian noise, Lee et. al. (2016, AS) proposed conducting hypothesis testing for coefficients selected by the Lasso by conditioning on a specific event determined by its KKT condition. In this talk, we develop their result to high dimensional classification problem. Concretely speaking, we consider a logistic regression model after feature selection based on marginal screening, and derive a high dimensional statistical behavior of the post-selection estimator by utilizing seminal results. This enable us to control a type I error rate conditional on some specific event of a certain hypothesis testing, asymptotically. Because individually extracting statistically significant features is important in practice, we attempt to detect these features by considering multiple testing. Although there are many approaches to adjust multiplicity, we simply utilize Bonferroni correction to control family-wise error rate. In addition, we conduct several simulation studies to confirm the statistical power of a significance test, and compare our proposed method with data splitting and other methods.

Keywords: Family-wise Error Rate; High Dimensional Asymptotics; Hypoth- esis Testing; Logistic Regression; Post-Selection Inference

Yuta Umezu
Nagoya Institute of Technology e-mail: umezu.yuta@nitech.ac.jp

Model based clustering for tensor-valued data strcutures

Kohei Uno

Abstract Nowadays tensor-valued data are observed frequently especially in brain imaging research and recommendation systems. Many methods for tensor-valued data have recently attracted great attention not only in machine learning but in chemometrics and psychometrics. In this study, we develop a new clustering method for tensor-valued data using tensor normal distribution. Tensor normal distribution is an extension of ordinary multivariate (vector) normal distribution, based on a parallel extension of vectors to tensors. Since tensor normal distribution express the correlations of each order of the tensor, mixtures of tensor normal distribution provide a tool for investigating tensor data structure. By applying our proposed method to artificial data sets and real data examples, we show that the proposed method is superior besides in classification accuracy, in that the method can express mode-specific covariance matrix.

Keywords: Mixture Models; Tensor-variate distribution; EM-algorithm

Kohei Uno
Osaka University e-mail: uno.stats@gmail.com

Classification of japanese graduate schools: in terms of educational practices and the grown globalization competencies by the policies

Taiyo Utsuhara, Masaki Uto, Asana Ishihara, Atsushi Yoshikawa and Maomi Ueno

Abstract This study undertakes classification of overall Japanese graduate schools in terms of educational practices and the grown globalization competencies by the policies. We divided globalization competencies of graduate schools into the following two: the competencies for 1) global academic activities and 2) global industry. We administered a mail-in survey to acquire data for this study. We analyzed educational practices and the grown globalization competencies in Japanese graduate schools using correspondence analysis, which enables classification of their similar patterns and which reduces high-dimensional Euclidean spaces of multiple variables to low dimensions. Results show that 1) enforcement of educational practices interrupts students' autonomous learning, 2) educational practical academic activities tend to decrease competencies for global industry, and 3) the main policies of universities are broadly divisible into basic academic training and cross-cultural understanding. Subsequently, we clustered the institutions hierarchically in three dimensions. Results also suggest broad divisions of the institutions into groups that (A) tend to enforce educational practices and (B) enhance self-regulated learning. Furthermore, group (A) divides into (A-1), which tends to develop competencies for global industry and which tends to require training for cross-cultural understanding, and (A-2), which enhances competencies for practical academic activities and which requires fundamental training for global academic activities. Group (B) divides into (B-1), which develops competencies for practical academic activities and which requires training for cross-cultural understanding, and (B-2), which tends to develop competencies for global industry and which tends to require fundamental training for global academic activities. Finally, we applied one-way ANOVA and post-hoc analyses. Results suggest that the following set of educational practices enhances the both globalization competencies: 1) self-regulated learning such as programs for logical thinking, 2) educational practices for global industry including financial support for internship abroad programs, and 3) fundamental academic training including academic writing.

Keywords: Global human resources; Correspondence analysis; Hierarchical cluster analysis

Taiyo Utsuhara
Doctoral course, The University of Electro-Communications e-mail: utsuhara@ai.is.uec.ac.jp

Masaki Uto
Assistant Professor, The University of Electro-Communications e-mail: uto@ai.is.uec.ac.jp

Asana Ishihara
Japan Institute of Lifelong Learning e-mail: ishihara@shogai-soken.or.jp

Atsushi Yoshikawa
Visiting Professor, Tokyo Institute of Technology e-mail: at_sushi_bar@dis.titech.ac.jp

Maomi Ueno
Professor, The University of Electro-Communications e-mail: ueno@ai.is.uec.ac.jp

Inverse CA: retrieving a contingency table from a CA solution

Michel van de Velden, Wilco van de Heuvel, Hugo Galy, and Patrick J.F. Groenen

Abstract In correspondence analysis (CA) the aim is to obtain a low-dimensional representation that optimally depicts associations in a two-way contingency table. For inverse CA, this low-dimensional solution is given and the aim is to retrieve the original contingency table, or, if such is not possible, to identify possibly underlying data sets. In previous work on inverse methods for CA, it was shown that solutions can be found by taking convex combinations of a set of so-called vertices. For problems where the original data matrix was of relatively small dimensionality, the complete set of such vertices could be obtained using complete enumeration procedures. The proposed methods, however, did not take into account that a contingency table is integer-valued. In this paper, we take a different approach to the inverse CA problem by considering it from an integer programming perspective. We show that by using integer programming techniques, it becomes possible to retrieve the original contingency tables for several realistic scenarios. Moreover, we investigate whether a contingency table is unique to certain low-dimensional solution.

Keywords: Correspondence Analysis, Inverse Problems, Integer Programming

Michel van de Velden
Erasmus University Rotterdam e-mail: vandevelden@ese.eur.nl

Wilco van de Heuvel
Erasmus University Rotterdam e-mail: wvandenheuvel@ese.eur.nl

Hugo Galy
Erasmus University Rotterdam e-mail: hugo.galy@gmail.com

Patrick J.F. Groenen
Erasmus University Rotterdam e-mail: groenen@ese.eur.nl

Latent class trees

Mattis van den Bergh

Abstract Researchers use latent class analysis to derive meaningful clusters from sets of categorical observed variables. However, especially when the number of classes required to obtain a good fit is large, interpretation of the latent classes in the selected model may not be straightforward. To overcome this problem, we propose an alternative way of performing a latent class analysis, which we refer to as latent class tree modelling. For this purpose, we use a recursive partitioning procedure similar to those used in divisive hierarchical cluster analysis; that is, classes are split until the model selection criterion indicates that the fit does no longer improve. The key advantage of the proposed latent class tree approach compared to the standard latent class analysis approach is that it gives a clear insight into how the latent classes are formed and how solutions with different numbers of classes are linked to one another. We also propose measures to adjust the tree in certain conditions, how to apply the approach to longitudinal data and how to assess relations between cluster membership and external variables.

Keywords: Latent Class Analysis; Classification Trees; Divisive Hierarchical Clustering; Mixture Models

Mattis van den Bergh
Tilburg University e-mail: m.vdnbergh@uvt.nl

Multiclass classification and meta-learning with bayes error estimates

Gerrit J.J. van den Burg and Alfred O. Hero

Abstract Recent advances in information theory have provided a method for accurately assessing the difficulty of a binary classification problem.This is done using a nonparametric estimate of the Bayes error rate based on the minimal spanning tree of the data points. Here, the accuracy of this method is improved further, its use in multiclass classification problems is investigated, and the computational characteristics of this approach are analyzed. Next, the estimate is applied to two practical cases: meta-learning in multiclass classification problems and the design of a hierarchical classification method. The latter methodis compared to existing multiclass classifiers in the context of the support vector machine. Extensive simulations on empirical benchmark datasets show that the proposed classification method is often much faster than existing techniques, while achieving competitive accuracyon most datasets.

Keywords: Classification, Support Vector Machines, Meta-Learning

Gerrit J.J. van den Burg
Econometric Institute, Erasmus University Rotterdam e-mail: burg@ese.eur.nl https://gertjanvandenburg.com

Alfred O. Hero
University of Michigan e-mail: hero@eecs.umich.edu http://web.eecs.umich.edu/ hero/

Decomposing information-theoretic validity indices

Hanneke van der Hoef

Abstract In (semi-)supervised clustering, external validity indices are used to quantify how well a found partition matches a true or 'golden' standard. Over the past years, many external validity indices have been developed, which can be categorized into three approaches: pair-counting, information-theoretic and set-matching measures. While external validity indices quantify how well a partition matches a (golden) standard, most external validity indices are overall measures, which only provide a general overview of the recovery of clusters. Little information is provided on the recovery of individual clusters.By decomposing overall measures into information chuncks corresponding to individual clusters, more insight can be provided into the recovery of individual clusters. While the decomposition of overall indices can be applied to all three approaches of external validity indices, difference exists in the weighting of overall measures between the approaches.In this presentation, the focus is on information-theoretic indices, in particular: two asymmetric versions of the Normalized Mutual Information (NMI) and the ratio of Mutual Information to joint entropy (NMI1). While in information theory normalization is a commonly agreed property to take into account the effect of cluster size, we show that even these normalized indices are still affected by cluster size imbalance. More specifically, we show that information-theoretic indices are weighted means, which preponderate medium-size clusters and underestimate small clusters.

Keywords: validation; information-theory; validity indices; NMI

Hanneke van der Hoef
University of GroningenPsychometrics and Statistics.Research Master student, Research Master student at the University of Groningen (The Netherlands). Member of the Dutch VOC. e-mail: H.van.der.hoef@rug.nl

Clustering patients with non-specific low back pain

Hanneke van der Hoef

Abstract Patients with non-specific low back pain (LBP) form a highly heterogeneous population. Treatment effects for LBP are modest and a 'one-size fits all' approach may be inappropriate. Hence, an important topic in current research is to find subgroups (clusters) within this heterogeneous population of patients with low back pain. The current study aimed at finding these subgroups by using cluster analysis. As a base, thirteen key variables were selected, which showed considerable differentiation between patients in prior studies. Next, an iterative process of determining the optimal number of clusters and the variable impact led to a final clustering with K-medoids, twelve variables and five clusters. Briefly, the groups are characterized as (1) pain in the legs, (2) acute, intense low back pain, (3) sleeping problems and bothered by pain, (4) no activity limitations, (5) long-term episodes of LBP. The clustering is validated on both internal measures and outcome (i.e., longitudinal) data.

Keywords: Clustering;K-Medoids;clara

Hanneke van der Hoef
Faculty of Behavioral and Social Sciences, Research Master student (Psychometrics and Statistics), University of Groningen e-mail: h.van.der.hoef@student.rug.nl

Sparse principal covariates regression for (ultra-)high-dimensional data: some computing issues

Katrijn Van Deun, Elise Crompvoets, and Eva Ceulemans

Abstract Sparse principal covariates regression (SPCovR) is an attractive method for prediction in a context of high-dimensional data, this is data with many more covariates than observations. SPCovR predicts on the basis of a limited set of variables that simultaneously account for the structural variation in the predictors, hence giving insight in the underlying processes. We recently developed an efficient estimation procedure that can deal with a large number of variables without requiring specialized computing infrastructure. In this presentation we will discuss issues related to efficient implementation and to (lack of) convergence of the coordinate descent procedure as presented in the statistical literature.

Keywords: Regularization, Dimension Reduction

Katrijn Van Deun
Tilburg University e-mail: k.vandeun@uvt.nl

Elise Crompvoets
Tilburg University e-mail: E.A.V.Crompvoets@uvt.nl

Eva Ceulemans
KU Leuven e-mail: eva.ceulemans@kuleuven.be

Optimal treatment regime estimation: a clustering problem with a well-defined, pragmatic optimality criterion

Iven Van Mechelen and Aniek Sies

Abstract When multiple treatment alternatives are available for a certain disease, one may wonder which of these alternatives is most effective. As the most effective alternative may differ between patients, one may further wish to look for a rule that specifies the preferable treatment alternative for each patient, based on that patient's pattern of pretreatment characteristics. Such a rule is called a treatment regime. An important challenge for biostatisticians is optimal treatment regime estimation. Given data from some randomized clinical trial or observational study, this estimation comes down to looking for a treatment regime that would yield the best possible expected (potential) outcome if it were used for treatment assignment of all patients in the target population of interest. In this paper we will argue that optimal treatment regime estimation comes down to a clustering problem with a well-defined, pragmatic optimality criterion. Next, we will address the problem of assessing the actual quality of estimated optimal treatment regimes. At this point, we will summarize the results of a benchmarking study with simulated datasets in which the performance of several methods for such an assessment will be investigated. Throughout the talk we will illustrate making use of empirical data from a randomized clinical trial in which two antidepressant treatments were compared.

Keywords: Clustering; Optimal Treatment Regimes; Benchmarking

Iven Van Mechelen
Professor of Quantitative Psychology and Individual Differences, University of Leuven, Belgium e-mail: Iven.VanMechelen@kuleuven.be http://ppw.kuleuven.be/okp/people/Iven_Van_Mechelen/

Aniek Sies
University of Leuven, Belgium e-mail: Aniek.Sies@kuleuven.be http://ppw.kuleuven.be/okp/people/Aniek_Sies/

Introduction to the IFCS Cluster Benchmark Data Repository, the two challenges connected with it, and the target data set for the second challenge

Iven Van Mechelen

Abstract The IFCS Cluster Benchmarking Task Force recently launched a Cluster Benchmark Data Repository (http://ifcs.boku.ac.at/repository/). The aim of this repository is to stimulate better practice in benchmarking (performance comparison of methods) for cluster analysis by providing a unique resource comprised of a wide variety of well-documented, high quality datasets and simulation routines for use in practical benchmarking. The repository will include datasets with, as well as datasets without, given "true" clusterings. A unique feature of the repository is that each dataset will be supplied with comprehensive meta-data, including documentation on the specific nature of the clustering problem and on characteristics that useful clusters should fulfill (with scientific justification). The Task Force subsequently called for individuals to contribute data sets to this repository as part of a first challenge connected with IFCS-2017. The winning data set of this challenge was a data set on baseline and outcome assessment of low back pain patients (which, along with associated meta-data and variable descriptions, can be downloaded at http://ifcs.boku.ac.at/repository/challenge2/). Next, the Task Force called for individuals to contribute cluster analyses of this winning data set as part of a second challenge connected with IFCS-2017. Out of the submissions, six contributions have been shortlisted as potential winners of the challenge. These contributions will be presented during two connected invited sessions at IFCS-2017. The present talk is an introduction into the two sessions in question. In this talk, I will briefly introduce benchmarking in the clustering area along with the IFCS Cluster Benchmark Data Repository. Next, I will briefly introduce the low back pain data set.

Keywords: Clustering; Benchmarking; Challenge

Iven Van Mechelen
Professor of Quantitative Psychology and Individual Differences, University of Lueven e-mail: iven.vanmechelen@kuleuven.be http://ppw.kuleuven.be/okp/people/Iven_Van_Mechelen/</url >

Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis

Vincent Vandewalle

Abstract In model based clustering of quantitative data it is often supposed that only one clustering variable explains the heterogeneity of all the others variables. However, when variables come from different sources, it is often unrealistic to suppose that the heterogeneity of the data can only be explained by one variable. If such an assumption is made, this could lead to a high number of clusters which could be difficult to interpret. A model based multi-objective clustering is proposed, is assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data on some clustering projection. In order to estimate the parameters of the model an EM algorithm is proposed, it mainly relies on a reinterpretation of the standard factorial discriminant analysis in a probabilistic way. The obtained results are projections of the data on some "principal clustering components" allowing some synthetic interpretation of the principal clusters raised by the data. The behavior of the model is illustrated on simulated and real data.

Keywords: Model based clustering, dimension reduction

Vincent Vandewalle
Inria & Université de Lille e-mail: vincent.vandewalle@inria.fr

Cluster distance-based regression

José Fernando Vera and Eva Boj

Abstract In distance-based regression analysis, the vector of continuous responses is projected in a Euclidean space given by multidimensional scaling (MDS). The MDS configuration is obtained by considering a dissimilarity matrix, typically of Euclidean distances, measured between the observed elements in the predictor space. One of the main problems in distance-based regression analysis for mixed variables is that of determining the number of dimensions or of latent predictor variables, in particular when a large dissimilarity data set instead of Euclidean distances is employed. To have Euclidean distances, the observed proximities are translated by means of the well-known procedure of the additive constant. Nevertheless, this methodology usually increases the number of dimensions in MDS and therefore the number of latent predictor variables, in particular when the sampling size is larger with respect to the original number of predictor variables. To reduce the number of elements to be represented using dissimilarities, the use of cluster analysis in conjunction with MDS is an advisable procedure. The main aim in this methodology consists on the classification of the objects into clusters while simultaneously the cluster centres are represented in a low dimensional space. To determine a reduced number of latent predictor variables a cluster distance-based regression procedure is proposed. In this methodology, not the observed sample elements by itself but the coordinates of cluster centres are employed to determine a reduced space of latent predictor variables using a cluster-MDS procedure. Thus, given a dissimilarity matrix obtained from the original data set, a combination of a k-means procedure for dissimilarities and MDS is employed to determine a classification of the observed elements and to determine a reduced latent predictor space. The performance of the proposed procedure is illustrated with the analysis of real data sets.

Keywords: Distance-based regression; dissimilarity; cluster analysis; latent predictor variables.

José Fernando Vera
Department of Statistics and O.R. University of Granada. Prof. Dr., Multivariate Statistics and Classification Group of the Spanish Society of Statistics and Operations Research (SEIO-AMyC). University of Granada. Spain e-mail: jfvera@ugr.es

Eva Boj
Department of Economic, Financial and Actuarial Mathematics. University of Barcelona., Multivariate Statistics and Classification Group of the Spanish Society of Statistics and Operations Research (SEIO-AMyC). University of Granada. Spain e-mail: evaboj@ub.edu

External logistic biplot for nominal and ordinal data

Jose Luis Vicente-Villardon

Abstract Classical Biplot methods allow for the simultaneous representation of individuals and continuous variables in a data matrix. When variables are binary, nominal or ordinal, a classical linear biplot representation is not suitable. Recently, biplots for categorical data based logistic response models have been proposed. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The method is related to logistic regression in the same way as Classical Biplot Analysis (CBA) is related to linear regression, thus we refer to the method as Logistic Biplot (LB). In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. Alternated generalized regressions are used to estimate the row and column markers of the representation. Those estimation methods are suitable for matrices in which the number of individuals is much higher than the number of variables. When the number of variables is high, external logistic biplots could be used; row coordinates are obtained by Principal Coordinates Analysis and then logistic regression are fitted to obtain the variables representation. In this work, external logistic biplots for binary data are extended to nominal and ordinal data using parametric and nonparametric logistic fits. The main results will be applied to several sets of real data.

Keywords: Logistic Biplot; Nominal and Ordinal Data; Latent Traits

Jose Luis Vicente-Villardon
ProfessorDepartamento de Estadística, Department of Statistics. Universidad de Salamanca e-mail: villardon@usal.es http://biplot.usal.es

Clustering by moving centroids using optimization heuristics

Mario A Villalobos-Arias, Juan José Fallas, and Jeffry Chavarria

Abstract In this paper, we present new results of the new approach to clustering of numerical data, in which the centroids are moved, and then the individuals are assigned to the nearest centroid, instead of making movements of individuals between clusters and then calculate the centroids, as has traditionally been done in the algorithms that use similar optimization heuristics.
We present the results of different optimization heuristics, such as:

- simulated annealing,
- particle swarm optimization and
- tabu search.

We obtained very good results, that are compared with the classical clustering algorithms.

Keywords: clustering, optimization heuristics, moving centroids

Mario A Villalobos-Arias
Universidad de Costa Rica e-mail: mario.villalobos@ucr.ac.cr

Juan José Fallas
Instituto Tecnológico de Costa Rica e-mail: jfallas@itcr.ac.cr

Jeffry Chavarria
e-mail: jchavarria@itcr.ac.cr

Classification by quantiles

Cinzia Viroli

Abstract Classification with small samples of high-dimensional data is important in many application areas, but it can be computationally demanding because of the curse of dimensionality. In supervised classification, one way to address this problem is to rely on portions of the conditional distribution of the features given the class labels. For instance, distance-based classifiers use the partial information of the class-conditional distributions, such as the means or the medians, under the implicit assumption that the 'core' of the distribution contains the major source of discriminatory information. But other quantities, such as the tails, may contain information relevant for classification. It may be therefore fruitful to go beyond the central moments. One way is to consider the general quantiles. Quantile classifiers will be defined as distance-based classifiers that require a single parameter, regardless of the dimension, and classify observations according to a sum of weighted component-wise distances of the components of an observation to the within-class quantiles. The optimal parameter for the quantiles can be chosen by minimizing the misclassification error in the training sample. This choice is consistent for the classification rule and the optimal quantile classifier gives low misclassification rates compared to other classification methods. A generalization to unsupervised classification is also proposed, by adapting the quantile-based distances to develop model-based and distance-based clustering methods.

Cinzia Viroli
Department of Statistical Sciences, IT e-mail: cinzia.viroli@unibo.it

Latent markov factor analysis for exploring within-subject measurement model differences in experience sampling studies

Leonie V.D.E. Vogelsmeier, Jeroen K. Vermunt, and Kim De Roover

Abstract Experience sampling, in which participants are questioned repeatedly via smartphone apps, is popular for studying psychological constructs (e.g., wellbeing, depression) within subjects over time. The validity of such studies, e.g., regarding decisions about treatment allocation over time, may be hampered by distortions of the measurement of the relevant constructs over time, e.g., by the onset of response styles or altered interpretations of questionnaire items. Such distortions can be traced as changes in the 'measurement model' (MM) which represents a certain factor structure underlying a participant's answers. In ES studies, it is common practice to assume invariance of the MM over time based on psychometric evaluations of the questionnaires in cross-sectional or survey research or to apply methods that test for invariance of an a priori hypothesized MM only. However, typically we have no prior information on (changes in) the MMs. Therefore, an exploratory approach is required to investigate changes in MM. Also, when they are found, it would be most useful and substantively interesting to learn from MM differences and existing methods provide no insight in these differences. Therefore, we present a method called latent Markov factor analysis (LMFA) which disentangles the distortions from the actual construct measurements in ES studies by modeling MM differences across time points, without the need for prior assumptions on the MMs. In LMFA, a latent Markov chain captures the changes in MMs over time by clustering observations per subject into a few states and, per state, the data are factor-analyzed. Within-subject MM differences are then easily explored by comparing the state-specific MMs and the states indicate for which time points the construct measurements may be validly compared. A simulation study is presented evaluating parameter recovery under a wide range of conditions and the value of LMFA is illustrated with an empirical example.

Keywords: experience sampling; measurement invariance; factor analysis; latent Markov modeling

Leonie V.D.E. Vogelsmeier
Tilburg University, The Netherlands e-mail: l.v.d.e.vogelsmeier@uvt.nl

Jeroen K. Vermunt
Tilburg University, The Netherlands e-mail: j.k.vermunt@uvt.nl

Kim De Roover
Tilburg University, The Netherlands e-mail: k.deroover@uvt.nl

Fractionally-supervised classification using the weighted likelihood

Irene Vrbik

Abstract This talk discusses a model-based approach for classification using a weighted-likelihood framework. In this context, semi-supervised classification involves using labelled and unlabelled observations to fit a finite mixture model which in turn is used as a classifier for the unlabelled observations. In the typical framework labeled and unlabelled data are weighted equally in the model fitting algorithm. As will be discussed in this talk, some improvements in classification performance can be achieved by allowing weights to vary on these two subsets of data. In essence, this approach—termed fractionally-supervised classification—allows labelled data to inform the classifier more heavily than unlabelled data or vice versa. This talk will address the complicated matter of weight specification, and demonstrate the efficacy of this approach on some benchmark data sets.

Keywords: Fractionally-supervised classification, semi-supervised learning, weighted-likelihood

Irene Vrbik
Postdoctoral Fellowship with Natural Sciences and Engineering Research Council of Canada (NSERC) under the supervision of Professor Jason Loeppky, University of British Columbia Okanagan e-mail: irene.vrbik@ubc.ca https://people.ok.ubc.ca/ivrbik/</url >

A supervised multiclass classifier for the family income and expenditure survey

Kazumi Wada and Yukako Toko

Abstract The Family Income and Expenditure Survey (FIES) aims at providing comprehensive data on income and expenditure of households in Japan. About 9,000 sampled households are asked to keep daily accounts of all the transactions, and the resulted statistics are published approximately within a month.
The experts of the survey classification have sorted out a variety of entries containing orthographical variance and local dialects into approximately 600 labels for compiling statistics with limited time and resources. To alleviate the workload, the electronic questionnaire and a rule-based auto-coding system has been introduced.
So far, the coding by experts are still necessary since a large part of the households submit handwritten account books; however the future spread of the electronic questionnaire will facilitate the compilation of statistics together with the auto-coding system. The rule-based auto-coding system has approximately 15,000 rules. Those rules are prepared and managed by the experts.
With the aim to increase further productivity, we are developing a multiclass classifier for the auto-coding system by a simple white-box algorithm using an idea of naïve Bayes. The classifier creates coding rules automatically based on previously coded data. In this study, the details of the system, its performance, and comparison with an existing supervised multiclass classifier will be presented.

Keywords: Classification, Naïve Bayes, Automated coding, Natural language processing

Kazumi Wada
Senior Researcher, Research and Development Division, National Statistics Center e-mail: kwada@nstac.go.jp

Yukako Toko
Researcher, Research and Development Division, National Statistics Center e-mail: ytoko@nstac.go.jp

Critical thinking ability scale development on item response theory

Noboru Wakayama, Yoshimitsu Miyazawa, Shinji Kajitani, and Maomi Ueno

Abstract Critical Thinking, as shown in the 21st century skill (Griffin et al. 2010), is essential to the life of modern society. Critical Thinking is the ability and willingness to think logically and to derive rational decisions without being confined by prejudice (Wakayama 1997). Critical Thinking test requires a long time consideration for test candidates, so it is difficult to display many items in the test. A test should cover the whole area of Critical Thinking and also measure Critical Thinking ability with as few items as possible. Therefore, the purpose of this research is to develop and evaluate the scale of Critical Thinking ability. In order to develop Critical Thinking ability scale, we employed Item Response Theory (IRT), which allowed us to evaluate all items on the only one scale with the same criteria. Various Critical Thinking ability scales have been developed all over the world thus far. This study qualitatively analyzed all previous Critical Thinking scales to identify the measured abilities in each scale. The results derived that each previous Critical Thinking scale addresses parts of the following three ability scales: (1) analytical thinking, (2) logical and reasoning, and (3) reading and understanding. We also analyzed the data of these three Critical Thinking scales tests conducted for 736 college students. Based on the result of these three Critical Thinking scales' correlation, scatter-plot, test information function and scree-plot, these three Critical Thinking scales (analytical, reasoning and reading) were statistically almost independent and confirmed to be reliable. In addition, we conducted this ability scale tests of analytical thinking on three groups: university researchers, college students, and high school students. Consequently this scale confirmed its validity because the performance level on each group indicated significant differences.

Keywords: critical thinking; item response theory (IRT); scale development

Noboru WAKAYAMA
Teikyo University e-mail: cct60660@syd.odn.ne.jp https://www.e-campus.gr.jp/staffinfo/public/staff/detail/125/20Teikyo University

Yoshimitsu MIYAZAWA
Tokyo Gakugei University, Tokyo Gakugei University e-mail: miyazawa@u-gakugei.ac.jp

Shinji KAJITAN
The University of Tokyo, The University of Tokyo e-mail: shinjip@kfx.biglobe.ne.jp

Maomi UENO
The University of Electro-Communications, The University of Electro-Communications e-mail: ueno@ai.is.uec.ac.jp

A data-driven method for deriving shorter DSM symptom criteria sets: The case for alcohol use disorder

Melanie Wall

Abstract With the release of the newest psychiatric disorder classification system (DSM-5), sponsors describe the new guide as a 'living document' hoping to break the lengthy revision process of its predecessors. Under a new continuous improvement model, the APA has established a new DSM web portal (www.dsm5.org) to field proposed changes. One possible improvement could come from shortening lengthy criteria sets. The DSM-5 provides a fixed-length set of 11 criteria associated with Alcohol Use Disorder (AUD) combining Alcohol Dependence and Alcohol Abuse from DSM-IV. The goal of the present study is to develop an empirical method incorporating factor analysis and variable selection to systematically identify subsets of the 11 criteria and associated cut-off rules that will yield diagnosis as similar as possible to using all 11 criteria. We do this by: (1) maximizing the association between the sum scores of all 11 criteria with newly constructed subscales from subsets of criteria, (2) optimizing the similarity of AUD prevalence between the current DSM-5 rule and newly constructed diagnostic short-forms, (3) maximizing sensitivity and specificity of the newly constructed diagnostic short-forms against the current DSM-5 rule, and (4) minimizing any differences in the accuracy of the short-form across chosen covariates (i.e. age, sex, race, psychiatric comorbidity). In our application to DSM-5 AUD, there were more than 11,000 possible diagnostic short-forms that could be created and using our method we were able to narrow the optimal choices down to 16. Results found that "Neglecting Major Roles" and "Activities Given Up" could be dropped from the criteria set with practically no change in terms of who is diagnosed. The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) was used as the data source.

Keywords: factor analysis; psychiatric disorders; diagnosis

Melanie Wall
Professor of BiostatisticsDirector Division of Biostatistics in Psychiatry , Columbia University e-mail: mmwall@columbia.edu

Mixtures of common t-factor analyzers for high-dimensional data with missing information

Wan-Lun Wang

Abstract Mixtures of common t-factor analyzers (MCtFA) have emerged as a sound parsimonious model-based tool for robust modeling of high-dimensional data in the presence of fat-tailed noises and atypical observations. This paper presents a generalization of MCtFA to accommodate missing values as they frequently occur in many scientific researches. Under a missing at random mechanism, a computationally efficient expectation conditional maximization either (ECME) algorithm is developed for parameter estimation. The techniques for visualization of the data, classification of new individuals, and imputation of missing values under an incomplete-data structure of MCtFA are also investigated. Illustrative examples concerning the analysis of real and simulated data sets are presented to describe the usefulness of the proposed methodology and compare the finite sample performance with its normal counterparts.

Keywords: Clustering; Common factor loadings; Dimension reduction; ECME algorithm; Heavy tails; Missing data

Wan-Lun Wang
Professor of Department of Statistics, Graduate Institute of Statistics & Actuarial Science, Feng Chia University, Taichung 40724, Taiwan e-mail: wlunwang@fcu.edu.tw

Active learning with simultaneous variable and subject selection

Zhanfeng Wang

Abstract Modern computing and communication technologies can make data collection procedures very efficient.However, our ability to analyze and extract information from large data setsis hard-pressed to keep up with our capacity for data collection.Especially, when these data are not collected for any particular research purpose or we are interested in extracting information which was not considered inthe original data collection procedure, there is usually a lack of label information which is essential for learning a classification/prediction rule. In addition, there will be no confirmed information about which variables are essential for constructing an efficient classification rule.The concept of active learning allows subjects to be selected sequentially without using label information is a possible outlet for this situation. In this paper, we study an active learning procedure with sequential variable selection scheme. Both theoretical and numerical results are presented.

Keywords: Active learning; Area Under ROC curve; Classification; Shrinkage Estimate; Stopping time

Zhanfeng Wang
Department of Statistics and Finance, University of Science and Technology of China e-mail: zfw@ustc.edu.cn

Understanding external validity indices

Matthijs J. Warrens

Abstract To evaluate the performance of a clustering method researchers typically assess the agreement between a reference standard partition that purports to represent the true cluster structure of the objects, and a trial partition produced by the method that is being evaluated. High agreement between the two partitions indicates good recovery of the true cluster structure. Agreement between the partitions can be assessed with so-called external validity indices.Researchers tend to use and report validity indices that quantify agreement between two partitions for all clusters simultaneously. Commonly used examples are the Rand index and the Hubert-Arabie adjusted Rand index. Both indices are based on counting pairs of objects. Since these overall measures give a general notion of what is going on, their value (usually between 0 and 1) is often hard to interpret (except, perhaps, for values close to 0 or 1). In this presentation we present an algebraic analysis of several families of overall indices (based on counting pairs). The indices are weighted means (variously defined) of indices that reflect agreement on individual clusters. In many cases the weights of these means are quadratic functions of the cluster sizes. It will be shown that measures like the Dice index and the Hubert-Arabie adjusted Rand index tend to reflect how well the larger clusters are recovered. They provide little to no information on the recovery of smaller clusters. Furthermore, the value of the Rand index is determined to a large extent by the number of pairs of objects that are not joined in either of the partitions. It is a moot point whether this actually assesses agreement.

Keywords: External validity indices; cluster validation; Rand index; Dice index; adjusted Rand index

Matthijs J. Warrens
Groningen Institute for Educational Research University of Groningen e-mail: m.j.warrens@rug.nl

Distances, Bayes errors, and classification method performance in high dimensions

Claus Weihs, Tobias Kassner

Abstract Distances between classes play an important role for the size of Bayes error and classification quality. In theory, for linear discrimination and normal features with invertible covariance matrix the Mahalanobis class distance is directly related to the Bayes error. In practice, in the case of less observations n than features p, one can show that the classification error is converging to the worst result 0.5 for increasing p and two classes with limited class distance. However, if class distance is sufficiently increasing with p, then one can even expect perfect classification for distance based classifiers. In this paper, we will study these results in more detail based on an experimental design varying the number of feature dimensions p, the covariance between features, the share of the classes, and the expected error rate in certain subspaces for the classifiers linear discriminant analysis, independence rule, 1NN, naïve Bayes, decision tree, and linear support vector machine. This will further illuminate the power and weaknesses of standard classifiers in high dimensions.

Keywords: Distance; Bayes Error; Classification

Claus Weihs
TU Dortmund University, Germany e-mail: claus.weihs@tu-dortmund.de

Tobias Kassner
TU Dortmund University, Germany e-mail: tobias.kassner@tu-dortmund.de

Quantifying similarity in brain responses based on multi-subject eeg data: a simulation study to compare the inter-subject correlation (isc) measure to novel methods

Tom F Wilderjans and Wouter D Weeda

Abstract Recently, neuroscientists started focusing on how and the extent to which brains of different subjects react similarly to natural stimuli (e.g., movie trailers). Some evidence exists that stimuli that evoke similar brain responses across people are good predictors of human behavior. For example, advertisements that elicit high neural reliability were more liked, or movies scenes with faces showed high brain reliability across participants.
To quantify neural reliability, the Inter-Subject Correlation (ISC) has been proposed. Although the ISC showed promising results for an empirical EEG data set, the performance of this measure has not been evaluated yet for data with systematically manipulated data characteristics (i.e., amount of neural reliability, noise level, etc.). It is, for example, unknown to which extent the true level of neural reliability is masked by the (large amounts of) noise that is usually present in EEG data. Furthermore, given the high-dimensionality of EEG data, it is unclear whether the ISC method efficiently uses that part of the data that contains information on similarity in brain responses. To explore these issues, we propose two methods. First, a wavelet de-noising procedure for EEG data is proposed that should be applied before calculating the neural reliability measure. Second, novel approaches to estimate neural reliability are introduced. In particular, methods based on matrix correlations, namely Tuckers congruence and the (modified) RV-coefficient, are proposed, along with a Generalized Canonical Correlation Analysis (GCCA) method.
The goal of this talk is three-fold. First, to present a formal framework to evaluate the performance of measures for neural reliability. Second, to evaluate the wavelet-based de-noising procedure. Third, to compare the ISC to the novel proposed measures in terms of the extent to which the true level of neural reliability is recovered. To achieve this goal, the results of an extensive simulation study are presented.

Keywords: Canonical Correlation Analysis; matrix correlations; Neural reliability; EEG data; brain similarity

Tom F Wilderjans
Assistant Professor, Methodology & Statistics Research Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University (Leiden, the Netherlands)Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven (Leuven, Belgium) e-mail: t.f.wilderjans@fsw.leidenuniv.nl

Wouter D Weeda
Assistant Professor, Methodology & Statistics Research Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University e-mail: w.d.weeda@fsw.leidenuniv.nl

Incorporating family disease history in risk prediction models with large-scale genetic data substantially dissolves unexplained variability

Sungho Won and Jungsoo Gim

Abstract **Motivation:** Despite the many successes of genome-wide association studies (GWAS), the known susceptibility variants identified by GWAS have modest effect sizes, leading to notable scepticism about the effectiveness of building a risk prediction model from large-scale genetic data. However, in contrast to genetic variants, the family history of diseases has been largely accepted as an important risk factor in clinical diagnosis and risk prediction. Nevertheless, the complicated structures of the family history of diseases have limited their application in clinical practice.
Results: Here, we developed a new method that enables incorporation of the general family history of diseases with a liability threshold model, and propose a new analysis strategy for risk prediction with penalized regression analysis that incorporates both large numbers of genetic variants and clinical risk factors. Application of our model to type 2 diabetes (T2D) patients in the Korean population (1846 cases and 1846 controls) demonstrated that single nucleotide polymorphisms accounted for 32.5% of the variability of risk in T2D cases, and incorporation of family history led to an additional 6.3% improvement in prediction. Our results illustrate that the family medical history is valuable information on the variability of complex diseases and improves prediction performance.
Availability: The R source codes and the analysis script can be downloaded at http://healthstat.snu.ac.kr/software/revealing_MH

Sungho Won
Department of Public Health Science, Seoul National University e-mail: sunghow@gmail.com

Jungsoo Gim
e-mail: iedenkim@gmail.com

Fused sliced average variance estimation

Sungmin Won, Hyoin Ahn, and Jae Keun Yoo

Abstract We propose an approach to combine the kernel matrices constructed by sliced average variance estimation (SAVE) with various numbers of slices. The proposed approach is called fused sliced average variance estimation (FSAVE). By fusing the information by usual SAVE applications with different slice numbers, the sensitivity to slices can be reduced, so the structural dimension estimation can be improved. Numerical studies confirm this, and two real data analysis with $n > n\ p$ and $n < p$ are presented.

Keywords: Fusing; Inverse regression; Sliced average variance estimation; Sufficient dimension reduction

Sungmin Won
Graduate Student, Department of Statistics, Ewha Womans University e-mail: goretti.won@gmail.com

Hyoin Ahn
Graduate Student, Department of Statistics, Ewha Womans University e-mail: anhyoin93@gmail.com

Jae Keun Yoo
Associate Prefessor„ Department of Statistics, Ewha Womans University e-mail: peter.yoo@ewha.ac.kr

The accelerated hyperbolic smoothing sum-of-distances clustering method: New computational results for solving very large instances

Adilson Elias Xavier and Vinicius Layter Xavier

Abstract The work considers the minimum sum-of-distances clustering problem, which has a equivalent formulation to the Fermat-Weber problem. Its mathematical modeling leads to a min-sum-min formulation which is a global optimization problem with a bi-level nature, nondifferentiable and with many minimizers. To overcome these hard difficulties, we use the Hyperbolic Smoothing methodology in connection with a partition of observations into two groups: data in frontier and data in gravitational regions, which drastically simplify the computational tasks. For the purpose of illustrating both the reliability and the efficiency of the method, we perform a set of computational experiments making use of traditional instances described in the literature. Apart from consistently presenting similar or even better results when compared to related approaches, the novel technique, considering the strict Fermat-Weber formulation, was able to deal planar instances, never tackled before, with up to 1243088 observations, more than 1000 times the previous largest one presented in the literature. The same problem can be defined in spaces with any number of components. In this case, the technique was able to solve even larger clustering problems, up to 1842292 patterns with 16 components.

Keywords: clustering; smoothing; nondifferentiable optimization

ADILSON ELIAS XAVIER
Systems Engineering and Computer Sciences Department PESC/COPPE/UFRJ, Federal University of Rio de Janeiro - Brazil e-mail: adilson.xavier@gmail.com

Vinicius Layter XAVIER
State University of Rio de JaneiroRio de Janeiro, BRAZIL e-mail: viniciuslx@gmail.com

Association rule analysis, and comparison of visualization by the quantification technique

Sanetoshi Yamada and Yoshiro Yamamoto

Abstract About the questionnaire data of multiple choice, we have proposed a visualization of the order to extract the features specific to the attribute using the correspondence analysis and association rule analysis (Yamada and Yamamoto, 2016). In the correspondence analysis, high response rate answer items are gathered in the center, and low response rate answer items have been edge on the outside. However, there are also a number of quantification techniques for visualization in addition to correspondence analysis. Therefore, in this study, we did visualizations of Multidimensional scaling and Hayashi's Quantification Theory Type III in place of the corresponding analysis. And we made a comparison with the corresponding analysis by examining the characteristics of their quantification techniques. We converted the answer and the media layer into the binary data. Then, we calculated the cross-tabulation for the corresponding analysis, and the distance matrix for the multidimensional scaling, where the media layers are six attributes of M1 (male from 20 years old to 34 years old), M2 (male from 35 years old to 49 years old), M3 (male over 50 years old), F1 (female from 20 years old to 34 years old), F2 (female from 35 years old to 49 years old) and F3 (female over 50 years old). In Multidimensional scaling, low response rate answer items are gathered in the center, and high response rate answer items have been edge on the outside. In Quantification Theory Type III, interpretive contents changed by an axial combination. In this study, questionnaire data we use is the data of the survey of QPR monitor of Macromill INC. (QPR-SCAPE).

Keywords: visualization; Quantification Theory; Multidimensional scaling

Sanetoshi Yamada
Graduate School of Science and Technology, Tokai University e-mail: S.Yamada@star.tokai-u.jp

Yoshiro Yamamoto
School of Science, Tokai University e-mail: yama@tokai-u.jp

Analyze the health consciousness of yogurt buyers

Saya Yamada and Yumi Asahi

Abstract In Japan, health consciousness is growing. There is a standard "food for specified health uses". People calls "Tokuho". There are many foods and drinks certified to it.The market size of "Tokuho" in 2015 was 639.1 billion yen. The growth rate from the previous year is 109.3%. Dairy products occupy more shares than other types.A company can appeal efficacy if there get certified by "Tokuho". We often see a TV commercial that appeals the effect. Advertising expenditure on terrestrial TV in 2016 was 1,837.4 billion yen. It's 93.4% of the total. It's an important advertising medium. This time, there is focus on Meiji Yogurt R-1 and Meiji Provio Yogurt LG21. These are yogurt that one of the dairy product. We used TV viewing survey, purchasing process data and attribute data of September and October 2013. As of October 2013, about Meiji Yogurt R-1, 8.97% people who have bought one or more times, and 80.96% people who have know it, but have not buy it. About Meiji Provio Yogurt LG21, 12.0% people who have bought one or more times, and 84.71% people who have know it, but have not buy it. People who wanted to buy, in order 40.41%, 45.42%. Despite that people wants to buy it, there are so many people who do not buy it. We think about a process to have these people purchase it. What kinds of TV programs are people that they bought "Tokuho" commodity of yogurt type watching? Compare the people who bought it with those who did not buy it and saw what kind of difference. Customers were segmented and analyzed by multivariate analysis. What type of program TV commercials would be effective if companies advertise it? We were clarify these.

Keywords: Factor analysis, Questionnaire data, Commercial data

Saya Yamada
Tokai University Graduate School e-mail: yamada.s.3849@gmail.com

Yumi Asahi
Tokai University e-mail: asahi@tsc.u-tokai.ac.jp

Explanatory research of fashion behavior by bayesian network analysis

Keiko Yamaguchi and Hiroshi Kumakura

Abstract Fashion behavior, i.e. consumer behavior toward choice, purchase, and consumption of clothing, is diverse and intricate because it relates to consumer's complicated contexts. For example, while consumers follow social norms of dress in some occasions, they differentiate their appearances from others in other occasions. And fashion behavior affects consumer's psychological conditions, while it is also affected by consumer's attributes, demographics, or psychographics. It is important for practitioners to understand the whole picture of fashion behavior and take actions so that they can keep their customer's interests in fashion alive and promote fashion products appropriately. However, as far as we know, there is few comprehensive research on such a diverse and intricate chain reaction in fashion behavior.In light of this gap, our research aims to describe the chain reaction upon various factors of fashion behavior and identify key factors changing consumer's psychology and behavior in fashion contexts. Firstly, authors learned and estimated the structure of Bayesian networks with "R: bnlearn" under theoretical assumption based on fashion psychology, life course, and consumer experience theory. In this network, consumer's factors such as purchase records, consumer's demographics attributes, psychological and social factors surveyed by questionnaire are included as nodes. These purchase and survey data are provided from an anonymous fashion E-commerce site via JASMAC. Consumer's social factors such as getting married, changing job are extracted from open-ended questions by morphological analysis with "Mecab" and latent semantic analysis. Secondly, from significant paths among nodes, authors find out some key nodes that affect subsequent fashion behavior, such as a demographic attribute arousing consumer's interest in fashion. Finally, authors suggest managerial implications in promoting customer's fashion E-commerce site usage by sensitivity analysis. Our research reveals a part of the whole picture and contributes to theoretical and practical knowledge of consumer's diverse and intricate fashion behavior.

Keywords: Bayesian Network; Survey Data Analysis; Causal Inference

Keiko Yamaguchi
Assistant Professor, School of Management, Tokyo University of Science e-mail: keiko.yamaguchi@rs.tus.ac.jp

Hiroshi Kumakura
Professor, Faculty of Commerce, Chuo University e-mail: kumakura@tamacc.chuo-u.ac.jp

Clustering of multivariate categorical data with dimension reduction via nonconvex penalized likelihood maximization

Michio Yamamoto

Abstract In clustering of categorical data, the dimension reduction provides huge benefits for interpretation and visualization of the data, and several techniques to achieve clustering and dimension reduction simultaneously have been developed. We propose a novel simultaneous technique that is based on the ordinary latent class analysis formulation and is supposed to have cluster-specific component scores in a low-dimensional space and weights for categorical variables. This formulation enhances the interpretability of the cluster structure and avoids a computational burden compared with some existing methods. One of features of the proposed model is to have an indeterminacy about orthogonal transformation of component scores like rotational indeterminacy in principal component analysis. Then, to obtain interpretable result, some rotation methods should be implemented after parameter estimations. Instead of this two-step approach, we develop a penalized likelihood procedure that imposes a nonconvex penalty on the weights for categorical variables, which can be viewed as a generalization of the two-step approach and be expected to provide more interpretable solutions. An efficient algorithm using the EM algorithm with gradient projection and coordinate descent is introduced. Properties of the proposed model are shown by theoretical and empirical analyses.

Keywords: Clustering; Categorical data; Mixture model

Michio Yamamoto
Kyoto University e-mail: michyama@kuhp.kyoto-u.ac.jp

Interactive visualization of characteristics of groups

Yoshikazu Yamamoto, Junji Nakano, and Nobuo Shimizu

Abstract We often have huge amount of data expressed by both continuous and categorical variables. In such circumstances, we cannot see each individual data anymore, but divide them into some natural groups and hope to find different characteristics among groups. Symbolic data analysis is a way to analyze such groups of individual data. Symbolic data is typically described by intervals, histograms for continuous variables, or barcharts for categorical variables. We express each group by up to second order descriptive statistics, for example, means, standard deviations and correlation coefficients of continuous variables, and contingency tables for pair of categorical variables. We use an extended parallel coordinate plot for visualizing these statistics of groups at the same time. We propose to use interactive operations such as selection and linked highlighting to explore characteristics of each group intuitively and have developed a software written by the Java language.

Keywords: Symbolic data; An extended parallel coordinate plot; Java language

Yoshikazu YAMAMOTO
Tokushima Bunri University e-mail: yamamoto@is.bunri-u.ac.jp

Junji NAKANO
The Institute of Statistical Mathematics e-mail: nakanoj@ism.ac.jp

Nobuo SHIMIZU
The Institute of Statistical Mathematics e-mail: nobuo@ism.ac.jp

A trial of diagnostic cut-off point selection in three-class classification for health questionnaire

Kazue Yamaoka, Yoshinori Nakata, Mutsuhiro Nakao, Kei Asayama, Mariko Inoue, and Toshiro Tango

Abstract

BACKGROUND: It is a big issue to prevent a disease beforehand to maintain the health of people, and to improve it. For this, to predict "Ahead sick, so cold ME-BYO)" status is important. ME-BYO is an intermediate level exists between diseased and non-diseased status. Examining diagnostic cut-off points $(c1, c2)$ in 3-class classification (to classify the status for Healthy [H]/ ME-BYO [M] / Disease [D]) is valuable.

OBJECTIVE: To examine diagnostic cut-off points selection method in 3-class classification problems using simulation data corresponding to 3 dimensions of health questionnaire (physical, psychological, and social health).

MEHODS: We assume that a test result is a mixed distribution of three groups (H/M/D) with independent normal distributions. The discrimination between the three distributions and obtain optimal pairs of cut-off points (or thresholds) $c1$ and $c2(c1 < c2)$ in the sense that the sum of the correct classification proportions will be maximized. Some simulation studies are conducted to evaluate the performances of proposed interval estimates. Multivariate normal random numbers with a given variance covariance of the 3 dimensions of health were generated by the multiplicative congruential method.

RESULTS &DISCUSSION: In this study, we examined the method of classification into three classes (stages H/M/D) in a specific monotone ordering without relapses (irreversible), such as H $<$ M $<$ D. The parametric approach was considered for estimating the cut-points $c1$ and $c2$. It is necessary to recognize the intermediate stage for the sake to gain the optimal timing window for medical interventions. Through such timely treatments, we can reduce the loss of social wealth and improve the quality of life of the patients. Therefore, the diagnostic tests is valuable, and the evaluating methodologies for such tests are necessary.

Keywords: health questionnaire; diagnostic cut-off points selection; 3-class classification problems

Kazue Yamaoka
Teikyo University Graduate School of Public Health e-mail: kazue@med.teikyo-u.ac.jp

Yoshinori Nakata
Teikyo University Graduate School of Public Health e-mail: ynakata@med.teikyo-u.ac.jp

Mutsuhiro Nakao
Teikyo University Graduate School of Public Health e-mail: mnakao@med.teikyo-u.ac.jp

Kei Asayama
Teikyo University School of Medicine e-mail: kei@asayama.org

Mariko Inoue
Teikyo University Graduate School of Public Health e-mail: inoue-ph@med.teikyo-u.ac.jp

Toshiro Tango
Center for Medical Statistics e-mail: tango@medstat.jp

Layered multivariate regression with its applications

Naoto Yamashita and Kohei Adachi

Abstract Multivariate regression is known as a multivariate extension of multiple regression, which explain/predict the variations in multiple dependent variables by multiple independent variables. Recently, various procedures for Sparse Multivariate Regression (SMR) have been proposed, in which a sparse regression coefficient matrix (having a number of zero elements) is obtained aiming to facilitate its interpretation. The procedures for SMR can be classified into the following two types; penalized and cardinality-constrained procedures. In them, the resulting number of zeros in the regression coefficient matrix is controlled/constrained by a prespecified penalty parameter or cardinality value. In this research, we propose another approach for SMR, referred to as Layered Multivariate Regression (LMR). In LMR, the regression coefficient matrix is assumed to be a sum of several sparse matrices, which is called *layer*. Therefore, the sparseness of the resulting coefficient matrix is controlled by how many layers are used. In LMR, k-th layer can be viewed as the coefficient matrix in the regression of a partial residual (i.e., the residual for all but k-th layer) on independent variables, and thus the variance explained by LMR gets closer to that for the unconstrained regression as the number of layers increases. We present an alternating least squares algorithm for LMR and a procedure for determining how many layers should be used. LMR is assessed in a simulation study and illustrated with a real data example. As an application of LMR, procedures for sparse estimation in some multivariate analysis techniques (e.g., principal component analysis) are also presented.

Keywords: Sparse Analysis; Regression; Multivariate Analysis

Naoto Yamashita
Graduate School of Human Sciences, Osaka University e-mail: nyamashita@hus.osaka-u.ac.jp

Kohei Adachi
Graduate School of Human Sciences, Osaka University e-mail: adachi@hus.osaka-u.ac.jp

The longitudinal and cross-national values survey: Cultural manifold analysis of national character

Ryozo Yoshino

Abstract The Institute of Statistical Mathematics has been conducting a longitudinal survey on Japanese national character since 1953. From 1971, this survey was extended to include cross-national comparative surveys and people of Japanese ancestry in Hawaii, the West Coast of the United States, and Brazil. The cross-national survey primarily focuses on comparing social values, ways of thinking and feeling, and other relevant characteristics of people from various nations. This study investigates conditions under which meaningful cross-national comparability of social survey data is guaranteed, despite differences in languages and statistical sampling methods. In this presentation, I discuss the development of our research paradigm, termed cultural manifold analysis (CULMAN), and provide an overview of our past surveys. Among others, I review past research on people's sense of trust as reflected in the data from longitudinal and cross-national comparative surveys by the Institute of Statistical Mathematics, utilizing Hayashi's Quantification Method III. To overcome the limitations of the studies based mostly on the items of the General Social Survey or the World Values Survey, I explore more basic social values on human bonds that may underlie people's sense of trust beyond differences in countries or time.

Keywords: cross-national comparison; cultural manifold analysis; family; Japanese immigrant; interpersonal relations; Japanese national character; longitudinal survey; social survey; trust

Ryozo Yoshino
The Institute of Statistical Mathematics e-mail: yoshino@ism.ac.jp http://researchmap.jp/read0000926/?lang=english

Morphological characterization of 3d shape with curvature flow-based spin transformation and spherical harmonics decomposition

Ryo Yamada, Fujii Yosuke, Kohei Suzuki, Ayako Iwasaki, Takuya Okada, and Kazushi Mimura

Abstract With the advance of imaging technology such as two-photon microscopy, computed tomography and magnetic resonance, 3D shape of the object is available. Although parameterizing the shape of objects is one of the important research topics in three dimensional data processing, systematic and data-driven evaluation of 3D shape, such as cellular shapes, is difficult. Various methods have been proposed to parameterize their surface, however, some of them depend on the particular morphological patterns or are not applicable to the shapes with significant inclusions and/or protrusions.
Methods
Cell shapes were obtained from 4D time-lapse two-photon microscopy and z-axis images were stacked into voxel. Isosurface was reconstructed by Marching Cubes algorithm. Curvature flow-based conformal spin transformation transformed arbitrary shapes into unit sphere without collapsing the surface pattern. During this procedure all local features of the original shapes were mapped as scalar/vector fields in one-to-one correspondence to unit sphere. After mapping the surface coordinate system on unit sphere into the original shape, feature values (e.g. mean and Gaussian curvature) were decomposed into spectra through spherical harmonics. Distances between paired shapes were estimated by computing L2 norm between two spectra of spherical harmonics coefficients and this distance matrix was embedded into reduced dimensional space through multi-dimensional scaling (MDS).
Results and Conclusion
Simulation dataset and real cell data from two-photon microscopy were parameterized successfully and their sequential shape changing were visualized in MDS space as trajectories. This framework extracted quantitative feature spectra from complex 3D shapes and would integrate them with biological measurements such as transcriptomes.

Keywords: 3D; spherical harmonics; differential geometry;quantification; multi-dimensional scaling

Fujii Yosuke
Statistical Genetics, Kyoto University e-mail: fujii@genome.med.kyoto-u.ac.jp

Kohei Suzuki
Kyoto University e-mail: k.suzuki0604@gmail.com

Ayako Iwasaki
Kyoto University e-mail: iwasaki.ayako.38n@st.kyoto-u.ac.jp

Takuya Okada
Kyoto University e-mail: okokada.takuya@gmail.com

Kazushi Mimura
Faculty of Information Sciences, Hiroshima City University e-mail: mimura@hiroshima-cu.ac.jp

Ryo Yamada
Statistical Genetics, Kyoto University e-mail: yamada.ryo.5u@kyoto-u.ac.jp

Order selection for high-dimensional non-stationary time series

Shu-Hui Yu

Abstract Most model selection methods for high-dimensional time series are devised for the stationary processes. We consider the problem of model selection for high-dimensional autoregressive (AR) models whose integration orders can be positive or zero, and hence both stationary and non-stationary cases are included. Combining the strengths of AIC and BIC, we propose a two-stage information criterion (TSIC), and show that TSIC is asymptotically efficient in predicting integrated AR models, regardless of whether the underlying AR coefficients obey the strong or weak sparsity condition. We also conduct a simulation study to compare the performance of AIC, BIC, HQIC, TSIC, Lasso and the adaptive Lasso. Our study reveals that TSIC performs favorably compared to other methods in various scenarios.

Keywords: Asymptotic efficiency; high-dimensional time series; Lasso; non-stationary processes; TSIC

Shu-Hui Yu
Associate Professor, Institute of Statistics, National University of Kaohsiung e-mail: shuhui@nuk.edu.tw

Classification based on dissimilarities towards prototypes

Beibei Yuan, Willem Heiser, and Mark de Rooij

Abstract We introduce the δ-machine, a statistical learning tool for classification based on dissimilarities or distances between profiles of the objects on the predictor variables. The first step is to compute dissimilarities between the objects and a set of selected exemplars or prototypes. Afterwards, these dissimilarities take the role as predictors in a logistic regression to build classification rules. This procedure leads to nonlinear classification boundaries in the original predictor space.In this presentation we discuss the δ-machine with mixed data. Two dissimilarity measures are distinguished: the Euclidean distance and the Gower measure. The first is a general distance measure, while the second is a tailored dissimilarity measure for mixed type variables. Using simulation studies we compared the performance of the two dissimilarity measures in the δ-machine using three types of artificial data. Furthermore, we investigate the performance under the following conditions: the types of predictor variables (mixed or purely discrete); the number of levels of categorical predictors (binary or multi-category). The simulation studies show that overall the Euclidean distance is competitive with the Gower measure, and in some conditions it outperforms the Gower measure.

Keywords: Dissimilarity; Nonlinear Classification; the Lasso; Monte Carlo simulations; Mixed data

BEIBEI YUAN
Institute of Psychology, Methodology and Statistics Department, LEIDEN UNIVERSITY
e-mail: B.YUAN@FSW.LEIDENUNIV.NL

Willem Heiser
Institute of Psychology, Methodology and Statistics Department, LEIDEN UNIVERSITY
e-mail: HEISER@FSW.leidenuniv.nl

Mark de Rooij
Institute of Psychology, Methodology and Statistics Department, LEIDEN UNIVERSITY
e-mail: ROOIJM@FSW.leidenuniv.nl

A comparative evaluation of feature selection methods

Wanwan Zheng and Mingzhe Jin

Abstract In studies of classification, the first requirement is to identify the most useful features that discriminate the attribution. Feature selection is a strategy that aims at making text classifiers more efficient and accurate. Due to the long history of feature selection studies, many feature selection methods are available in the literature. A feature selection method provides us with a way to reduce the dimensionality of a dataset by removing irrelevant features from the classification task, so as to reduce computation time, improve prediction performance, and supply a better understanding of the data. However, it remains difficult to determine an appropriate feature selection method in real applications, as so many feature selection methods are available. The focus in this paper is the evaluation and comparison of feature selection methods regarding authorship attribution problems. We firstly provide an overview of feature selection methods and then discuss the general versatility that can always be used to select useful features. We use different languages (Japanese, Chinese, and English), and different types of features (Japanese: word-unigram, tag-unigram, tag-bigram; Chinese: the function words Xuci, word-bigram; English: spam dataset) to measure the general versatility of these methods. Finally we use SVM (Support Vector Machine) to measure the effectiveness of feature selection methods. In addition, we use different systems to calculate the accuracy rate of SVM and determine the rank of feature selection methods based on different criteria. Based on the integrated analysis of four tables, the validity and effectivity of the feature selection methods are evaluated. If one feature selection method can be placed in the first half of the top-ranking methods over three times, we can state that the feature selection method is effective and universally valid.

Keywords: classification, feature selection, effective, universally valid

WANWAN ZHENG
Graduate School of Culture and Information Science, Doshisha University e-mail: teiwanwan@gmail.com

MINGZHE JIN
Faculty of Culture and Information Science, Doshisha University e-mail: mining.jin@gmail.com

Cross-national comparison on collective characteristics of cultural symbols in Asia-Pacific area

Yuejun Zheng

Abstract Acculturation, a term synonymous with assimilation, refers to the process of concrete cultural change and psychological change as a result of continuous contact with different cultures. The phenomenon of acculturation can happen at both the group level and the individual level. Many contemporary studies have primarily focused on different strategies of acculturation and how variations in acculturation affect how well in individuals adapt to their society in the fields of psychology, anthropology, and sociology. Unfortunately, there were still few results which concentrate on the process of acculturation from the viewpoint of cross-cultural comparison based on survey data. The purpose of this paper is to identify collective characteristics of symbolic culture and representative values in the Asia-Pacific area, and quantitatively clarify the similarities and dissimilarities in the different societies based on a cross-cultural survey dataset. Data analysis has focused on the modern people's concerned themes such as cultural symbol, belief, tradition, and subjective well-being etc. Typical patterns of acculturation have been identified depending on demographic attributes such as gender, age, education and household income. The results derived from categorical data analysis have revealed the trends of acculturation and significant difference among Asia-Pacific countries today.

Keywords: Cross-national survey; Categorical data analysis; Pettern calssification

Yuejun ZHENG
Faculty of Culture and Information Science, Doshisha University e-mail: yzheng@mail.doshishac.jp