Analysis and Modeling of Complex Data in Behavioural and Social Sciences

Book of abstracts

JCS – CLADAG 12

Joint meeting of the Japanese Classification Society and the Italian Classification and Data Analysis Group

Villa Orlandi – Anacapri, Italy
September 3-4, 2012

Japanese Classification Society

Preface

Following a biennial tradition of organizing joint meeting with classification societies, the CLAssification and Data Analysis Group of the Italian Statistical Society – CLADAG – hosts the Japanese Classification Society – JCS – in an international meeting in Anacapri (Capri Island, Italy).

The past editions have been held in Caserta (2008) with the French Classification Society – SFC – and in Florence (2010) with the German Classification Society – GfKl.

The conference focus is on the Analysis and Modeling of Complex Data in Behavioural and Social Sciences, including theoretical developments, applications and computational methods. The joint meeting aims at enhancing the scientific cooperation between Italian and Japanese data analysts, and at establishing new cooperation between members of the two societies.

This volume contains the Abstracts of all the presentations - keynote and specialized lectures, as well as solicited and contributed session talks - presented at the joint meeting. We hope that the broad range of topics in the wonderful and gorgeous scenery of the Capri Island will stimulate the scientific discussion and exchange among participants for new and interesting challenges.

Naples, July 2012

Donatella Vicari (CLADAG Co-Chair)
Akinori Okada (JCS Co-Chair)
Giancarlo Ragozini (LOC Chair)

Scientific Program Committee

CLADAG Members

Donatella Vicari (co-chair), *Sapienza University of Rome*
Vincenza Capursi, *University of Palermo*
Giuseppe Giordano, *University of Salerno*
Stefania Mignani, *University of Bologna*
Giancarlo Ragozini (LOC chair), *Federico II University of Naples*
Piercesare Secchi, *Politecnico di Milano*

JCS Members

Akinori Okada (co-chair), *Tama University*
Yasumasa Baba, *The Institute of Statistical Mathematics*
Koji Kurihara, *Okayama University*
Kazunori Yamaguchi, *Rikkyo University*
Ryozo Yoshino, *The Institute of Statistical Mathematics*

IASC Member

Vincenzo Esposito Vinzi, *ESSEC Business School*

Local Organizing Committee

Giancarlo Ragozini (chair), *Federico II University of Naples*
Antonella Bisceglia, *Federico II University of Naples*
Rita Cimmino, *Federico II University of Naples*
Daniela D’Ambrosio, *Federico II University of Naples*
Giorgio Infante, *Federico II University of Naples*
Alfonso Piscitelli, *Federico II University of Naples*
Concetta Scolorato, *Federico II University of Naples*
Agnieszka Stawinoga, *Federico II University of Naples*
Cristina Tortora*, Stazione Zoologica Anton Dohrn of Naples*
Domenico Vistocco, *University of Cassino and Southern Lazio*
Maria Prosperina Vitale, *University of Salerno*

TABLE OF CONTENTS

ABSTRACTS

ABSTRACTS

Spatial Clustering of Earthquakes by Data Depth

Claudio Agostinelli, Ca' Foscari University of Venice
Mario Romanazzi, Ca' Foscari University of Venice

Abstract

Earthquake catalogues can be interpreted as samples from the (unknown) parent distribution of seismicity for the reference region and period of time. Data include time coordinate of shocks, spatial coordinates of epicenters, depth below epicenters and magnitudes and are modelled by multidimensional stochastic processes with a space dimension corresponding to location on the Earth's crust and two scalar dimensions corresponding to time and energy release of the shock (Vere-Jones, D., 1995). The goal of such investigations is to measure the seismic risk. Within the general framework, a more specific problem involves the spatial distribution of seismic events, the existence of different clusters and possible space-time interactions. Data depth is a nonparametric method able to rank multivariate data according to the degree of centrality (Liu, R., Parelius, J. and Singh, K., 1999). With seismic data it provides spatial maps of the (estimates of) probability of occurrence of new events. A major application is the investigation of space clustering and even space-time clustering, through comparison of space clusterings of different periods. This result is obtained by using local depth functions (Agostinelli, C. and Romanazzi, M., 2011) which allow the partial centers of a possibly multimodal distribution to be recognized. Local depth has some contact points with kernel smoothing (Stock, C. and Smith, E., 2002). The present work deals with an Italian earthquake catalogue including events recorded after 1600 A. D. with magnitude not less than 4.5 (data kindly provided by Basili, R. and Rotondi, R.). The spatial clustering of earthquakes issued by data depth is compared with the classification based on homogeneous tectonic regions.

Globally Optimal Scoring Weights and their Application in Item Response Theory Models

Sayaka Arai, The National Center for University Entrance Examinations
Shin-ichi Mayekawa, Tokyo Institute of Technology

Abstract

In educational assessment, tests are used to gather information about students' level of understanding or learning achievement, and to make decisions about course placement, university admission, etc. There are several types of test scores. The most popular is the number right score that is defined as the number of items answered correctly by the examinee. Another type is the weighted total score which is defined as the weighted sum of the items answered correctly. The weighted scores are easy to calculate, although the statistical properties of the score depend on the weights. On the other hand, in item response theory (IRT), an examinee score is provided as the estimate of his/her latent trait, ϑ. Since IRT has numerous advantages over classical test theory, IRT methods are used in a number of testing applications. IRT scores are estimated, in most cases, by either the maximum likelihood estimation (MLE) or the expected a posteriori estimation (EAP) method. However, because these estimation methods are complicated, it is almost impossible for a layman to comprehend how the scores are calculated. Mayekawa (2008) developed the globally optimal scoring weights as a new way of weighting items. Globally optimal scoring weights are a set of weights which maximize the expected test information under IRT. The results showed that they reduced the posterior variance when evaluating the posterior distribution of latent ability, ϑ, given the weighted total scores (Mayekawa, 2008; Arai & Mayekawa, 2009, 2011). In this study, we applied the globally optimal scoring weights to real data. The real data we used were the National Admission Test for Law Schools (NATLaS) data. Since NATLaS consists of both dichotomous and polytomous items, we used graded response models. Further, because NATLaS has a set of built-in item weights, we calculated three types of test scores: (1) number right scores (no weight), (2) original built-in weight (original weight), and (3) globally optimal scoring weight (GO weight). We compared the discrimination power of the three, and found that GO weight had the highest power to classify examinees accurately.

Multilevel Mixture IRT Models: an Application to the University Teaching Evaluation

Silvia Bacci, University of Perugia
Michela Gnaldi, University of Perugia

Abstract

In the educational and evaluation contexts, data usually present several characteristics that need to be taken into account to correctly model the heterogeneity of the analyzed phenomenon. Firstly, the variables of interest are not directly observable and we rely on the observed responses to a set of items, which may be differently scored (e.g., binary or ordinal polytomously scored items). Secondly, data usually have a hierarchical structure with lower-level units aggregated in higher-level units (e.g., students in courses). In this situation, we expect a higher correlation between lower-level units belonging to the same higher-level unit than the correlation between lower-level units belonging to different higher-level units. In other words, the global heterogeneity may be decomposed at the different hierarchical levels. The multilevel mixture factor models represent a wide class, which allows to incorporate the two above mentioned elements in the same frame. This class of models presents high flexibility, allowing (i) to specify discrete, continuous, and mixed latent variables, that may differ at the different levels of the hierarchy, and to (ii) define measurement models with different item parameterizations. Moreover, (iii) observed covariates may be considered to explain part of the global heterogeneity. In this contribution, we rely on a specific type of multilevel mixture factor models, characterized by (i) an Item Response Theory (IRT) parametrization and (ii) discrete latent variables at all hierarchical levels. The main aim of the proposed multilevel mixture IRT model is the classification of higher-level units into a small number of homogeneous classes, so that units in the same class will receive the same type of treatment. Other than simplifying the decisional process through the classification of units, the discreteness assumption of latent variables is convenient from a computational point of view. Indeed, it does not require any restrictive assumption about the distribution of latent variables and it allows to skip the well-known problem of the intractability of multidimensional integral which characterizes the marginal log-likelihood function in continuous multilevel models. To select the number of latent variables characterizing each hierarchical level and to allocate each item to the corresponding latent variable, we suggest to adopt a model-based hierarchical clustering procedure, based on a multidimensional mixture IRT model. The procedure will be separately applied for each hierarchical level, so as to allow a different latent structure for the different levels. The proposed model is illustrated through an application to university teaching evaluation. More precisely, we consider data coming from questionnaires on students' satisfaction about courses of the Faculty of Political Sciences of the University of Perugia in the academic year 2009-2010. Our aim is to propose a classification of courses in classes, which are homogeneous with respect to the quality of teaching on the different latent variables measured by the questionnaire. Data have a hierarchical structure, with students aggregated in courses. The resulting model is a three-level mixture IRT model, with item responses as first-level units, students as second-level units, and courses as third-level units.

A Procrustes-based Analysis for Comparing Relational Datasets

Simona Balbi, Federico II University of Naples
Michelangelo Misuraca, University of Calabria
Agnieszka Stawinoga, Federico II University of Naples
Adele Medaglia, University of Calabria

Abstract

A huge and complex set of relationships marks out everyday life. Relational systems are the object of the interest of Social Network Analysis (SNA) which considers relations not as the properties of individuals, but of systems of individuals, in a larger relational system (Scott, 2000). Typically in a defined circle of actors, such as relatives, friends or work colleagues, a unique kind of relation is observed. However, in many cases it is also possible to consider multiple types of connections measured at the same set of actors. A theoretical framework for analysis of multiple relations has been developed in the field of Social Network Analysis (Wasserman and Faust, 1994; Pattison and Wasserman, 1999). The aim of the paper is to analyse, in the frame of a factorial approach, two (or more) different kinds of relations existing among actors. In the framework of SNA, Giordano and Vitale (2007) proposed to use a factorial technique, namely the contiguity data analysis in order to explore the structural properties of a network. In this paper, we propose to apply Multidimensional Data Analysis for studying multi-relational data. In a statistical standpoint a factorial analysis is performed separately on the adjacency matrices representing respectively each of considered relations. Subsequently, the Procrustes rotations (Gower, 1975) are proposed to be used for the joint plot of the multiple networks on a common space. The proposed approach gives the possibility to analyse the positions of actors and to compare the multi-relational structures in a joint factorial representation.

References

1. Giordano G., Vitale M.P. (2007). Factorial Contiguity Maps To Explore Relational Data Patterns, Statistica Applicata , Vol. 19, n. 4.
2. Gower J.C. (1975). Generalised Procrustes Analysis. Psychometrika, vol.(40) : 33-51.
3. Pattison, P.E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. British Journal of Mathematical and Statistical Psychology, 52, 169-193.
4. Scott J. (2000). Social Network Analysis: A Handbook. Sage. London.
5. Wasserman, S., & Faust, K. (1994). Social network analysis: methods and applications. New York: Cambridge University Press.

Time-frequency Filtering for Seismic Waves Clustering

Antonio Balzanella, Second University of Naples
Giada Adelfio, University of Palermo
Marcello Chiodi, University of Palermo
Antonino D'Alessandro, National Institute of Geophysics and Volcanology
Dario Luzio, University of Palermo

Abstract

This paper introduces a new technique for clustering seismic events based on processing, in time-frequency domain, the waveforms recorded by seismographs. The detection of clusters of waveforms is performed by a k-means like algorithm which analyzes, at each iteration, the time-frequency content of the signals in order to optimally remove the non discriminant components which should compromise the grouping of waveforms. This step is followed by the allocation and by the computation of the cluster centroids on the basis of the filtered signals. The effectiveness of the method is shown on a real dataset of seismic waveforms.

Modeling Longitudinal Data with Application to Educational and Psychological Measurement

Francesco Bartolucci, University of Perugia

Abstract

I review a class of models for longitudinal data, showing how it may be applied in a meaningful way for the analysis of data collected by the administration of a series of items finalized to the educational or psychological measurement. In this class of models, the unobserved individual characteristics of interest are represented by a sequence of discrete latent variables, which follows a Markov chain. Inferential problems involved in the application of these models are discussed considering, in particular, maximum likelihood estimation based on the Expectation-Maximization algorithm, model selection, and hypothesis testing. Most of these problems are common to hidden Markov models for time-series data. The approach is illustrated by different applications in education and psychology.

Clustering by Considering the Local Density of Units

Vladimir Batagelj, University of Ljubljana

Abstract

The standard (from the book) rule for determining the clusters in hierarchical clustering: cut the dendrogram at selected height - the corresponding subtrees determine the clusters is not valid when there are big differences in local densities in the space of units. There are two approaches to obtain better clusterings in such cases. The first approach is to replace the height in the dendrogram with a clustering level and apply the rule on it. The second approach is to base the clustering procedures on criterion functions that consider the local density. In the paper we present an elaboration of the first approach and illustrate it with clustering results obtained for some artificial and some real-life data sets. The developed algorithms are implemented in *R*.

The Aggregate Association Index and its Extensions

Eric J. Beh, University of Newcastle
Duy Tran, University of Newcastle
Irene L. Hudson, University of Newcastle
Linda Moore, Statistics New Zealand

Abstract

Recently, an new index was proposed that identifies how likely two dichotomous variables may be associated, given only the aggregate, or marginal, information. Such an index is referred to as the Aggregate Association Index, or the AAI. In this presentation we shall consider some extensions of the AAI that demonstrates its methodological and practical links with other simple, but popular, measures of association for 2×2 tables.

Santa Maria del Fiore Dome Behavior: Statistical Models for Monitoring Stability

Bruno Bertaccini, University of Florence
Antonio Giusti, University of Florence

Abstract

The paper describes the work in progress about the analysis of the behaviour of the web cracks on the Brunelleschi’s Dome of Santa Maria del Fiore. The web cracks in the Dome have always given rise to concern about the stability of the monument. The analyses performed show a steady increase in the size of the main cracks and, at the same time, a relationship with the environmental variables. None of the studies presented in the past involves all the variables simultaneously detected by the monitoring system. Understanding the relationships among all the endogenous and exogenous variables characterizing the phenomenon under study would help to predict the static evolution of the monument and to automatically detect any abnormal behaviour, whose identification would, of course, of paramount importance for the preservation of the monument itself.

A Competing Risks Analysis of Italian University Careers

Matilde Bini, European University of Rome
Margherita Velucchi, European University of Rome

Abstract

During the last decades, the Italian university system has been characterized by low levels of professors' career progression and mobility (both geographical or sectoral). Only recently, Italian professors have faced a process of structural change encompassing the entire system of higher education. We study this process focusing on the career progression as an exit from a status, using a competing risk approach to directly compare alternative exit states in a common framework. We use micro data from Italian Ministry of Research on academic personnel (1998-2008) and we distinguish between two competing exit types: professors remains in the same university/sector/department or he/she switches. A Cox regression, stratified by failure type, is then run with a set of covariates interacted with each risk type. Consider the usual survival analysis where time-to-failure is measured as a function of experimental or observed factors. The term competing risk refers to the chance that the event of interest (promotion) may be affected by a competing event, say, university/department/sector switch. This event changes the probability of the occurrence of the event of interest (promotion). This is not simply right-censoring: while censoring merely obstructs you from observing the event of interest, a competing event affects the way and the time at which the event of interest occurs. When competing risks are present, we focus on the failure function, the cumulative incidence function, instead of focusing on the event of interest survivor function (promotion). Because we are interested in the promotion incidence, we use a Cox model that models two sub-hazard functions depending on the competing risks (switch of university/sector/department), because the promotion cumulative incidence function will likely depend on both. The model covariates may likely affect the two sub-hazards differently, and the cumulative incidence function will turn out to be a nonlinear function of these effects and of the baseline hazards. By this method, we can identify how the covariates impact upon each competing risk. We find that: gender difference strongly affects the promotion probability, associate and full professors have a higher probability to be hired from another university than lecturers (position effect). On the contrary, lecturers have a higher probability of being promoted when she/he switches from a scientific sector to another, while an associate professor switching from one sector to another has a lower probability of being promoted (dropout effect). The role, indeed, has an opposite effect when the competing event is university or sector switch: being an associate professor switching from one university to another increases the probability of being promoted, while being an associate professor switching from a scientific sector to another reduces the probability of being promoted.

Scale Reliability Evaluation for *a-priori* Clustered Data

Giuseppe Boari, Catholic University of the Sacred Heart of Milan
Gabriele Cantaluppi, Catholic University of the Sacred Heart of Milan
Marta Nai Ruscone, Catholic University of the Sacred Heart of Milan

Abstract

According to the classical measurement theory (Bollen, 1989), the reliability of the relationship between a latent variable describing a true measure and its corresponding manifest proxies can be assessed through the Cronbach’s alpha reliability index. The Cronbach’s alpha index can be used for parallel measures and represents a lower bound for the reliability value in presence of congeneric measures, for which the assessment can properly be made only ex post, after having estimated the loading coefficients. Let us assume the existence of an a-priori segmentation, based upon a categorical variable Z. We want to test the reliability of the construct over all the groups defined by Z. This corresponds to the null joint hypothesis that the loadings are equal within each group as well as they do not vary among groups. Otherwise different measurement models want to be defined over groups. Raykov (2002) presents a test for measuring group differences in reliability basing on differences of loading estimates in a structural equation model with latent variables (SEM-LV) framework. Our proposal considers a procedure based only on values of the item responses, avoiding the preliminary SEM-LV estimation. Here we consider a formulation of the Cronbach's alpha coefficient according to the decomposition of pairwise covariances in a clustered framework. The rejection of the null hypothesis may be interpreted, in the case the Cronbach's alpha index is not significantly high for all groups, as the possible presence of congeneric measures at the group level (different loading settings among groups). The group differences may also stem from the presence of some zero item loadings in a group.

Use of ICC for Defining the Optimal Clustering Solution under Normality Assumption

Giuseppe Boari, Catholic University of Sacred Heart of Milan
Marta Nai Ruscone, Catholic University of Sacred Heart of Milan

Abstract

Many kinds of data have a clustered structure, usually identified by means of cluster analysis, that classifies objects with reference to a set of user selected characteristics. The resulting groups should exhibit high internal (within clusters) homogeneity and high external (between clusters) heterogeneity.

However, the typical clustering algorithms do not suggest a definitely optimal solution and alternative criteria are, in general, available. For example, one may consider a sort of merging cost deriving from the internal dissimilarity of the identified groups.

On the other hand, a clustered structure may also be defined with reference to a simple hierarchical model, like the random effect one. As previously underlined, also in this case the observations within a group are differently correlated with respect to the other clusters (among distinct groups they are actually independent). However, a typical clustering problem has a multivariate structure, making use of a multivariate data set, generally assumed to be normally distributed.

In the univariate case, the intraclass correlation coefficient, ICC, is usually suggested, in order to identify the presence of a group structure. A univariate F test is the procedure usually adopted (Sheffè, 1959; Rao, 1973). The extension to the multivariate problem has received lower attention.

Our proposal is essentially based on a sort of multiple comparison procedure, for simultaneously testing the presence of positive ICC, for each of the k observed variables.

The optimal number of clusters is then defined to be the one corresponding to the highest significant multiple test.

Innovation and Quality in Italian Local Communities (IQuEL). An Evaluation of Performance of Territorial Services Center (TSC) by a Nonparametric Combination Ranking Method

Mario Bolzan, University of Padua
Livio Corain, University of Padua
Valeria De Giuli, University of Padua
Luigi Salmaso, University of Padua

Abstract

The work presents some results about a national project IQuEL-2010, aimed to solve some problems associated to the digital divide by Territorial Services Center (TSC). A survey was carried out by sample of local operator in the 3 Italian provinces (Padova, Parma, Pesaro-Urbino). We applied a nonparametric combination (NPC) ranking method on a set of nine dimensions related the public services supplied. The results show important differences among the three Italian provinces, at least for six out of nine TSC abilities or performances and producing a Global satisfaction ranking where Parma is less appreciate than Padova and Pesaro-Urbino.

A Modified Weighted Kappa for the Comparison of Different Measurement Scales

Andrea Bonanomi, Catholic University of the Sacred Heart of Milan

Abstract

In psychometric sciences, the choice of a good response scale is a common problem. Several studies (see Bonanomi, 2004) have shown that different measurement scales can lead to highly dissimilar evaluations of goods/services, in particular in the measurement of observable variables in latent variables models. Every scale has, in its very nature, a tendency or propensity to lead a respondent to mainly positive (or negative) ratings. This paper investigates possible causes of the discordance between two ordinal scales evaluating the same good or service. The question we would like to answer is the following: Is the eventual discordance random, or on the contrary, is it the result of a systematic tendency to assign mainly positive (or negative) evaluations?

In psychometric literature, Cohens Kappa (Cohen, 1960) is one of the most important index to evaluate the strength of agreement (or disagreement) between two nominal variables. Kappa does not take into account the degree of disagreement between different scales and all disagreement is treated equally as total disagreement. Therefore when the categories are ordered, it is preferable to use Weighted Kappa (Agresti, 2002), and to assign different weights to subjects for whom the raters differ by i categories, so that different levels of agreement can contribute to the value of Kappa.

In this paper a Modified Weighted Kappa (MWK) is proposed. In a survey an item was presented to n subjects. Each respondent had to assign the evaluation answering two different measurement ordinal scales. Three matrices are involved, the matrix of observed frequencies, the matrix of expected scores based on chance agreement and the weight matrix. Weight matrix cells on the main diagonal represent agreement and thus contain zero. Off-diagonal cells contain weights indicating the seriousness of that disagreement. For the cell (i,j) the correspondent weight is the absolute difference between i and j, and it measures the distances from the main diagonal (situation of perfect agreement). The proposed MWK measures the ratio among the difference between the weighted relative frequencies in the upper triangle of the matrix and in the lower triangle and the sum of the weighted relative expected frequencies. MWK is normalized in the set [-1;1], where -1 indicates a systematic tendency to assign more positive evaluations using the scale with categories by rows, vice versa +1 a systematic tendency for the scale with categories by columns and 0 the situation of indifference (agreement) between the two scales. A proper procedure to determine the lower and upper triangle in a non-square table is also implemented, so to generalize the index in order to compare two scales with different number of categories. With the aim to verify the tendency of a scale to have a mainly positive or negative rating compared to a different one, a parametric test is set up. The null hypothesis is referred to the perfect concordance between the two considered scales. A study with real data is conducted to compare the proposed index to classical agreement measures.

Asymmetries in Organizational Structures

Giuseppe Bove, Roma Tre University

Abstract

Relationships in organizational structures are frequently asymmetric (e.g., the number of e-mail messages that an employee sends to a colleague is usually different from the number of e-mail messages he received from that colleague). So organizational data are usually represented by asymmetric square matrices that cannot be analyzed by standard symmetric approaches. For this reason methods based on Singular Value Decomposition and Asymmetric Multidimensional Scaling were proposed to analyze these types of matrices (e.g., Freeman 1997, Okada 2008, 2010, 2011). In many situations information concerning hierarchies or aggregations in the organizational structure is available and can be used in the analysis of the data (e.g., professional levels or departments in a company). In this paper some methods taking in consideration this additional information will be considered and applied to Krackhardt's (1987) data on advice-giving and getting in an organization.

References

Freeman, L.C. (1997). Uncovering Organizational Hierarchies. Computational & Mathematical Organization Theory, 5-18.

Krackhardt, D. (1987). Cognitive Social Structures. Social Networks, 109-134.

Okada, A. (2008). Two-dimensional centrality of a social network. In: C. Preisach, L. Burkhardt, & L. Schmidt-Thieme (Eds.), Data Analysis, Machine Learning and Applications, (pp. 381-388). Heidelberg, Germany: Springer.

Okada, A. (2010). Two-dimensional centrality of asymmetric social network. In: N.L. Lauro et al. (Eds.), Data Analysis and Classification, (pp. 93-100). Heidelberg, Germany: Springer.

Okada, A. (2011). Centrality of Asymmetric Social Network: Singular Value Decomposition, Conjoint Measurement, and Asymmetric Multidimensional Scaling. In: Ingrassia et al. (Eds.). New Perspectives in Statistical Modeling and Data Analysis (pp.219-227). Heidelberg, Germany: Springer.

THEME-SEER: THEmatic Model Estimation through Structural Equation Exploratory Regression

Xavier Bry, Université Montpellier 2
Patrick Redont, Université Montpellier 2
Thomas Verron, SEITA-ITG Centre de recherche SCR
Pierre Cazes, Université Paris IX Dauphine

Abstract

We present a new path-modelling technique, based on an extended multiple covariance criterion: System Extended Multiple Covariance (SEMC). SEMC is suitable to measure the quality of any Structural Equations System. We show why SEMC may be preferred to criteria based on usual covariance of components, and also to criteria based on Residual Sums of Squares (RSS). Maximizing SEMC is not straightforward. We give a pursuit algorithm ensuring that the SEMC increases and converges. When one wishes to extract more than one component per variable group, a problem arises of component hierarchy. To solve it, we define nesting principles of component models that make the role of each component statistically clear. We then embed the pursuit algorithm in a more general algorithm that extracts sequences of locally nested models. We finally provide a specific component backward selection strategy. The technique is applied to cigarette data, to model the generation of chemical compounds in smoke through tobacco combustion.

A Generalized Additive Model for Binary Rare Events Data: An Application to Credit Defaults

Raffaella Calabrese, University College Dublin
Silvia Angela Osmetti, Catholic University of Sacred Heart of Milan

Abstract

We aim at proposing a Generalized Additive Model (GAM) for binary rare events, i.e. binary dependent variable with a very small number of ones. GAM is an extension of the family of Generalize Linear Models (GLMs) by replacing the linear predictor with an additive one defined as the sum of arbitrary smooth functions. In the GLMs the relationship between the independent variable and the predictor is constrained to be linear. Instead the GAMs do not involve strong assumptions about this relationship, which is merely constrained to be smooth. We extend the Generalized Extreme Value (GEV) regression model proposed by Calabrese and Osmetti (2011) for binary rare events data. In particular, we suggest the Generalized Extreme Value Additive (GEVA) model by considering the quantile function of the generalized extreme value distribution as a link function in a GAM. In order to estimate the smooth functions, the local scoring algorithm (Hastie and Tibshirani, 1986) is applied.

In credit risk analysis a pivotal topic is the default probability estimation. Since defaults are rare events, we apply the GEVA regression to empirical data on Italian Small and Medium Enterprises (SMEs) to model their default probabilities. We compare on these data the performance of the GEVA model with the one of the most used regression model for binary dependent variable, the logistic additive model. By reducing the sample frequencies of rare events (defaults), the predictive performance of the logistic additive regression model to identify the rare events becomes worse. On the contrary, the GEVA model overcomes the underestimation problem and its accuracy to identify the rare events improves by reducing the sample percentage of rare events. Finally, we show that the GEVA model is a robust model, unlike the logistic additive regression model.

The Meaning of *Forma* in Thomas Aquinas. Hierarchical Clustering from the *Index Thomisticus* Treebank

Gabriele Cantaluppi, Catholic University of the Sacred Heart of Milan
Marco Carlo Passarotti, Catholic University of the Sacred Heart of Milan

Abstract

Started in 1949 by father Roberto Busa (1913-2011), the Index Thomisticus represents the first digital corpus of Latin, and contains the opera omnia of Thomas Aquinas (118 texts) as well as 61 texts by other authors related to Thomas, for a total of around 11 million tokens. The corpus is morphologically tagged and lemmatized. The Index Thomisticus Treebank (IT-TB: http://itreebank.marginalia.it), an ongoing project aiming at performing the syntactic annotation of the whole IT corpus, is a dependency-based treebank consisting of around 150,000 annotated tokens for a total of approximately 8,000 sentences from three works of Thomas Aquinas: Scriptum super Sententiis Magistri Petri Lombardi, Summa contra Gentiles and Summa Theologiae. We analyze here the lemma forma. This lemma has 18,357 occurrences in the IT corpus. Presently, 5,191 occurrences of forma have been annotated in the IT-TB. We produced two data sets, one from IT-TB and one from IT: - For each occurrence of forma in IT-TB the observations report the lemmas of: (a) its parent and grandparent in the dependency tree, (b) all its attributives, (c) all its coordinated nodes in the tree, (d) up to 2 words preceding and 2 words following the occurrence of forma concerned in the observation. - For each occurrence of forma in IT the lemmas of up to 3 words preceding and 3 words following the occurrence are reported. We carried out a divisive hierarchical clustering analysis (Kaufman & Rousseeuw, 1990) starting from a dissimilarity matrix generated by considering a modification of the simple matching distance, taking into account the similarity in groups of variables. We performed several analyses, by choosing different settings for grouping the variables. We also excluded specific kinds of words (like function words, pronouns and some verbs) when computing dissimilarities. The same experiments were always performed on both all the 18,357 observations provided by the IT-based matrix and on a subset consisting of the 16,525 observations of forma in Thomas' works only. In order to evaluate the results, we considered: - Gold standard A (GsA): we manually annotated the meaning of 672 randomly chosen occurrences of forma. We used a tagset of 10 different values that were defined according to Deferrari & Barry, Latin Wordnet and lexico-syntactic information from IT-TB; - Gold standard B (GsB): among the observations of GsA, we selected a subset of 356 featuring a clear meaning of forma. The best performing setting on 5,191 observations is the following: - function words, pronouns and verb sum excluded from computing similarity/dissimilarity; - grouping setting: 2 separate groups, namely syntactic information and textual information. With this setting, we reached the best f-score of 0.93 (precision 0.95; recall 0.90), for the GsB observations tagged with label 6 (forma materialis; forma connected with materia). The GsB observations tagged with other labels show lower scores (ranging from 0.86 to 0.5). If GsA observations are concerned, the f-score ranges from 0.8 (precision: 0.93; recall: 0.7) for label 6, to 0.28 for label 2 (forma as anima corporis).

Multivariate Analysis for the Abuse of New Technologies by Young People

Claudia Cappello, University of Salento
Giuseppina Giungato, University of Salento
Daniela Pellegrino, University of Salento

Abstract

In recent years, the rapid increase in using new technologies, such as Internet and mobile phone, by all segments of the population, has determined many benefits to their users. In particular, these tools favour new ways of communication and social relations. On the other hand, an abuse of these technologies could determine adverse health effects (i.e. social isolation and other forms of psychological disorders), especially on young people, which is considered to be high risk for pathological technologies use.

Debates on the influence of behavioural disorders are common and several works have been produced in the last decades. In this paper, a case study on the use of Internet and mobile phone among teenagers, is discussed. In particular, a questionnaire has been submitted to a sample of students (398), aged between 11 and 13 years, in some schools of Lecce and Brindisi districts (Apulia Region). Multivariate analysis has been applied in order to assess the relationships among behaviours and attitudes of teenagers with reference to the using of Internet and mobile phone.

Bayesian Classification Models for Financial Evaluation

Paola Cerchiello, University of Pavia
Paolo Giudici, University of Pavia

Abstract

In this contribution we propose to estimate the probability of financial default of companies and the correlated rating classes, using efficiently the information contained in different databases. In this respect, we propose a novel approach, based on the recursive usage of Bayes theorem, that can be very helpful in integrating default estimates obtained from different sets of covariates. Our approach is ordinal: on one hand, the default response variable is binary; on the other hand, covariates that induce partitioning of companies are measured on an ordinal scale. We use our approach not only in a Bayesian variable averaging perspective but also to binarize ordinal variables in the most predictive way. The method is based on a mixture of Binomial and Beta random variables since we model the proportions of default companies in each level of the covariate as independent Binomials with a Beta prior distribution. The application of our proposal to an Italian credit risk database shows that it performs quite efficiently, allowing to predict for each company the probability of default by averaging the covariates contribute.

Interrupted Careers: the Determinants of University Outcomes in a Competing Risks Approach

Renata Clerici, University of Padua
Anna Giraldo, University of Padua
Silvia Meggiolaro, University of Padua

Abstract

Obtaining an university degree results in important outcomes for subsequent life course. However, choosing to start the path of university does not guarantee that the student will actually graduate: university withdrawal is one of the major problems. In fact, highly complex educational histories are observed in the learning process. In this paper we study the different academic outcomes (withdrawal, course changes, delay, and degree completion) in three-year degree courses at the University of Padova. Data come from the university administrative archives, integrated by some information on degree courses organizational and structural characteristics. We consider about 24,000 students enrolled from 2002/03 to 2004/05 academic years in 84 undergraduates courses. In a previous study on the same data, we found that withdrawal, course changes, and degree completion show great variability across the various degree courses. In this paper we analyze the factors that influence these different outcomes of university path using some methodological attentions. First of all, they are dependent competing risks that lead to not longer being enrolled at university. If the event of interest is degree, the various kinds of drop-out cannot be considered as censored observations. Thus, a competing risks model is required. In addition, in this setting, time is organized into academic year, and a discrete time modeling approach provides appropriate estimates. Finally, the characteristics of different degree courses have to be considered. Thus, we apply a discrete-time method for competing risks event history analysis in a multilevel framework, which takes into account not only student characteristics, but also distinctive features of the degree courses.

Archetipal Functions

Paola Costantini, National Agency for the Evaluation of Universities and Research Institutes
Giovanni C. Porzio, University of Cassino and Southern Lazio
Giancarlo Ragozini, Federico II University of Naples
Juan Romo Urroz, Carlos III University of Madrid

Abstract

Archetypal analysis is a mathematical procedure for decomposing a multivariate datasets as a function of a set of underlying archetypes or ideal types. The aim of archetypal analysis is to find ideal types, the archetypes, within a set defined in a specific context. Here, the problem is to find a few, not necessarily observed, points (archetypes) in a set of multivariate observations such that all the data can be well represented as convex combinations of the archetypes. This methodology was first introduced by Cutler and Breiman (1994) and while used for many years in the physical sciences, these methods have been applied in many fields. In this work we extend this methodology to the case of functional data. Functional data analysis is a collection of techniques in statistics for the analysis of curves or functions. In their work, Cutler and Breiman (1994) have already introduced archetypal functions, using the Euclidean distance as metric to estimate the distance between curves. Our idea is to use basis functions to model the shape of each smooth curve observed over time.

Forward Search for Data Analysis Ported to Octave Environment

Christophe Damerval, European Commission

Abstract

The Forward Search is an approach allowing to perform an efficient and robust statistical analysis of complex and heterogeneous datasets. It allows to detect outliers in different contexts (univariate, multivariate, time series). Besides it provides tools for data transformation and model selection. Its main characteristics are the adaptability to the data, the sorting of observations from the most regular to the strongest outlier, and a precise monitoring of statistics of interest. To ensure that such tools can be used by the statistical community, a considerable work was done in terms of software implementation. In this regard, a toolbox for Matlab environment was released: Forward Search for Data Analysis (FSDA), which makes up the state-of-the-art. In addition to the Forward Search related functions, this toolbox integrates main traditional robust estimators, in-depth exploratory data analysis and dynamic visualization tools. Besides the adoption of algorithmic strategies improved its computational efficiency. Our contribution consists in porting this Matlab toolbox (Matlab being under proprietary licence) to a free software and development environment: Octave, which is distributed under the terms of the GNU General Public License. Such a porting provides an extended availability to a greater number of users in the communities of statistics and computer science. Thanks to the similarity between Matlab and Octave, a certain degree of automatic portability is feasible using ad hoc scripts. A fully-working package for Octave can be further obtained using specific and manual interventions. We highlight the differences in terms of available features and visualization tools. We also compare their respective performances, taking into account different platforms.

Weighted Rank Correlation Measures in Hierarchical Cluster Analysis

Livia Dancelli, University of Brescia
Marica Manisera, University of Brescia
Marika Vezzoli, University of Brescia

Abstract

Cluster analysis aims at identifying groups of individuals or objects that are similar to each other but are different from individuals in other groups. This is useful, for example, in the context of marketing to segment consumers in order to better understand market relations and improve industrial competitiveness. In those studies, consumers' preferences are often expressed using grades, leading to rankings of objects provided by individuals. Standard clustering techniques are not appropriate, because they measure intersubject similarity by distance-type measures, including correlation coefficients, thus requiring quantitative data. Instead, when consumers express their preferences by means of rankings, ordinal data are obtained: ranking numbers convey the location/preference of the "object" in a given ordered list and distance-type measures must be replaced by matching-type measures. Among these, rank-based correlation coefficients, as the Spearman's rho, can be considered. In particular, we think that Weighted Rank Correlation (WRC) measures are remarkably useful, evaluating the agreement between two rankings by emphasizing the contribution of top ranks. For example, consider 3 consumers A, B and C that ranked 6 aspects of a restaurant attributing ''1'' to the most important aspect and ''6'' to the least important one. If the three rankings are: A: [1, 2, 3, 4, 5, 6], B: [1, 2, 3, 4, 6, 5], C: [2, 1, 3, 4, 5, 6] and we want to emphasize top ranks, A and B preferences are more similar than A and C. This is because, with respect to A, B exchanges the lowest ranks while C exchanges the top ones. Unweighted Spearman's rho provides the same result when computed between A and B and between A and C while WRC measures capture the greater similarity between A and B. We focus on WRC measures related to the Spearman's rho adopting weights that are function of both ranks and treat them symmetrically. The WRC indices existing in the literature introduce weights in the simplified formula of rho. Also, we have recently proposed a new class of WRC measures, introducing weights in the unsimplified formula of rho. Generally, the two classes of indices lead to different results when computed on the same pair of rankings. We have compared the performance of 5+5 indices of the two classes and identified the WRC measure that seems to perform better. We now propose to employ this measure to evaluate the similarity between rankings in a hierarchical cluster analysis, in order to segment consumers when they express their preferences by rankings. The procedure has been applied to real data involving more than 1,000 consumers of several McDonald's restaurants, who ranked different aspects of the restaurant service (food quality, neatness, quality-price ratio, etc.). Two hierarchical cluster analyses were performed on this dataset by measuring similarity between consumers using (a) unweighted Spearman's rho and (b) the chosen WRC measure. Results show that groups obtained with the WRC measure really contain subjects whose preferences are more similar on the top ranks.

The Classification Problem in Multi-Criteria Decision Analysis: the Electre Tri method

Renato De Leone, University of Camerino
Valentina Minnetti, Sapienza University of Rome

Abstract

In this work we will address the classification problem in which categories are in a strict preference relation using a multi-criteria approach. To solve it we use the well-known classical Electre Tri method which requires the elicitation of preferential and technical parameters (weights, thresholds, profiles, cutting level) in order to construct a preference model which best matches the Decision Makers preference information. We propose a new methodology in two phases, taking into account that the core of the analysis is the profiles estimation made by linear programming problem.

Churn Prediction in Telecommunication by Boosting Regression Models

Ermelinda Della Valle, Second University of Naples
Rosaria Lombardo, Second University of Naples
Barbara Masiello, Second University of Naples
Sanja Kon, Vodafone Enterprise
Alessandro Bernorio, Vodafone Enterprise

Abstract

In data mining literature great attention has been given to the Customer retention study which is generally perceived as a cornerstone of successful Customer Relationship Management (CRM) (Payne and Frow, 2005). Among various tools of customer retention, our focus is on the customer churn analysis (Neslin, et al., 2006). As the cost of maintaining an existing customer is usually lower than the cost of acquiring a new customer (Reinartz and Kumar, 2003), preventing customer churn is critical to the survival of a company. Therefore, the high acquisition cost makes it imperative to predict customer churn behavior and execute appropriate proactive actions. In particular in this paper, we aim to predict customer churn in mobile telecommunication company. It is undoubted that the quality of a customer churn prediction model is directly influenced by the available input data (demographic, historical transactional, financial data; Glady, Baesens, and Croux, 2009) and by the data mining techniques used to identify customers that are likely to churn. Neslin et al. (2006) point out that the choice of the modeling technique gets a significant impact upon the return on company investment. To model churn marketing in the current work, we propose to look at different boosting models (Hastie, Tibshirani, Friedman, 2001). Boosting is one of the most powerful learning ideas introduced in the last ten years, originally designed for both classification and regression problems. Boosting iteratively fits a model to a data set with weights that depend on the accuracy of the earlier iterated models and uses a linear combination of these models for producing the final prediction. In the family of boosting algorithms (Lombardo, Durand, Leone, 2012), we will compare linear partial least squares and non-linear PLS algorithms (Lombardo, Durand, De Veaux, 2009), addressing towards the best model capable to identify key determinants of customers churn behaviors.

Dynamic Clustering of Financial Assets

Giovanni De Luca, Parthenope University of Naples
Paola Zuccolotto, University of Brescia

Abstract

Several approaches to time series clustering are present in the literature. After the first studies, where dissimilarities between time series were merely derived by the comparison between observations or some simple statistics computed on the data (see for example Bohte et al., 1980), more complex solutions have been proposed. Piccolo (1990) and Corduas and Piccolo (2008) proposed a distance measure for time series generated by ARIMA processes, based on the comparison between the parameters of the corresponding Wold decomposition. Otranto (2008) extended this approach to GARCH models, Galeano and Peña (2000) considered the generalized distance between the autocorrelation functions of two time series while Caiado et al. (2006) introduced a metric based on the normalized periodogram. Alternative methods, employing parametric and non-parametric density forecasts, was discussed in Alonso et al. (2006) and in Vilar et al. (2010), respectively. To give an idea of the great variety of approaches in this framework, it's finally worth recalling the frequency domain approach of Kakizawa et al. (1998) and Taniguchi and Kakizawa (2000), the use of two-dimensional singular value decomposition by Weng and Shen (2008), the procedure using a robust evolutionary algorithm of Pattarin et al. (2004). But the list of citations could be even longer.

In this contribution we apply time series clustering to financial returns, using a procedure clustering assets in groups which are homogeneous in the sense that their joint bivariate distributions exhibit high association in the lower tail. The empirical motivation of this approach is to identify assets whose extreme losses tend to occur simultaneously, due to the rapid contagion characterizing financial markets. This can be important from a risk management perspective, as it allows to protect investments from parallel extreme losses during crisis periods. To this aim, we exploit the clustering procedure proposed by De Luca and Zuccolotto (2011), using a dissimilarity measure based on tail dependence coefficients, estimated by means of copula functions, extending its formulation to a dynamical context. More specifically, through the definition of time-varying tail dependence coefficients, we obtain groups whose composition is dynamically adapted to the variations due to the changes in the volatility of the market. So, the dynamical clustering solutions allow the construction of time-varying portofolios protecting investments from the effects of a financial crisis. In addition, the analysis of the dynamic pattern of tail dependence coefficients can give information about contagion, which is known as a situation when assets are characterized by a significant increase in cross-correlation during the crisis periods.

Worthiness based Scaling of Social Agents. A pseudo-Bayesian Approach for Standardizing

Giulio D'Epifanio, University of Perugia

Abstract

The social behavior of an agent is intended here as that emerging in driving the behaviors of the individuals that the agent should govern, toward achieving increasingly challenging goals that were scheduled, by the policy-maker (PM), on a chain. Using an ordinal indicator Y of outcome, constructed on the chain, a type of index is proposed which takes into input the distributions, realized by the agents on the levels of Y, for evaluating the agents performance. It inherits formal properties of coherence, in rational choices, from the rank dependent expected utility theory. But, in this paper, the value increases, between any two adjacent levels of the ordinal outcome are re-interpreted as increases of worthiness in a probabilistic setting based on the informal criterion of the intrinsic worthiness. The PM may want justify difference of performance, among the agents, on the basis of difference in reference conditions of their governed individuals. In order to quantify these conditional increases, in such a way that they are fully standardized on the behavioral data of a given reference agent, a pseudo-Bayesian approach is outlined. Firstly, it models, up to hyper-parameters on a structural probabilistic framework, interrelated evolutionary-processes, latent behind the goals chain, which are guided by manifest conditions through latent increases of worthiness. Then, it regulates the hyper-parameters of the model by adapting it on the performance data of the reference social agent, using a pseudo-Bayesian method which minimize residuals from updating. Thus, the conditional increases of worthiness, standardized on the reference agents, are elicited which enters conditional evaluation index.

Markov Models for Two Group Transition Data

Mark de Rooij, Leiden University

Abstract

Having longitudinal data researchers are often interested in the differential development of two or more groups. Such development can be assessed based on trajectories or based on transitions. For longitudinal multinomial data we describe a methodology for the statistical analysis based on a distance model. Various hypothesis are translated into different statistical models. Model selection is performed with the QIC statistic, a variant of the AIC for dependent data, and a clustered bootstrap procedure.

On the Use of Multiple Correspondence Analysis to Visually Explore Affiliation Networks

Domenico De Stefano, University of Trieste
Maria Rosaria D'Esposito, University of Salerno
Giancarlo Ragozini, Federico II University of Naples

Abstract

Correspondence analysis (CA) is a data analytic technique designed for contingency tables and which is applicable in a variety of ways to relational data. Indeed, it has been frequently used in social network analysis to analyze and graphically represent two-mode networks. Its use is principally due to the similarity between the affiliation matrix associated to a two-mode network and the usual contingency table. However, the use of CA in social network analysis can been criticized because the nature of relational data do not fit the feature of the method. In this paper we propose a more suitable approach, based on the use of multiple correspondence analysis (MCA), that allows to take into account the nature of the relational data, the intrinsic asymmetry of actors/events in the affiliation matrices, and the attributes of nodes.

Diagnostics for Meta-analysis Based on Generalized Linear Mixed Models

Marco Enea, University of Palermo
Antonella Plaia, University of Palermo

Abstract

Meta-analysis is the method to combine data coming from multiple studies. The aim is to provide an overall event-risk measure of interest summarizing information coming from the studies, such as proportions, difference of proportions, odds or odds ratios. Meta-analysis of Individual Patient Data (IPD) is the gold-standard in evidence-based synthesis. Meta-analysis of IPD assumes that all the information, i.e. data, are available at both study and patient level. To take into account heterogeneity, generalized linear mixed models (GLMMs) are often employed in meta-analysis of IPD, usually assuming normal or binomial outcome distribution, with normal random effects. Even if IPD meta-analysis is the gold standard, very often IPD are available only for some of the studies, while for the other only aggregated data (AD) can be used. In order to take into account the between-study heterogeneity with AD, meta-regression is used. Meta-regression is a weighted regression, whose weights are given by the precision of the studies, in which the outcomes are the effect sizes, i.e. the measures of interest averaged on the whole study, usually assumed to be normal, with covariates values (also said moderators) provided at aggregate level. Although its simplicity, this method do not consider the analysis of raw data, is prone to bias and lacks of statistical power. Thus, when it is possible, one should opportunely combine IPD with AD in such a way to exploit all the available information. Outliers detecting and influence diagnostics analyses are a natural final step in making meta-analysis. However, due to the lack of standard methods for detecting influential studies and/or individuals for non-normal GLMMs, meta-analysts often neglect the diagnostics analysis, reducing it only to a checking of publication bias. In this work, influence diagnostics for GLMMs-based meta-analyses is considered for both IPD/AD and their combinations. Surprisingly, this latter option is little considered, and influence diagnostics for meta-analyses combining IPD and AD has not been addressed yet. Here, we address the problem by applying a known likelihood-oriented influence measure at both individual and study level. A simulated example with artificially created outliers will be used to assess the method performance.

An Extension and a New Interpretation of the Rank-based Concordance Index

Pier Alda Ferrari, University of Milan
Emanuela Raffinetti, University of Milan

Abstract

The ordinal data management is assuming a relevant role in many application areas, in order to capture information about phenomena which are not directly observable. This interesting topic arises in important fields such as health, education and customer satisfaction assessment. More precisely, the main occurring problem regards the study of dependence relations in a quanti-qualitative context, when for example the response variable is quantitative and the covariates mostly assume ordinal nature. In fact, ordinal variables can not be specified according to a metric scale and for this reason the existing standard dependence measures, such as the R^2 determination coefficient, could not be successfully applied because based on the euclidean distance and then more appropriate in a quantitative setting. Therefore, the problem of defining new dependence measures in a multivariate context particularly useful for model goodness of fit and model selection appears. Recent literature presents some interesting suggestions in this direction by resorting to specific statistical tools such as the concordance curve and the Lorenz curves. In fact, the definition of a novel dependence and model selection criterion ("Multivariate Ranks-based Concordance Index"- MRbCI) based on the Lorenz curves and ranks has been developed. However, the classical Lorenz curves employment implies that the underlying variable is characterized by non-negative values: this condition clearly detects a strong restriction with regard to the real application context. In this contribution, the idea is focused on building the corresponding response variable Lorenz curve, through an y-axis values translation of the set of points representing the Lorenz curve. Also a novel Gini measure definition with a translational component has been provided. In this scenario, by denoting with Y(t) the translated response variable, the MRbCI role consists in capturing information about the concordant or discordant relation existing between Y(t) and the corresponding linear estimates $\hat{y}(t)$. Our proposed index performance has been validated by a simulation study and its applicability has been evaluated through the analysis of real data. Note that the extended MRbCI can be further on overdrawn to the context where the response variable assumes only negative values, by implementing the same translational procedure. This last step completes all the possible instances of analysis.

Modelling Facial Expressions through Shape Polynomial Regression

Lara Fontanella, University G. d'Annunzio
Caterina Fusilli, University G. d'Annunzio
Luigi Ippoliti, University G. d'Annunzio
Alfred Kume, University of Kent

Abstract

The reasons for the interest in Automatic facial expression analysis - FEA - are multiple, but they are mainly due to the advancements accomplished in related research areas such as face detection, face tracking and face recognition. Given the significant role of the face in our emotional and social lives, it is not surprising that the potential benefits from efforts to automate the analysis of facial signals, in particular rapid facial signals, are varied and numerous, especially when it comes to computer science and technologies brought to bear on these issues. The analysis of facial expressions has a great relevance in sociological, medical and technological researches. The automated analysis of facial expressions is a challenging task because everyones face is unique and interpersonal differences exist in how people perform facial expressions. In recent years, due to the availability of relatively cheap computational power, automatic facial expression analysis has been investigated as facial pattern recognition using imaging techniques. In this framework, many different techniques have been developed which claim to provide means for objectively measuring facial expressions. Some examples are represented by Expert Systems, Hidden Markov Models, Gabor filters and Optical Flow analysis. The approach presented in this paper uses statistical shape analysis to synthesize facial expressions. We work with high-dimensional data sets such as video sequences and summarize the expressions through a set of landmarks on the face. The shapes of the main features and the spatial relationships between them are then represented by a shape polynomial regression model which provides a compact, parameterized description of shape for any instance of a face. We shall show that the model, estimated using the Expectation Maximization (EM) algorithm, is useful to represent the dynamics of an expression as well as for the construction of an accurate classifier for shape (expression) allocation.

Statistical Graphics for Spatial Clustering Based on Inter-temporal Variation

Tomokazu Fujino, Fukuoka Women's University
Yoshiro Yamamoto, Tokai University

Abstract

We have developed statistical graphics called ranking comparison plot based on parallel coordinate plot with individual text labels colored by the factors of the other attributes. This graphics could be use for finding relationship of the order fluctuation of the specific variable to the other factors. For example, this graphics can visualize the relationship between variation of the car sales rankings of two car dealers and the retail price. At first, we thought that this graphics could be used in the process of exploratory data analysis, especially in the marketing field. In this study, we propose a new use of the graphics, which is for spatial clustering based on inter-temporal variation combining the ranking comparison plot and choropleth map.

Three-mode Analysis of Compositional Data: Some Graphical Displays

Michele Gallo, L'Orientale University of Naples

Abstract

Compositional data (CoDa) consist of vectors of positive values summing to a unit, or in general, to some fixed constant for all vectors. They appear as proportions, percentages, concentrations, absolute and relative frequencies. The CoDa are commonly present in all experimental fields, and their analysis pose a number of difficulties, the first of which is the spurious correlation. Thus, we require a careful consideration of the relationships between the parts of a composition so as they are properly incorporated into statistical modeling. There are several approaches to incorporate CoDa into statistical modeling when it is not realistic to assume a multinomial distribution of the data. Based on the log-ratio transformations, Aitchison (1986) proposed preprocessing the compositional data by means of some log-ratio transformations, and successively analyzing them in a straightforward way by traditional methods. Following Aitchison's approach, more multidimensional techniques have been adapted to analyze compositional data, for example, Principal Component Analysis (Aitchison, 1983), Partial Least Squares (Hinkle and Rayens, 1995; Gallo, 2003), Discriminant Partial Least Squares (Gallo, 2010), Hierarchical Cluster (Martìn-Fernàndez et al., 1998) are only some of multivariate techniques proposed in literature. Sometimes, the CoDa are arranged into three-way arrays, i.e. proportions of household expenditures on commodity groups of single professions in several countries; proportions of fixed and variable costs in bank budgets for several years; proportions of leucocytes in blood samples determined by different methods; proportions of different chemical compounds in several rivers in different seasons. In these cases, one of the main purposes of using three-way analyses is exploring the interrelations between data, where Tucker models (Tucker, 1966; Kroonenberg and de Leeuw, 1980) are the methods particularly suitable for this aim. They yield a low-dimensional description of the three-way arrays and, like the PCA on two-way matrices of compositional data (Aitchison, 1983), and they have proved difficult to handle statistically because of the awkward constraint that the components of each vector must sum to unity. Gallo (2011) has proposed a reexamination of the Parafac/Candecomp analysis (Harshman, 1970; Carol and Chang, 1970) in case of compositional data. It is based on a certain choice of prerequisites that the results should reasonably be expected to satisfy, i.e. scale invariance and subcompositional coherence, where scale invariance merely reinforces the intuitive idea that compositional data provide information only about relative and not about absolute values. The subcompositional coherence requires using full compositions and the subcompositions of these full compositions should have the same relations within the common parts (for more detail see Aitchison, 1986). The principal aim of this work is to describe how to do a Tucker analysis of compositional data and how to read the results correctly in some low-dimensional space.

Network Data as Complex Data Objects: An Approach Using Symbolic Data Analysis

Giuseppe Giordano, University of Salerno
Maria Paula Brito, Universidade do Porto

Abstract

Symbolic Data Analysis (SDA) aims at extending statistics and data mining methods from first-order (i.e. micro-data) to second-order objects (often obtained by aggregation of micro-data into more or less large groups), taking into account variability that is inherent to the data. In this work, we focus on network data, as defined in the framework of Social Network Analysis and Graph Theory. We represent each network as a complex data object, using the SDA approach, in order to establish general rules that allow defining a graph structure and the underlying network data, within this framework. The definition of a graph structure as a complex data object should consider the different structural information that can be of interest to retrieve. We may start from simple descriptive statistical measures that provide a first insight into the network structure. The basic idea is to aggregate information attached to each node in terms of its centrality and role in the network and express it as symbolic data by means of interval or histogram-valued variables so that the whole network could be expressed through the logical union of such different measurements. The final output should allow building a symbolic data table where each row pertains to a different network and columns to the network indices. That is, each row defines a Network Symbolic Object (NSO). Symbolic data analysis of NSO could be applied for the sake of comparisons among several networks emerged at different occasions in time, computing similarities among networks, or representing networks as "points" on a reduced embedding (metric space). A simulation study as well as an empirical case study will be provided.

A Proposal for Categorizing Nonuniform DIF in Polytomous Items

Silvia Golia, University of Brescia

Abstract

DIF is understood to be present when something about the characteristics of a test taker interferes with the relationship between ability and item responses. In a DIF analysis, the population is divided in two subgroups, that is reference and focal group. Two types of DIF can be identified: uniform and nonuniform. Uniform DIF (UDIF) occurs when the relative advantage of one group over another on a test item is uniform, favoring only one group consistently across the entire scale of ability. Nonuniform DIF (NUDIF) exists when there is interaction between ability level and group membership and the relative advantage of one group over another is not uniform across the entire ability continuum. In the case of UDIF, Penfield (2007) proposed a system for categorizing the severity of UDIF in polytomous items analogous to the ETS system for characterizing dichotomous items. The aim of the present study is to identify a way to assign a severity grade to DIF when NUDIF is present.

In order to do this, a simulation study is arranged. Different combinations of percentages of DIF items, DIF magnitude and test length will be considered. In order to study a possible categorization of the severity of NUDIF, the percentage of rejection of the null hypothesis underlying the two-sample Kolmogorov-Smirnov test applied to the real and estimated abilities, calculated from 100 data sets, and the value of the first eigenvalue of PCA on Rasch residuals for the focal group are used. The evaluation of both these two tools joined to the magnitude of DIF, allows a possible categorization of the severity of NUDIF.

Local Independence Statistics Applied to Testlet Items

Takamitsu Hashimoto, The National Center for University Entrance Examinations

Abstract

National Center for University Entrance Examinations in Japan (NCUEE) administrates the National Center Tests (NCTs), and there are two types of NCT English tests: the written test and the listening test. The written test contains about 50 items, and the listening test contains 25 items. These items construct some testlets. While items in late testlets are related to leading essays in the testlets, early testlets don't contain leading essays and items in these testlets don't related to each other items. However, independencies of such items have not been confirmed by data. Usually items are highly correlated with each other, because all items depend on examinees' ability variables. Therefore, conditional independencies given ability variables have to be confirmed. In order to confirm such conditional independencies, many methods have been proposed, and the latent conditional independence (LCI) test (Hashimoto and Ueno, 2011) is one of these methods. The feature of the LCI test is that the distribution of the test statistics is not affected by dependency structure of items. Using the LCI test, this study has confirmed whether early items of the NCT English tests are conditionally independent given ability variables, and has explored unexpected dependencies. Data in this study are responses of about 400 freshmen in 5 universities in Tokyo. Results of the LCI tests have shown that most items of the NCT written English tests are conditionally independent given ability variables, many items of the NCT English listening tests are conditionally dependent. Although these results have been obtained from university freshmen's data instead of high school student's data, the results suggest that the NCT written English tests might measure one-dimensional ability while the NCT English listening might measure multi-dimensional ability.

Quasi-exact Tests for Rasch Models

Reinhold Hatzinger, Vienna University of Economics and Business

Abstract

Many statistical and mathematical problems, like establishing a probabilistic framework for tests or removing the effect of nuisance parameters on a test involve sampling matrices with given row and column sums. Also when the asymptotic distribution of a test statistic is unclear or unknown it is useful to built quasi-exact tests by approximating the null distribution (the distribution if the model is valid) of the statistical model. In this talk some theoretical applications for simulating binary and ternary data matrices with given row and column sums will be presented. The aim is to sample discrete data matrices with fixed marginals to build quasi-exact tests for applications in the Rasch model context.

Statistical Sensitivity Analysis for Risk Prediction of Endoleak Formation after Thoracic Endovascular Aortic Repair

Kuniyoshi Hayashi, Okayama University
Fumio Ishioka, Okayama University
Bhargav Raman, Stanford University School of Medicine
Daniel Y. Sze, Stanford University School of Medicine
Hiroshi Suito, Okayama University
Takuya Ueda, St. Luke's International Hospital
Koji Kurihara, Okayama University

Abstract

During the past decade, thoracic endovascular aortic repair has gained popularity as a therapy for thoracic aneurysm. However, it has been observed that this procedure often causes a clinical side effect called "endoleak". In this study, to contribute to therapy planning, we evaluated the risk prediction of endoleak by statistical sensitivity analysis using linear discriminant analysis. Based on this evaluation, we identified the patients who were notable with regard to risk prediction and investigated their characteristics in detail.

Mixed Effect Models for Provider Profiling in Cardiovascular Healthcare Context

Francesca Ieva, Politecnico di Milano
Anna Maria Paganoni, Politecnico di Milano

Abstract

Provider profiling is the evaluation process of the performance of hospitals, doctors, and other medical practitioners, aimed at increasing the quality of medical care. Within this context, performance indicators for assessing quality in healthcare contexts have drawn more and more attention over the last years, since they enable researchers to measure all the relevant features of the healthcare process of interest, including the performances of healthcare providers, clinical outcomes and disease incidence. The purpose of this work is to highlight how advanced statistical methods can be used to model complex data coming from a clinical survey, in order to assess hospitals and healthcare providers performances in treating patients affected by STEMI (ST segment Elevation Myocardial Infarction), a disease with a very high incidence all over the world and where being well timed makes the difference in terms of patients' survival and later quality of life. The same methods are also used to classify hospitals according to the evaluation of their performances, compared with gold standards and guidelines. To these aims, we fit different models, trying to enhance the grouping structure of data, where the hospital of admission is the grouping factor for the statistical units, represented by patients. In these models we introduce performance indicators in order to adjust for different patterns of care and to compare their effectiveness in treating patients. Also the adjustment for case-mix (i.e., patients' features at admission) is considered. The clustering structure induced by the model assumptions is then investigated. In particular, we propose three different methods to evaluate hospitals performances: in the first one we estimate the in-hospital survival rates after fitting a Generalized Linear Model, using suitable indexes for testing the presence/absence of outliers; in the second one we fit a Generalized Linear Mixed Effects Model to explain in-hospital survival outcome by means of suitable patient's covariates and process indicators, with a parametric random effect accounting for hospital influence; then we perform an explorative classification analysis implementing a clustering procedure on the point estimates of hospital effects provided by the model. Finally, in the third case we classify hospitals according to the variance components analysis of the random effect estimates, where nonparametric assumptions have been considered. In fact, the discreteness of the random effect induces an automatic clustering that can be interpreted as identifying hospitals with similar effects on patients' outcome. The survey we consider for the case study, named STEMI Archive, is a clinical observational registry concerning patients admitted with STEMI diagnosis in any hospital of our regional district, i.e., Regione Lombardia. This registry has been designed and funded within a scientific project, named Strategic Program, aimed at the exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction. The nearly unanimous agreement of results obtained implementing the three methods on data supports the idea that a real clustering structure in groups exists. Such methods can provide useful decisional support to people in charge with healthcare planning.

Fuzzy c-means for Web Mining: the Italian Tourist Forum Case

Domenica Fiordistella Iezzi, Tor Vergata University
Mario Mastrangelo, Sapienza University of Rome

Abstract

The objective of fuzzy c-means is to partition a data set into c homogeneous fuzzy clusters. This algorithm is a version soft of the popular k-means clustering. As well known, the k-means method begins with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors. k-centers clustering is moderately sensitive to the initial selection of centers, so it is usually rerun many times with different initializations in an attempt to find a good solution. We propose a new method to initialize the fuzzy c-means, using a map grid. We apply this method to 525 posts published by the Italian tourist forums.

Direct Scaling of Asymmetric Similarity Matrix by a Communication Channels Model

Tadashi Imaizumi, Tama University

Abstract

We can find many asymmetric relation from an objects to another object of n objects, for example, a trading data between nations, frequency of talking of n persons, journal citation data etc. Then we want to explore the complex relationship of n objects in data matrix by some model and method. Okada and Imaizumi (1997) proposed a nonmetric scaling model and method for analyzing asymmetric similarity matrices. This model consists of a common object configuration and two kinds of weights, i.e., for both symmetry and asymmetry. In the common object configuration, each object is represented by a point and a circle (sphere, hypersphere) in a Euclidean space. The common object configuration represents pairwise proximity relationships between pairs of objects for the group of all sources. Symmetry of data is represented as Euclidean distances between two objects. Asymmetry of data is represented as circles, sphere, and hypersphere, etc. around the point of object. And this asymmetry is related to the characteristics of each object. Zeilman and Heirs's slide-vector model (1996) assumes the asymmetric data are induced by sliding of latent dimensions. And how to represent the asymmetry of data is big difference between these two models, one is by objects and another is by dimensions. As Okada and Imaizumi's model do not utilize any information of diagonal elements of similarity matrix in analyzing data for simplicity, their model will be applied to analyze dissimilarity matrices too. But, some important information about objects is also contained in the diagonal elements of similarity matrix as we know. So this model is lock of utilizing information in data. A redefined and extended model of Okada-Imaizumi will be proposed to account the diagonal elements and off-diagonal elements of asymmetric similarity matrix. Similarity data will be represented by a function of Euclidean distance and the number of communication channels, input channels and output channels, between corresponding two objects in this new model. A comparison with previous model and an application to real data set will be shown.

Spatial Classification and Clustering Approaches Based on Echelon Analysis

Fumio Ishioka, Okayama University
Koji Kurihara, Okayama University

Abstract

There are several approaches to classify the different kinds of spatial data. However, there are few classification methods based on spatial structure of data. Echelon analysis is a useful technique for investigating the phase-structure of spatial data systematically and objectively. In this paper, we perform a classification method based on the peaks of echelon structures. In addition, we detect spatial clustering using echelon's spatial scan statistic.

Validation of a Non-curriculum-based Ability Test through Factor Analysis with Consideration of External Performance Score Scales

Kei Ito, The National Center for University Entrance Examinations
Haruo Yanai, St. Luke's College of Nursing

Abstract

In the last decade in Japan, the social and educational background of applicants for admission to universities has been diversified, and the improvement of admission procedures has been discussed among teachers, researchers, universities, and interested parties. In 2000, the Council for Higher Education highlighted the necessity of evaluating students' performance in problem solving and task handling as well as their aptitude for higher education, which are difficult to measure through conventional subject tests. In this regard, the non-curriculum-based ability test (NCBAT) can be considered a practical tool for comprehensive evaluation from various points of view. Many studies have considered the basic properties of the NCBAT focusing factors such as difficulty, discrimination, and reliability. A validation study of the NCBAT through comparison with external criteria is, however, preferable in order to clarify the similarities and differences between the NCBAT and conventional subject tests. The comprehensive test measures non-curriculum-based abilities such as logical thinking ability, reading comprehension, expressiveness, and related abilities, whereas subject tests measure academic achievement. The moderate degree of correlation between both tests indicates that they measure not only their respective abilities but also common abilities. Therefore, we can determine the peculiar properties of the NCBAT by eliminating known information; for example, eliminating the subject tests' information from the NCBAT (Ito et al., 2010). Yanai (1970) proposed a factor analysis of predictors with external criteria on the basis of analysis of variance and regression analysis techniques. This method decomposes covariance matrix of predictors into two orthogonal parts by introducing the mathematical concept of projection, and makes it possible to extract two kinds of factors. One is closely related to information concerning the external criteria, and the other is independent of the same. On the basis of the scheme of the factor analysis with external criteria, we attempt to extract a distinctive factor of the NCBAT and to confirm its validity.

References

Kei Ito, Atsuhiro Hayashi, Kumiko Shiina, Masaaki Taguri, Ken-ichiro Komaki, and Haruo Yanai (2010). Validation of a Non-curriculum-based Ability Test by Comparison with Subject Tests and Self-evaluation Ratings, Japanese Journal for Research on Testing, Vol. 6, No. 1, 113-124 (in Japanese).

Haruo Yanai (1970). Factor Analysis with External Criteria: Application of Analysis of Variance and Regression Analysis Technique to Factor Analysis, Japanese Psychological Research, Vol. 12, No. 4, 143-153.

Observationally Equivalent Multi-level Models: Why Comparative Data often do not Discriminate between Theoretical Alternatives

Wolfgang Jagodzinski, University of Cologne

Abstract

While problems of measurement equivalence have been widely discussed in recent years particularly in the framework of SEM, problems of observational equivalence have been rarely examined. Two models with different constraints are called observationally equivalent if and only if they generate the same estimates of the sample variances, covariances and correlations. Multi-level models in comparative analysis often use a mixture of quantitative and dummy variables for specifying the context effects. While quantitative variables like variables GNPC per capita or HDI serve as indicators modernization, economic or human development, dummy variables are often interpreted as indirect measures of cultural, historical or political influences. The paper will demonstrate analytically and empirically that the resulting solutions are less exclusive than it is usually assumed. Largely different theoretical perspectives are supported by the data of international survey programs equally well - observationally equivalent solutions, in other words, can be found. The consequences for comparative research are drawn in the concluding section.

Spatio-temporal Clustering of Age Grouped Suicide Data in Japan

Takafumi Kubota, The Institute of Statistical Mathematics
Makoto Tomita, Tokyo Medical and Dental University
Fumio Ishioka, Okayama University
Toshiharu Fujita, The Institute of Statistical Mathematics

Abstract

We applied age grouped suicide data in Japan, which include four time periods, 354 areas and four age groups, to identify spatial/spatio-temporal clusters. We calculated spatial scan statistics of the suicide data, and then we used them to search highly concentrated areas. We discussed the comparison between the results of this study and the spatial/spatio-temporal cluster of total of all age groups in previous studies.

A SVM Applied Text Categorization of Academia-Industry Collaborative Research and Development Documents on the Web

Kei Kurakawa, National Institute of Informatics
Yuan Sun, National Institute of Informatics
Nagayoshi Yamashita, Japan Society for the Promotion of Science
Yasumasa Baba, The Institute of Statistical Mathematics

Abstract

To make a policy of science and technology research and development, university-industry-government (U-I-G) relations is an important aspect to investigate it. Web document is one of the research targets to clarify the state of the relationship. In the clarification process, to get the exact resources of U-I-G relations is the first requirement. The objective of this research is to extract automatically resources of U-I relations from the web. We set a target into “press release articles” of organizations, and make a framework to automatically crawl them and decide which is of U-I relations. Classification technique, i.e. support vector machine (SVM) is adapted to the decision. Japanese natural language processing is required to preprocess the web document and translate it into feature vector for SVM input for both learning and classifying. We have conducted an experiment for several combinations of feature vector elements and kernel function types. We will report essential summary of how the classifiers be effective in the experiment.

Improving Likelihood-based Estimates of the Rasch Model with ε-Adjustments

Tommaso Lando, University of Bergamo
Lucio Bertoli-Barsotti, University of Bergamo

Abstract

Within the framework of the item response theory, the maximum likelihood (ML) estimation method does not provide finite latent ability estimates for respondents who obtain an extreme score – when a Rasch Model (RM) is considered. To overcome this problem, computer programs usually produces finite ML estimates for null and perfect patterns by estimating the abilities that correspond to the scores *r* and *M–r* where *M* is the maximum possible score, and r is an arbitrarily specified real number (see, e.g., ConQuest by Wu, Adams, Wilson & Haldane, 2007). WINSTEPS, the most popular software for Joint ML estimation of the RM, offers two adjustments: the adjustment for estimation bias suggested by Wright and Douglas (1977) and the above mentioned ‘‘correction’’ for extreme values of the total scores (Linacre, 2009). To jointly solve these estimation problems at their source a new method is proposed: its key principle lies in the minimum fit function (or minimum divergence) method. More specifically, this estimation procedure is obtained by modifying the Kullback-Leibler divergence (on which the MLE is based). As the sensitivity of this adjustment depends on an arbitrarily small positive number **ε**, we called it the **ε**-correction method. The **ε**-correction warrants the existence of the estimates for both person and item parameters, not only in presence of extreme score, but also for ill-conditioned datasets (Fischer, 1981) and in presence of null-categories (Luo & Andrich, 2005). The right choice of **ε** is a little troublesome, it can depend on several factors, including the dimension of the dataset, anyway it can be found empirically and approximately. Simulation studies show that, setting **ε** conveniently, the new method is very effective in correcting the bias of the MLE, and for this purpose it seems to work even better than the other known method.

Dynamic Customer Satisfaction and Measure of Trajectories: a Banking Case

Caterina Liberati, University of Milano-Bicocca
Paolo Mariani, University of Milano-Bicocca

Abstract

Customer satisfaction is the most important asset of a company and, in the last years, it has been successfully applied also in banking sector. Such scenario leads banks to employ intelligent systems to monitor their own clients during the time, analyzing the evolution of customer satisfaction and tracking patterns of customer evaluations related to bank service features in order to build and preserve a robust relationships with them. The three-way analysis technique offers an effective solution to merge information about clients and evolutions of their preferences. Moreover, studying trajectories patterns, which drawn direction and intensity of the changing occurred in the sample, we derived a new measure to synthesize such information relatives to a single customer or cluster or typologies. However in literature already exist well known techniques which compare the temporal trajectories using dissimilarity measures and classification criteria, the innovative aspect of our index lies in the way we summarize characteristics of the trajectories such as distance covered, the shape and direction. Using a questionnaire, framed in according to SERVQUAL model, the case of a national bank with a spread Italian network has been analyzed. Information was obtained via a satisfaction survey repeated in 3 different temporal waves on a sample panel composed by 27.000 instances per wave. A multi-way factorial model has been carried out on a three-way data matrix X_{ijk} (professional segments (i), service attributes (j), waves (*k*)) of 9×24×3 dimensions. Using these results it will be illustrated an empirical study of our approach to time trajectories in order to highlight promising performances of our index.

Bootstrap Confidence Regions in Classical and Ordered Multiple Correspondence Analysis

Rosaria Lombardo, Second University of Naples
Trevor Ringrose, Cranfield University
Eric J. Beh, University of Newcastle

Abstract

Multiple correspondence analysis (MCA) is an exploratory technique for graphically representing the associations between categorical variables. In much of the literature on MCA, attention has largely focused on the determination and interpretation of points representing the categories of the variables in a low dimensional plot, including their proximities to the origin and each other. Little attention has been paid to various inferential aspects, in particular the sampling variation of the configuration of points. In the present paper, we discuss a particular inferential aspect concerning the configuration of points in classical MCA and ordered MCA (OMCA). Our focus is on deriving confidence regions (CRs) for each data point in a low-dimensional plot. Various strategies have previously been proposed that discuss a variety of different ways to obtain CRs for simple correspondence analysis. We can distinguish among regions derived in a purely algebraic way, which may be circular or elliptical, regions based on asymptotic statistics, such as confidence ellipses calculated by the Delta method, and bootstrap-based regions like convex hulls by the partial bootstrap. This paper considers an alternative bootstrap approach previously used for quantitative variables and recently proposed for categorical variables by Ringrose (2011). This approach was developed for simple correspondence analysis, taking into account the variation of the axes, and constructs CRs based on the variability in the difference between the sample point and the population point when both are projected onto the sample axes. The stability of the configuration of points is particularly important when MCA is considered as a tool for decision making. This is likely to be the case with OMCA, where the categorical variables are ordered and orthogonal polynomials may be considered as suitable transformation functions resulting in an OMCA of the data. Bootstrap confidence regions for biplot displays of both MCA and OMCA will be considered, adding information on the reliability of the analysis results.

Social Network Analysis for Social Planning Evaluation

Rosaria Lumino, Federico II University of Naples
Concetta Scolorato, Federico II University of Naples

Abstract

The paper focuses on the application of Social Network Analysis (SNA) for social planning evaluation, especially in regard of concertation, trough the analysis of a case study: Territorial Youth Plans (PTGs). These latter were introduced by the Campania Region in 2009 at enhancing young citizenship and participation in decision making through. In this paper we exploit the Social Network Analysis tools to study the concertation networks among public and private social actors in the local contexts and to analyze if and how the structure of interactions between actors affects quality of plans in terms of coherence, effectiveness and innovation. The relational data have been collected by taxonomic analysis of official documents. The analysis has three major objectives: (i) to describe the structure of networks on the base of topologic and morphologic parameters; (ii) we map each network into a multidimensional space and apply the usual cluster techniques to obtain cluster of networks that are similar in term of relational structure; and (iii) to relate variability of clusters to different forms of social planning, assuming that different relational structures of networks shape social planning activity and local innovation ability.

Preference Mapping by PO-PLS: Separating Common and Unique Information in Several Data Blocks

Ingrid Måge, Nofima AS
Elena Menichelli, Nofima AS
Tormod Næs, Nofima AS

Abstract

Multivariate data analyses such as Principle Component Analysis (PCA) and Partial Least Squares (PLS) regression are the foremost tools for interpreting relationships between many variables. In a situation with several data blocks, where each block is a collection of related variables, a straight-forward solution is to put all variables together into one large data matrix and analyse it with conventional PCA or PLS, depending on whether a predictive direction is present or not. The main drawback is that variables from different blocks are mixed together, which might obscure interpretation. Results will also depend heavily on how the different variable blocks are scaled relative to each other. Furthermore there may be problems in case of different dimensionality within each of the blocks. An approach which solves the problem of different scale is Generalized canonical Correlation Analysis (GCA), but this method has other drawbacks related to over-fitting and instability when the number of variables is large. In the chemometrics area new methods have recently been developed for solving these problems. They are based on sequential use of PLS regression on matrices that are orthogonalised with respect to each other. These methods are invariant with respect to the relative scale of the data blocks, allow for different dimensionality of the blocks, allow for high collinearity and enhance interpretation. One of these approaches is the so called Parallel Ortogonalised Partial Least Squares (PO-PLS) method. It can first identify a redundant information in the blocks and then detect the information in each block that is non-overlapping with the other. The algorithm uses a combination of PLS regression and GCA to identify the unique components across multiple predictor data blocks. The idea is to first identify a subspace that is common for the input blocks and then orthogonalise the individual blocks with respect to this space in order to identify the unique information in the blocks.

The present study is on the use of PO-PLS in the area of preference mapping. The approached situation is product development, where actual prototypes are often assessed by several measurement principles. In consumer science they can be chemical analysis, descriptive sensory analysis and various types of consumer liking or choice tests. In particular sensory variables are often divided into several categories, such as visual appearance, smell, taste and texture. The focus of the study is on the relation between sensory attributes and consumer liking, with a special focus on how the different blocks of sensory data relate to each other and to the consumer preference data. One can also incorporate instrumental data together with sensory data in preference mapping using one single analysis. It is shown that this type of modelling can be used for obtaining more interpretative information than standard preference mapping by PLS regression. While PLS only tells us the dominant components in all predictor blocks, PO-PLS shows how the variability is distributed between blocks.

Data Analysis of the AsiaBarometer Survey: Well-being, Trust and Political Attitudes

Kazufumi Manabe, Aoyama Gakuin University

Abstract

The purpose of this presentation is to illustrate the utility of Facet Analysis, especially, Smallest Space Analysis (SSA) developed by Louis Guttman for the data analysis of cross-national survey using the example of the AsiaBarometer survey. The AsiaBarometer is a large scale multi-national questionnaire survey conducted at regular intervals (every year from 2003) within the Asia region. This presentation analyzes the data from the 4th AsiaBarometer survey which was conducted from June to August 2006 in seven countries: China, Hong Kong, Japan, South Korea, Singapore, Taiwan, and Vietnam. The AsiaBarometer survey is a cross-national survey. In cross-national surveys, equivalence of measurement and comparability must be examined very carefully. When comparing the countries surveyed, it is important not to compare a single question item, but to compare the structure of the relationship between multiple question items using a method of multivariate analysis. This presentation uses a technique of Facet Analysis that is, Smallest Space Analysis (SSA) developed by L. Guttman. SSA is a type of multidimensional scaling method, and illustrates graphically the structure of relationship between n items shown in a correlation matrix by the size of the distance between n points in an m-dimensional (m<n) space. The higher the correlation between two variables, the smaller the distance between them on the map, and the lower the correlation, the larger the distance. Usually a two-dimensional (plane) or three dimensional (cube) space is used to visually depict the relationship between items. his shows that SSA is the most appropriate method of visually depicting the overall structure and relationships among question items. SSA is applicable as a very effective tool in examining equivalence of measurement when conducting cross-national surveys. In cross-national surveys that can yield SSA maps showing the same spatial structures, it is highly likely that commonality of meanings can likewise be established in the countries compared. This is an important reason for using SSA. The AsiaBarometer survey contains question items that measure the respondents' attitudes and behaviors in various aspects of everyday life, as well as items intended to measure their opinions regarding various relationships between people and society, their values related to freedom, human rights, and democracy, and their political behaviors. Thus, the question of which specific question items will be addressed in this analysis is the next point. I will try to deal with the following three groups of question items: (1) question items on well-being, (2) question items on social trust, and (3) question items on political attitudes.

Causal Models for Monitoring University Ordinary Financing Fund

Salvatore Marcantonio, University of Palermo
Antonella Plaia, University of Palermo

Abstract

Recently iterated decreasing government transfers and an increasing proportion of budget allotted basing on competitive performances, took Italian Universities started struggling with competition for funds, in particular for the University Ordinary Financing Fund (FFO). For this reason many universities have decided to monitor the FFO indicators, in order to know which are the variables responsible for them, their present and past values and how they relate with national references, what could be the future values of these variables, what are the components of strength and weakness, in order to look at measures for the correction of weaker performances and what strategies can be taken to increase the value of the indicators. Aim of this paper is monitoring variables responsible for FFO indicators, where monitoring means: describing, analyzing retrospectively, predicting and intervening on indicators. All this aims can be achieved by statistical techniques and models that should be theoretically equipped with the distinction between predicting under observation and predicting under intervention, in order to provide correct answers to the distinct tasks of pure out of sample extrapolation and policy making. They should be also capable of encoding not only information arising from empirical data, but also from extra knowledge, such as expert opinions, in order to quickly adapt to new possible scenarios and keep a genuine uncertainty about a priori information. A suitable statistical technique for this task is time series analysis, embedded into the Bayesian framework for two main reasons: because for each quantity to be forecast there are only a few observations available, and because time dependent observations call for flexible models, able to adapt rapidly to system shocks or to external interventions, like models with (time) varying parameters, a basic feature included in the Bayesian modelling. Sometimes knowing that two events are associated each other is not sufficient to ask questions we are interested in. This is especially the case when one event is interpreted as a cause and the other as the effect, since a single association relationship gives rise to two distinct causal interpretations: a casual relationship from the cause to the effect and a diagnostic relationship from the effect to the cause. Pearl’s methodology formalizes such difference using the (standard) notation of conditional probability for describing observational relationships (given that you see), while using a (new) notation of do() operator for describing causal effects (given that you do). Causal analysis requires more knowledge of the data generation process than observational analysis. Such methodology combines features of structural equation models, with the explicit use of latent variables, with potential outcome and graphical models. It gives an operational definition of intervention, with use of a new notation, (do(X = x)). By using this approach, this paper will particularly focus on policy evaluation and intervention policy in order to find what variables a university need to intervene upon.

A Semiparametric Regression Framework for Modelling Correlated Binary Responses

Giampiero Marra, University College London
Rosalba Radice, London School of Hygiene & Tropical Medicine

Abstract

First proposed by Ashford and Sowden (1970, Biometrics), bivariate probit regression has been one of the first methods used to study correlated binary data within a regression framework. Since then, such a model has been applied to biological, medical, sociological and economic data. A bivariate probit model provides a framework to study correlated binary responses, endogeneity and sample selection. In this talk, I will show how this can be flexibly achieved. Specifically, I will consider the semiparametric version of this model and introduce a fitting procedure which permits to model a system of two binary outcomes as a function of parametric terms, smooth functions of continuous covariates and known linear functionals (usually dependent on covariates). The issues of inference and variable selection will be also briefly discussed. The methods will be illustrated using data from a survey on the impact of education on women's fertility and data from the American National Election Study on the quantification of public support for school integration, where the issues of endogeneity and sample selection arise.

Prior Information at Item Level in Item Response Theory Models

Mariagiulia Matteucci, University of Bologna
Bernard Veldkamp, University of Twente

Abstract

Recently, Bayesian estimation of item response theory (IRT) models via Markov chain Monte Carlo methods has become very popular. The main reason is that this method is free from the limitations of using Gaussian quadrature in marginal maximum likelihood estimation and it is more easily extendable for the estimation of models with complex structures. Moreover, a Bayesian approach allows the incorporation of dependencies among variables and sources of uncertainty. The role of prior distributions is very important in Bayesian statistics, and informative prior distributions can be used in order to improve the accuracy of parameter estimation under particular conditions, for example when the sample size is small. This last aspect is too often disregarded by researchers, who are used to include in the model flat, uninformative priors, especially for item parameters. Differently, this work shows how the introduction of informative prior distributions, even empirical, is effective in improving the accuracy of model estimation. In particular, we will consider the introduction of prior information on the item parameters, which are treated as random variables within a Bayesian approach. Including informative prior distributions at item level does not state the matter of fairness, as it is argued in case priors are used at ability level. The use of empirical priors is also discussed, with respect to intelligence test data.

Statistical and Non-statistical Models in Clustering: An Introduction and Recent Topics

Sadaaki Miyamoto, University of Tsukuba

Abstract

Relations between statistical methods such as the mixture of Gaussian distributions and fuzzy clustering are overviewed. We first introduce a basic K-means algorithm, fuzzy K-means, and the mixture of distributions. We show when the latter two give equivalent results. In contrast, we review agglomerative hierarchical clustering, where significance of fuzzy clustering is mentioned and no similar method is found in a statistical model. Namely, fuzzy clustering implies the transitive closure of a fuzzy graph that is equivalent to the well-known single linkage. A family of new methods can be derived from closer observation of fuzzy clustering. A more recent topic of kernel-based clustering is discussed where fuzzy K-means and the above equivalence result suggest a new method of clustering. Simple illustrative examples are shown in order to understand theoretical results intuitively.

Evaluation of Nursing Homes Using an Extended Latent Markov Model

Giorgio E. Montanari, University of Perugia
Silvia Pandolfi, University of Perugia

Abstract

We illustrate the use of an extended version of the latent Markov model with covariates to analyse data coming from a longitudinal study on the health status of elderly patients hosted in Italian nursing homes. In this context, the latent Markov model finds a natural application since it relies on a not directly observable Markov process, whose states represent different levels of the health status of the elderly. In particular, we consider a dataset collected by the repeated administration of a questionnaire aimed at measuring the quality of life of the patients during their period of residence in the nursing homes. The questionnaire is made up of a large number of polytomous items, concerning several aspects of the everyday life. In this paper we consider only a subset of items which provides an amount of information close to that of the full set. These items have been selected in a previous work of the same authors aimed at introducing a strategy of item reduction that can be useful in this longitudinal context of repeated measures. The survey considered in this work has been carried out since 2004, and the residents have been evaluated at the baseline and then re-evaluated at 6 and 12 month. Accordingly, the interval of time between consecutive occasions is, in general, equal to six month but there are several exceptions, due to the possibility of repeated charge and discharge in the same nursing home, or to the death of the patients. For these reasons, the number of occasions is not constant across patients. The approach we propose is based on an extended version of the latent Markov model for polytomous items where both the initial and transition probabilities of the latent process depend on time-constant and time-varying covariates. The proposed model also allows us to account for missing responses on the basis of the assumption of latent ignorability. We consider two types of missingness: (i) missing response to an item; (ii) missing observation. The latter may be due to discharge, death, or to other unknown reasons. Through the assumption of latent ignorability it is possible to characterize the latent trait also in terms of probability that a response to an item is given and in terms of probability that an occasion of administration is recorded for a given patient. We then illustrate how the extended version of the latent Markov model may be applied so as to estimate the effect of each nursing home on the probability of transition between latent states. The objective is to evaluate the nursing homes performance with respect to their ability in improving the health condition of their patients or in delaying the worsening of these conditions.

Examination of the Necessity of Analysis of Triadic Distance Model

Atsuho Nakayama, Tokyo Metropolitan University

Abstract

Some previous studies of triadic distance models have enabled analysis of one-mode three-way proximity data for showing the triadic relationships among three objects. However, the results from a triadic distance model are similar to those of a dyadic distance model. So, it would be necessary to reveal the reason for such similarity and to establish the method to examine the necessity for the analysis of triadic distance model. But the method to examine the necessity for the analysis of triadic distance model has never been established. The purpose of present study is to examine the necessity of the analysis of triadic distance model. It examined whether the analysis of one-mode three-way MDS is necessary or not.

Two cases would exist behind the relationships among three objects in one-mode three-way proximity data. One is explained by the dyadic relationships and the other is explained by the triadic relationships. The results from a triadic distance model would be very similar to those of dyadic distance model under strong influence of the dyadic relationships in three-way proximity data. The relationships among three objects in one-mode three-way proximity data can be treated by dividing it into the dyadic relationships. However, the results from a triadic distance model would be dissimilar to those of a dyadic distance model under strong influence of the triadic relationships. The relationships among triadic objects in one-mode three-way proximity data cannot be explained as dyadic relationships. It might be appropriate to analyze the triadic relationships using a triadic distance model.

The present study proposes the method to evaluate the need for triadic distance model and shows the similarities and differences between triadic and dyadic relationships. Finally, the proposed method were applied to purchase behavior data.

Time and Network Structure: an Exploration of Change in Organizational Constellations

Francesca Odella, University of Trento

Abstract

The explanation of change in networks' structure and contents has been dealt in literature with models that highlight the function of structural elements and more recently with models separating choice and selection of connections at the individual level. The first type of models consider time as one of the agents of change, implicating possibilities for specific forms of connections (e.g. evolution from dyadic to triadic forms, from simple linkage to reciprocity links): according to this view the evolution of a network structure is dependant on the possibilities that are present in the previous stage (the mathematical reference is a Markov chains), such as for the case of preferential attachment phenomena. The second type of models interpret time as the setting of changes: the pace of creation and dissolution of network connections is related to individuals choices and activated by social selection processes that happen in specific time intervals. Both these types of models have been empirically tested, giving satisfactory results; the type of data, however, still influences measurements and computational efficacy. The paper reviews the most typical approaches to longitudinal network approaches and provides empirical examples of analysis of multiple organizational data according to these methodological principles.

Evaluating the Effect of New Brand by Asymmetric Multidimensional Scaling

Akinori Okada, Tama University
Hiroyuki Tsurumi, Yokohama National University

Abstract

Introducing a new brand into a market is one of the most important issues in marketing. There are many researches that analyze the effect of newly introduced brand into a market. Many of these researches focus their attention on sales quantity or market share of a newly introduced brand. The market share or the sales quantity are important in evaluating a newly introduced brand, but they are results of brand switching caused by the introduction of a new brand. These studies cannot disclose the reason(s) which resulted in the market share or the sales quantity. It is needed to deal with not only the sales quantity or market share of a newly introduced brand, but also to disclose the reason(s) which caused them. In the present study, we focus our attention to the brand switching from the existing brands to a newly introduced brand. A procedure of evaluating the effect of a newly introduced brand into a market based on the brand switching among existing brands as well as from the existing brands to a newly introduced brand is presented. The procedure utilizes the asymmetric multidimensional scaling based on singular value decomposition. The brand switching among *n* existing brands and *m* newly introduced brands is represented by an *n*×(*n*+*m*) brand switching matrix, each element of the matrix represents the frequency of the brand switching from the brand corresponding to the row to the brand corresponding to the column. This means that the brand switching among existing *n* brands and that from the existing brands to the newly introduced brands are dealt with, but the brand switching from the newly introduced brands to the existing brands and that among newly introduced brands are not available which do not occur soon after the introduction of a new brand. The asymmetric multidimensional scaling decomposes an *n*×*n* brand switching matrix among *n* existing brands (among existing brands) by singular value decomposition into three matrices; a matrix having left singular vectors of the unit length as its column, a diagonal matrix whose diagonal elements are singular values in the descending order, and a matrix of right singular vectors of unit length as its column. Each diagonal element represents the weight of the corresponding dimension. Each element of the left singular vector represents the outward tendency which shows the strength to be switched from the corresponding brand to the other brands. Each element of the right singular vector represents the inward tendency which shows the strength to be switched to the corresponding brand from the other brands. The asymmetric multidimensional scaling thus gives the outward and inward tendencies of the existing brands. The inward tendency of a newly introduced brand is derived by the external analysis of the asymmetric multidimensional scaling based on the derived outward tendency of the existing brands. This makes it possible to evaluate the strength of the newly introduced brand in the brand switching with existing brands.

A Bayesian Asymmetric MDS for the Radius-Distance Model

Kensuke Okada, Senshu University

Abstract

Bayesian analysis of multidimensional scaling (MDS) using Markov chain Monte Carlo (MCMC) technique has received considerable attention. Although most of the existing methods employed the symmetric MDS models, recently Bayesian estimation in an asymmetric MDS was proposed (Okada, 2012) employing the hill-climbing model (Borg & Groenen, 2005). However, this model assumes that the "slope" of the configuration space is the source of the asymmetry; therefore, the asymmetric effect is constant for all objects. This can be a rather strong assumption, especially when the number of objects is not very small.

In this paper we propose a Bayesian approach to the Okada-Imaizumi (OI; Okada & Imaizumi, 1987) model for the asymmetric MDS. In the OI model, each object is represented as a point and a circle (sphere) whose center is that point in the configuration space. The radius of the circle tells the asymmetry of the corresponding object. From these characteristics, this model is also known as the radius-distance model. Maximum likelihood estimation method for this model was already developed by Saburi and Chino (2008). The proposed method can be seen as a Bayesian alternative to their method.

As compared to the standard Bayesian symmetric MDS (Oh & Raftery, 2001), the difference between the symmetric and asymmetric (OI) model lies in the radius parameter of the circle (or sphere) which is attached to each object. Therefore, we can use the same prior distributions for the other parameters. For the radius parameter, the prior distribution can be constructed from the perspective of circular data Bayesian modeling (Ferrari, 2010).

The proximity measure of the OI model reduces to the usual Euclidean distance when the radius vector r equals zero. Therefore, the appropriateness of introducing the parameter that expresses the asymmetric structure, the radius vector, can be evaluated by simply determining whether the posterior credible region of the radius vector includes the origin. If the (e.g.) 95% Bayesian credibility interval does not include the origin, an asymmetric relationship between the variables is suggested. However, if it includes the origin, the use of a standard symmetric model may be considered as a better alternative.

An MCMC algorithm is employed to generate samples from the posterior distribution. Parameter estimation of the proposed Bayesian asymmetric MDS is conducted with the BUGS MCMC engine (Lunn et al, 2000), which is a programming language-based software that is used to generate random numbers from the joint posterior distribution of the parameters. An Monte Carlo simulation study and an application to the real asymmetric data is presented to verify and illustrate the proposed method.

Conditional Likelihood of Stratified Categorical Data for Analysis on a Large Scale Examination

Tatsuo Otsu, The National Center for University Entrance Examinations

Abstract

In research of social and/or behavioral sciences, we frequently need to analyze stratified contingency tables. Many important research questions are represented as inquiries of statistical structures for observed contingency tables. One popular case is testing independency of rows and columns of stratified 2 by 2 tables. It is well known that profile likelihood based estimation of log-linear models suffers from large biases when the data are sparse. Several methods for avoiding these biases have been developed. One of the most well known statistical tests for this type of data is Mantel-Haenszel test, a classical method that was devised in 1950's by researchers in medical statistics. Another popular method is conditional logistic regression, which supposes existence of a common odds ratio between the stratified tables.

Both methods are based on hyper-geometric distribution, which is a conditional multinomial distribution that has constraints on margins. A property of exponential distribution family makes conditional likelihood method be free from severe biases. Although these methods for stratified 2 by 2 tables are popular in practical data analysts, methods for more general stratified contingency tables are currently rare. One reason of this unpopularity is technical difficulty for computing conditional distributions in more general conditions. Recent advancement in computing machinery and mathematical theory have enabled practical use of conditional distributions in more general contingency tables. One of most influential developments was the introduction of computational algebra for obtaining Markov basis for MCMC generation of conditional distribution on contingency tables. Here we will show that estimations based on conditional distributions of stratified contingency tables are practically useful with some computational tools. One is the Markov basis obtained by use of algebraic theory, and another is constraint logic programming on finite domains CLP(FD) for enumeration with complex constrains. We analyzed a data of examinees subject selection of a nationwide university admission test in Japan (the National Center Test) with these methods. A significant gender effect in subject selection for natural sciences was observed in the data.

Archetypal Variables

Francesco Palumbo, Federico II University of Naples
Maria Rosaria D'Esposito, University of Salerno
Giancarlo Ragozini, Federico II University of Naples

Abstract

Archetypal analysis, as proposed by Cutler and Breiman (1994), aims at synthesizing single-valued data sets through a few (not necessarily observed) data points that are called Archetypes, under the constraint that all points can be represented as a convex combination of the archetypes themselves and that the archetypes are a convex combination of the data. Recently archetypes have been extended to the case of interval-valued data (D'Esposito et al, 2012), and to the case of functional data (Costantini et al. 2012). In the present paper we switch the attention from data points to variables. We propose to apply archetypal analysis in the variable space to obtain archetypal variables. The interpretation and use of these latter is twofold. On one hand, as archetypal variables are convex combination of the original ones, they can be intended as few variables that summarize all the others, and can be used for variable clustering and variable selection. On the other hand, as all the other variables are convex combination of the archetypal variables, these latter can be interpreted also as latent factors, i.e. archetypal analysis in the variable space could represent a sort of factor analysis with special constrains on factor loadings.

P

Statistical Characterization of Virtual Water Trade Network

Alessandra Petrucci, University of Florence
Emilia Rocco, University of Florence

Abstract

The volume of water consumption required to produce a commodity is called the “virtual water” contained in the commodity. If one country (one region, company, individual, etc.) exports a water intensive product to another country, it exports water in virtual form. Virtual water flows between nations are then calculated by multiplying the international trade flow of a particular commodity by the associated virtual water content of that commodity in the country of export. Virtual water content of a commodity is usually estimated by hydrological models. Hanasaki et al. (2008) suggest, to evaluate the virtual water content of a commodity, a global hydrological model called H08 that operates on a 0.5°×0.5° grid spatial resolution with water and energy balance closure and requires two types of input data: meteorological forcing and land use. Amid an increasing water scarcity in many parts of the world, virtual water trade as both a policy instrument and practical means to balance the local, national and global water budget has received much attention in recent years. Nevertheless, according to our knowledge, a complete statistical characterization describing the network of virtual water flows on a global scale has only started (Konar et al., 2011; Suweis et al., 2011) and a model to explain its characteristics is still lacking. Konar et al. (2011) applied analytical tools of complex network analysis to characterize the global structure of the virtual water trade associated with the international food trade. The aim of this paper is to improve the statistical characterization of the “virtual water network” developing modelling approaches and extending the network methodology also to the study of domestic virtual water trades. Knowing the actual national virtual water balance is essential for developing a rational national policy with respect to virtual water trade: to improve global water use efficiency, to achieve water security in water-poor regions of the world and to understand systemic risk, particularly under the potential impacts of climate change, as well as opportunities for network optimization. But for some countries relatively dry in some parts and relatively wet in other parts domestic virtual water trade is a relevant issue.

An Ontology-based Mixed Method. Evidences from Cultural Tourism in Anacapri

Ilaria Primerano, University of Salerno
Stefania Spina, Federeco II University of Naples

Abstract

In social science methodology there exist two main epistemological paradigms the Qualitative and the Quantitative one. The current literature distinguishes the two paradigms on the basis of different criteria: i) the type of data used (nominal vs. continuous scale), ii) the logic employed (inductive vs. deductive), iii) the type of investigation (exploratory vs. confirmatory), iv) the method of analysis (interpretative vs. analytic) and v) the underlying approach (positivist or critical vs. rationalistic or naturalistic). Data collection qualitative methods (e.g.: focus groups, participant observations and open text) involve all empirical research techniques that are interested in understanding the outcomes in a few cases, i.e. case-oriented methods, by contrast, quantitative methods (surveys, questionnaires and meta analysis) are interested in identify relationships in a well-defined population, i.e. population-oriented methods. One of the current debates is the idea that quantitative and qualitative methods are compatible, i.e. they can both be used in a single research study. The answer to this problem of compatibility of the approaches is Mixed Methods Research (MMR). The underlying goal of MMR strategy is that to provide a link between the qualitative and quantitative approaches in order to bring an additional value to the investigated object. This value consists either to create new theories as interpretative applications or to better understand phenomena by getting more out of the data. The final aim is not to replace each of the two approaches but rather to draw from the strengths and minimize the weaknesses of both in single research studies and across studies. MMR are used in many disciplines, such as in management and organizational research, in health sciences, in sociology, in psychology and education. In this paper we consider the theoretical framework of MMR and wish put emphasis on a joint reading of qualitative and quantitative methods. Indeed, the main difficulty to carry out a MMR lies in the separation of the amount of information derived by qualitative and quantitative analyses. For instance, since databases of these two approaches usually do not communicate because they are built using different languages. It needs to define a common framework able to recode them in the same language. In order to represent a shared conceptualization, our proposed solution lies in building ontology, which are explicit specifications of the conceptualizations at a semantic level. We insert them in a common set of categories that are able to catalogue in a proper manner the information present in the two initial databases. The aim of this paper is to propose a method to combine data coming from typical qualitative research (e.g. focus groups or interviews) together with survey data typically addressed by means of factorial techniques (e.g. Multiple Correspondence Analysis), through a common statistical coding of their ontological reading. A case study will show the meaningful of the proposed method, while simulations will be used to validate the results.

Evaluating Cooperative Behaviours in Innovation Networks: a Pre-specified Blockmodeling Analysis

Laura Prota, University of Salerno
Maria Prosperina Vitale, University of Salerno

Abstract

Collaboration networks linking firms, universities and local governments have since long been considered the loci of innovation. To favour the development of such networks of knowledge, an increasing number of Technological Districts (TDs) has been established worldwide through policy interventions. A major problem emerged in relation to such policies, however, concerns the development of consistent evaluation criteria. Each TD has a different specialization field and governance structure, limiting the applicability of cross-cutting performance indicators.

The paper takes the perspective of behavioural additionality to understand to what extent the institution of districts has persistently altered collaborative behaviours among actors favouring knowledge spread.

Pre-specified blockmodeling is proposed as a strategy to verify the structural changes in the pattern of research collaborations. Ideal network configurations will be tested against observed data at different time points to evaluate the development trajectory of each district with regard to its previous stage. The collaboration network is constructed from the co-partecipation of actors in research projects publicly funded of a TD in Southern Italy.

Boxplot for Data Streams Summarized by Histogram Data

Lidia Rivoli, Federico II University of Naples
Antonio Irpino, Second University of Naples
Rosanna Verde, Second University of Naples

Abstract

In the framework of data stream analysis, we suggest to use histograms in order to summarize the data stream sub-sequences detected through non overlapping windows. We propose a method for defining the histogram order statistics: median, quantile, minimum, maximum histogram. The definition of these statistics is based on the Wasserstein distance already proposed in several analysis contexts when data are expressed by distributions (empirical by histograms or theoretical by probability distribution functions). Starting from the definition of order statistics for a set of histograms, a new box plot like representation as well as some variability measures are performed. Moreover, consistently with the requirements of data stream mining algorithms, the use of histograms as summaries has allowed to get some computational simplifications.

R

Automatic Contour Detection and Functional Prediction of Brain Tumor Boundary

Elvira Romano, Second University of Naples
Iulian T. Vlad, Universitat Jaume I
Jorge Mateu, Universitat Jaume I

Abstract

Detection of tumor boundary and a further prediction of the tumor dynamics are necessary to verify the results of a particular treatment. Manual alignment of the images is a procedure that is often used. However, this process is quite necessary but also in some cases, inexact. In this paper we provide a method to perform an automatic contour of the boundary of brain tumor obtained through magnetic resonance mages (MRI) and a functional linear model to monitor its evolution. The automatic determination of the contour points is based on the longitudinal resolution (number of points of vector defined in a region of interest ROI) and gives a number of points that reconstruct the shape of the tumor. The continuous functions that lie behind the detected contours are then reconstructed and used in order to monitor the tumor evolution by a functional linear model.

A Value Added Approach in Upper Secondary Schools of Lombardy by OCSE-PISA 2009 Data

Isabella Romeo, University of Milano-Bicocca
Brunella Fiore, University of Milano-Bicocca

Abstract

In the last decade, policy makers have expressed interest on standardized tests to provide a measure of the impact of both school and teacher in student achievement. Value added methodology (VAM) has emerged as an attractive method to determine the contribution of teacher and school in students' achievement, after controlling for student and school factors. To this purpose, longitudinal data on students' individual career are necessary. This paper considers value added models of school assessment and their implementation in the secondary schools of the Lombardy region using Ocse-Pisa 2009 data. These represent an extremely valuable source of information. They are collected to get information on 15-year-old students' abilities to use their knowledge in an international setting. Ocse-Pisa is a survey that takes place every three years with a different major subject area: reading, mathematics and science. In this paper, the analysis has concentrated on reading scale scores which represent the focus of Ocse-Pisa 2009 survey. It is not possible to exploit the several Ocse-Pisa surveys to get longitudinal information on students' achievement given the cross sectional nature of these surveys. For this reason, the measure of the initial reading level of each pupil is based on both the final evaluation in the leaving examination of lower secondary school (about two years before the Ocse-Pisa survey) and the marks reported in the first year of upper secondary school. These evaluations are summarized in one indicator calculated with the Rash analysis. We have estimated a two-level hierarchical linear model to properly take into accounts for the multilevel structure of the data, considering both students' and schools' characteristics. Along with the initial reading level of the pupils, this paper introduces in the model a motivational reading variable. Indeed, taking into account information on student motivation (which has been largely studied in the literature) resulted in more reliable student value-added measures. Value-added measures are obtained for each student as difference between the predicted and the observed reading scores. These measures represent contribution of schools to individual student academic growth. School effectiveness is obtained by averaging the students' value-added measures and it is used for ranking schools. This paper compares alternative school ranking obtained either using value-added measures or row reading scores for each school. The work analyzes to what extent the two rankings differ and finds out the schools that rank significantly below or above the average. In fact, in order to suitably evaluate the students' results and to correctly implement effective educational policies it is necessary to considerer both a starting reading level of the pupils and their motivation. We find that the school ranking obtained with the row reading scores are significantly different from those obtained using value-added measures. Differently from our expectation, technical schools and even some vocational schools show better value-added results compared with academic schools. This work suggests that in order to properly evaluate the impact that the single school has on students it is necessary to take into account a value-added measure to avoid not valid ranking.

Analysis of the Corporate Philosophy of Japanese Companies Using Text Mining

Akihiro Saito, University of Kitakyushu
Hiroshi Takeda, University of Kitakyushu

Abstract

Because of progress of globalization and business environment which requires that a company should achieve social responsibility, in these days, research of corporate philosophy attracts attention. However, empirical research of the corporate philosophy of Japanese companies is not progressing. In this research, we analyzed the corporate philosophy of 983 Japanese companies using text mining. From this analysis, we discovered the word which Japanese companies use for corporate philosophy frequently. Moreover, we discovered the feature of the corporate philosophy of each category of business from some methods of analyzing categorical data. From the results of these analysis, we discuss about how Japanese companies use corporate philosophy.

An Algorithm for Document Reconstruction

Gabriella Schoier, University of Trieste

Abstract

The need of reconstructing documents which have been destroyed by means of a shredder may arise in different fields. In a computer-based reconstruction, the pieces are described by numerical features which represent the visual content of the strips. We propose a solution to the problem of document reconstruction by considering an algorithm of variable reduction which can be used as a starting point for a following cluster analysis.

Prediction Model for Ratio of Students Passing New National Bar Examination for Each Law School

Kumiko Shiina, The National Center for University Entrance Examinations
Ken-ichiro Komaki, The National Center for University Entrance Examinations
Katsumi Sakurai, The National Center for University Entrance Examinations
Taketoshi Sugisawa, Niigata University

Abstract

In Japan, graduate schools specialized in training of legal professionals (hereafter referred to as 'law schools') were established as a new legal training system in 2004. Those who have completed the course at law schools are awarded the qualification for candidacy for the new national bar examination since 2005. As an admission examination to law schools, all applicants are given a test which evaluates applicants' abilities to judge, think, analyze and express themselves, whether or not they have already studied law. In order to evaluate such abilities demanded as preconditions for study at law schools, the National Admission Test for Law Schools (hereafter referred to as NATLaS) was administered to applicants between 2003 and 2010. The standard training term is three years, and students who have already studied law are allowed to complete the course in a shorter period (two years). The new national bar examination is designed to judge whether candidates are equipped with abilities to analyze matters, to think logically, and to interpret and apply laws by requiring candidates to demonstrate how to solve problems, how to prevent conflicts, how to design plans, and the like. Such abilities are common to the abilities evaluated by the NATLaS. It is expected that the ratio of students passing the bar exam to those who have completed the course at each law school (hereafter referred to as 'success ratio') relates to the NATLaS scores of enrolled students. A prediction model for the success ratio from the mean of the NATLaS score for enrolled students was constructed. The predictive validity of the NATLaS was examined on the basis of the estimated values of parameters of the model. One of the parameters was used to convert the raw NATLaS score into a value on the scale of the model. The parameter corresponds to the NATLaS score, which has a value of success ratio 0.5. The ratio of students passing the bar exam was assumed to increase continuously as a logistic function of the NATLaS score. The logistic function has a parameter, which indicates the sharpness of the increase. The values of the parameters for groups of students of each enrollment year were estimated by the method of least squares based on observed values of the mean score of the NATLaS of each law school and the success ratio. As for the NATLaS score parameter, the estimated value of this parameter for the students on the three-year course was higher than that for those on the two-year course. Thus, passing the bar exam was more difficult for students on a three-year course than for those on the two-year course. The estimated values of the sharpness parameter were positive for all enrollment years. It corresponds to that the students with higher NATLaS scores indicate a higher success ratio at the bar exam. The tendency of the two parameters suggests predictive validity of the NATLaS.

Analysis of National Center Test for University Admissions by Latent Rank Theory

Kojiro Shojima, The National Center for University Entrance Examinations

Abstract

Latent rank theory is a test standardization theory for evaluating examinees' academic status in an ordinal scale. In this study, a result of analyzing an English language test by an LRT model for applying to dichotomous (true/false) data. In addition, two polytomous LRT model: nominal model for nominal categories data and graded model for ordinal categories data (Likert-type data) is also introduced. Finally, a local dependence LRT model is referred.

Analysis of Double Bipartite Data by Asymmetric von Mises Scaling

Kojiro Shojima, The National Center for University Entrance Examinations

Abstract

In this study, we propose a technique for analyzing an asymmetric double bipartite data by asymmetric von Mises scaling (AMISESCAL), where the bipartite data is a rectangular matrix in which the intersection of the sets of the row and column elements is Ø and AMISESCAL is an asymmetric multidimensional scaling that uses a technique derived from directional statistics.

Algorithmic Imputation Techniques for Missing Data: Performance Comparisons and Development Perspectives

Nadia Solaro, University of Milano-Bicocca
Alessandro Barbiero, University of Milan
Giancarlo Manzi, University of Milan
Pier Alda Ferrari, University of Milan

Abstract

Missing data have always represented a hard-to-solve problem for researchers from every field. Unsuitable solutions could heavily affect the reliability of statistical results and lead to wrong conclusions. The increasing availability of data often characterized by missing values has paved the way for the development of new alternative methods for handling missing data. Among the others, three recent proposals seem most promising in terms of successfully detecting the actual values of missing data: (i) Stekhoven and Buehlmann's method (missForest), which uses an iterative imputation technique based on a random forest; (ii) Josse, Pages, and Husson's method (missMDA), which uses an iterative algorithm based on principal component analysis; (iii) Ferrari, Annoni, Barbiero, and Manzi's method (ForImp), which is based on alternating nonlinear principal component analysis and the nearest neighbour imputation method. Users are often faced with the dilemma of having to choose among many different imputation techniques, and, moreover, one is not always confident about the adequacy of the imputation exercise. It would therefore be important to find, for every different situation and possible missing data distribution, the best algorithm to be used, as well as to detect turning points where a given technique should be abandoned in favour of others. In this paper, potential solutions to these issues are explored. MissForest and missMDA along with an extended version of ForImp are compared on the basis of the results of a simulation study. Complete data matrices are generated according to different multivariate distributions (multivariate normal, Azzalini's skew-normal and multivariate exponential power distributions) and under different settings given by the number of variables, correlation or association structures, parameters related to skewness or kurtosis. Missing data in different percentages are then generated through a MCAR mechanism. The three methods are then compared through their RMSE with respect to the original complete data matrix. Advice to detect possible breakpoints to switch from one technique to another will be given according to the different experimental settings considered.

A Family of Methods for Asymmetric Nearest Neighbor Clustering

Satoshi Takumi, University of Tsukuba
Sadaaki Miyamoto, University of Tsukuba

Abstract

We study a family of linkage methods for asymmetric nearest neighbor clustering. Given an asymmetric graph with weights of edges, we can consider weak and strong transitive closures as extensions of the single linkage. When the concept of `dense points' is introduced, we have an asymmetric version of Wishart's mode analysis. Moreover, an asymmetric version of DBSCAN can be developed as a similar method to the asymmetric Wishart's mode analysis. Algorithms are developed and numerical examples are shown.

A Criterion-based PLS Approach to Multi-block and Multi-group Data Analysis

Michel Tenenhaus, HEC Paris

Abstract

On the one hand, multi-block data analysis concerns the analysis of several sets of variables (blocks) observed on the same set of individuals. On the other hand, multi-group data analysis concerns the analysis of one set of variables observed on a set of individuals taking into account a group-structure at the level of the individuals. Two types of partition of an *individuals × variables* data matrix **X** are then defined:

- In the multi-block framework, the column partition $\mathbf{X} = \left[\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_J\right]$ is considered. In this case, each block $\mathbf{X}_j$ is an $n \times p_j$ data matrix and represents a set of p_j variables observed on a set of n individuals. The number and the nature of the variables differ from one block to another but the individuals must be the same across blocks.
- In the multi-group framework, the row partition $\mathbf{X} = \left[\mathbf{X}_1^t, \ldots, \mathbf{X}_i^t, \ldots, \mathbf{X}_I^t\right]$ is considered. In this framework, the same set of variables is observed on different groups of observations. Each group $\mathbf{X}_i$ is an $n_i \times p$ data matrix and represents a set of p variables observed on a set of n_i individuals. The number of observations of each block could differ from one block to another.

Many methods exist for multi-block and multi-group data analysis. Regularized Generalized Canonical Correlation Analysis (RGCCA) was proposed in Tenenhaus & Tenenhaus (2011) and appeared to include an amazing large number of criterion-based multi-block data analysis methods as particular cases. In this paper, we intend to extend RGCCA so that it can also be a unifying tool for multi-group data analysis.

References

Tenenhaus, A. & Tenenhaus, M. (2011): Regularized generalized canonical correlation analysis. Psychometrika, vol. 76, n° 2, pp. 257-284.

Two-mode Three-way Asymmetric MDS with the Generalized Hyperellipse Model

Yoshikazu Terada, Osaka University
Hiroshi Yadohisa, Doshisha University

Abstract

Asymmetric MDS models for representing the asymmetric structure among the objects have been proposed by many researchers. Okada (1990) proposed the hyperellipse model for asymmetric MDS. In this model, each object is represented by a hyperellipse in Euclidean space. For two-mode three-way asymmetric proximities, Okada and Imaizumi (2002) proposed the weighted Euclidean model based on this model. Terada and Yadohisa (2012) extended the hyperellipse model to represent each object as a more general hyperellipse, which could be a rotated hyperellipse. In this paper, we propose a two-mode three-way asymmetric MDS by extending the generalized hyperellipse model of Terada and Yadohisa (2012). That is, we extend the weighted Euclidean model of Okada and Imaizumi (2002) to the generalized Euclidean model based on the generalized hyperellipse model for two-mode three-way asymmetric dissimilarities. In this model, the individual spaces are obtained first by rotating after that by stretching the common space, and each object is represented by a general hyperellipse in the both spaces.

Association Study and Comprehended Hierarchical Structures on DNA Markers

Makoto Tomita, Tokyo Medical and Dental University
Koji Kurihara, Okayama University

Abstract

In the statistical genetics, we have focused `haplotype' which is the major polymorphic marker on DNA data, therefore it has been estimated its relative frequency, examined it structure, and used for an association study since about 10 years ago. We considered that Echelon analysis can be applied not only to identify LD blocks but also to select tagging SNPs, and introduce these comprehensive methods with numerical examples (see Tomita et al., 2008; and Tomita et al., 2011 for details).

Robustness and Stability Analysis of Factor PD-clustering on Large Social Datasets

Cristina Tortora, Stazione Zoologica Anton Dohrn of Naples
Marina Marino, Federico II University of Naples

Abstract

Factorial clustering methods have been proposed in order to cluster large datasets, where large is referred to the number of variables. These methods are based on two main steps: linear transformation of original variables and clustering on transformed variables. The two steps are iterated until the convergence is reached. In literature this approach firstly appeared in 1994 as simple two-step clustering, the two-steps are iterated only once. However the two-steps optimize different criteria and the first factorial step can mask the real clustering structure. The main difference in factorial clustering methods is that the two steps optimize a common criterion. Among them Factorial Probabilistic Distance Clustering (FPDC) has been proposed. The two main steps are: a Tucker 3 decomposition of the distance matrix and a Probabilistic Distance (PD) clustering on the Tucker 3 factors. PD-clustering is an iterative, distribution free, probabilistic, clustering method; it assigns units to cluster according to their probability of belonging to a cluster, under the constraint that the ratio between the probability and the distance of each point to any cluster center is a constant. It has been demonstrated that Tucker 3 decomposition optimizes the same criterion of PD-clustering, consequently, the projection of units on Tucker 3 factors emphasize the real clustering structure. The method gives stable and interesting results dealing with different type of datasets. A simulation study has tested the performance of the method dealing with: outliers, different number of elements in each cluster and variance changing among clusters. The aim of this paper is to apply FPDC on behavioural and social datasets of large dimensions, to obtain homogeneous and well-separated clusters of individuals. The scope is to evaluate the stability and the robustness of the method dealing with real large datasets. Stability of results is referred to the invariance of results in each iteration of the method. Robustness is referred to the sensitivity of the method to errors in data, which can be outliers or deviations from assumptions. These characteristics of the method are evaluated using Bootstrap resampling.

Investigation of Mixture-modeling and Some Non-hierarchical Clustering Methods to Detect Heterogeneous Developmental Trajectories in Longitudinal Data Analysis

Satoshi Usami, University of Tokio

Abstract

In behavioral sciences, the mixture-modeling clustering approach has been effectively used to extract unobserved clusters for stratification and clustering of individuals. In the present research, the performance of finite mixture models and some non-hierarchical clustering methods are compared to estimate the number of clusters through a simple simulation study. Methods for non-hierarchical clustering include the K-means and the proposed restricted clustering method. The restricted non-hierarchical clustering method assumes a massive (but finite) number of clusters at first iteration. Statistically meaningful clusters are extracted through repetition, restricting the mixture proportion and degree of separation among clusters, thus allowing us to evaluate the possible patterns of change flexibly in accordance with the research purpose and knowledge that researchers have, with no emphasis on the specification of the number of clusters beforehand.

On Joint Dimension Reduction and Clustering

Michel van de Velden, Erasmus University Rotterdam
Alfonso Iodice D'Enza, University of Cassino and Southern Lazio
Francesco Palumbo, Federico II University of Naples

Abstract

There exist several methods for clustering high-dimensional data. One popular approach is to use a two-step procedure. In the first step, a dimension reduction technique is used to reduce the dimensionality of the data. In the second step, cluster analysis is applied to the data in the reduced space. This method may be referred to as the tandem approach. An important drawback of this method is that the dimension reduction may distort or hide the cluster structure. Vichi and Kiers (2001) showed in a simulation study how the tandem approach may fail to retrieve the clusters in low dimensional space. In the context of categorical data, Van Buuren and Heiser (1989) proposed a method in which object scores are restricted using cluster memberships. Hwang, Dillon and Takane (2006) proposed a joined multiple correspondence analysis (MCA) and K-means clustering method that uses user-specified weights for the clustering and dimension reduction parts. For binary data, Iodice D'Enza and Palumbo (2012) recently proposed a new method which they refer to as iterative factorial clustering for binary data (i-FCB). In this paper we review existing joint dimension reduction and clustering methods in a unified framework that facilitates comparison. Moreover, we propose a reformulation of Hwang et al.'s method that leads to some interesting new mathematical insights and opens the way to several adaptations. Also, we propose a reformulation of the i-FCB method and its extension to categorical data.

Treelet Analysis and Independent Component Analysis of Milan Mobile-Network Data: Investigating Population Mobility and Behavior

Simone Vantini, Politecnico di Milano
Valeria Vitelli, Électricité de France, École Centrale Paris
Paolo Zanini, Politecnico di Milano

Abstract

The metropolitan area of Milan is the fifth largest in Europe, it includes the provinces of Milan, Bergamo, Como, Lecco, Lodi, Monza-e-Brianza, Novara, Pavia, and Varese, and it is characterized by a high concentration of both working and residential activities. The OECD identifies housing, transport, and congestion as the bottlenecks to the future growth of the area. Indeed they seem to badly affect the well-being of the city from many perspectives: pollution (Milan is the second most air-polluted city in Europe), economy (the difficulty in managing people and goods mobility is estimated to damp more than 4% the output of the area), and society (while the population of the metropolitan area is growing, the population of the municipality of Milan is decreasing). In recent years a lot of initiatives have been undertaken to address these problems. The Green Move project, which the present research is part of, is among these initiatives. Green Move is an interdisciplinary research project financed by Regione Lombardia involving different research groups at the Politecnico di Milano and regarding the development of a vehicle sharing system based on the concept of "little, electric and shared vehicles". Our contribution to the project is to provide information about people mobility to find optimal places where to locate the docking stations. To this aim, we exploit the Telecom Italia database. In this database the metropolitan area of Milan is parted according to a uniform lattice of several thousands of sites of size 232m × 309m. In each site, the average number of mobile phones simultaneously using the network for calling at a given time is provided every 15 minutes for 14 days. At a first approximation, this quantity can be considered proportional to the number of active people in that site at that time, and thus able to provide information about people mobility. The data set at hand can be genuinely considered an instance of spatially-dependent functional data, because of the high within-unit sample size and the very high signal-to-noise ratio. Indeed two different functional data analyses are performed and the corresponding results thoroughly compared: a Treelet analysis and an Independent Component Analysis. Both analyses aim at describing the data set by a time-varying linear combination of a reduced number of time-invariant basis surfaces. Time-varying coefficients represent basic profiles of mobile network use recurring over the map, while site-evaluations of the surfaces measure the contribution of the corresponding basic profile to the signal observed in that site. We expect spatial dependence to be non-stationary and non-isotropic, being strongly related to the underlying road network. It is thus treated in a non-parametric way which relies on several random Voronoi tessellations of the investigated area, providing several sets of local representatives that are separately analyzed and then bagged together in a final aggregation step. The results of the two analyses are complementary disclosing both common and analysis-specific patterns amenable of a clear spatiotemporal interpretation such as: average population density, working and residential activities, universities, shopping, leisure, and morning/evening road/railways commuting.

Partitioning Asymmetric Dissimilarity Data

Donatella Vicari, Sapienza University of Rome

Abstract

When clustering asymmetric data, only the average amounts are often considered by assuming that the asymmetry is due to noise. But when the asymmetry is structural, as typically may happen for exchange flows, migration data or confusion data, this may strongly affect the partitioning because the directions of the exchanges are ignored. The clustering model proposed here relies on the decomposition of the asymmetric dissimilarity matrix into symmetric and skew-symmetric effects both decomposed in within and between cluster effects. The classification structures used here are generally based on two different partitions of the objects fitted to the symmetric and the skew-symmetric part of the data, respectively. The special case of the general model is also presented where the partition fits jointly both of them allowing for clusters of objects similar with respect to the average amounts and directions of the data. Parsimonious models are presented which allow for effective and simple graphical representations of the results.

Multimode Clustering

Maurizio Vichi, Sapienza University of Rome

Abstract

Multimode clustering is a field of research concerning the simultaneous clustering of the different dimensions of a multimode, multiway data matrix (array). For two way data, bi-clustering, co-clustering and two-mode clustering are the terms used for such joint classification of the rows and columns of a two-way, two-mode data matrix. These methodologies determine, for each mode: a single-partition; a multi-partition; a hierarchy. Furthermore, each cluster can be synthesized by the same functional, (mean vector, linear combination) symmetrically for each mode or by a different functional (asymmetrical treatment) for each mode. For example, clusters of rows and columns of a two way data matrix can be both synthesized by mean vectors, i.e., centroids for rows and columns. However, mean vectors for row clusters and linear combinations for column clusters may be more appropriate for a classical two-way two mode data matrix (*objects* × *variables*), because multivariate objects are generally synthesized by mean vectors, while observed variables are generally synthesized by non-observable latent variables (components, factors), i.e., linear combinations.

In this paper a new general model for multi-partitioning, by allowing a symmetrical or asymmetrical synthesis of the two-mode (units and variables) or three-mode (units, variables and occasions) data matrix is presented. For two way data it includes, as special cases: the double k-means, that identifies a single partition for each mode of the data; the Clustering and Disjoint Principal Component for identifying a partition of the objects and a partition of the variables together with a component of maximal variance for each class of the partition of the variables; and two-way Multimode Partitioning for specifying more partitions of one mode, conditionally to the partition of the other one. These three model are extended also the three-way data. The performance of such generalized multimode k-means has been tested by both a simulation study and an some applications on real data.

Similarity Measure and Clustering Algorithm for Candlestick Valued Data

Yoji Yamashita, Doshisha University
Hiroshi Yadohisa, Doshisha University

Abstract

Loss reduction in a stock market portfolio is possible by combining assets moving in opposite directions. Wittman (2002) proposed a method to grasp stocks moving in opposite directions by calculating the dissimilarity measure from the closing price. However, there are other values that can be used for this purpose, such as, the lowest value and highest values. Therefore, we propose a new dissimilarity using a candlestick. A candlestick represents four prices; opening, highest, lowest and closing. In this paper, we define candlestick-valued data from a given candlestick, and from opposite direction movements between candlesticks. In this paper, we propose a method for calculating the dissimilarity measure considering the opposite movements. We also propose a clustering algorithm on the basis of the dissimilarity measure.

Reference:

T. Wittman. Time-Series Clustering and Association Analysis o f Financial Data, CS 8980 Project, 2002.

Research Institute Analysis Based on Collaborative Research Activities

Nagayoshi Yamashita, Japan Society for the promotion of science
Yuan Sun, National Institute of Informatics
Kei Kurakawa, National Institute of Informatics
Yasumasa Baba, The Institute of Statistical Mathematics

Abstract

We propose a methodology for constructing research institute networks based on collaborative research activities. Relationships of research institutes in these networks are evaluated by the number of collaborative researches extracted from grant application data of a Japanese funding agency. Grant application data has been used for investigating trends of science. This format is organized and a hundred thousand applications from variety areas ranging from Humanities and Social Sciences to Medicine are proposed and each application has attributes, for example research fields, keywords, research institutes of each researcher. Each grant application data has collaborative researches: a principle investigator, several co-investigators and their belonging research institutes. Using these relationships, research institute networks are constructed. In this paper, we analyze collaborative researchers as these networks and show effectiveness of our proposed methodology by evaluating correlation between collaborative researches and research outcomes.

Y

Assessing Cooperation in Open Systems: an Empirical Test in Healthcare

Paola Zappa, University of Milano-Bicocca

Abstract

The diffusion of Web 2.0 technologies has made available a large amount of data to Social Network scholars. Appropriately treated, these data allow to study phenomena which previously required extensive survey data collection and could be affected by the related problems. This paper deals with the treatment of emails panel data and with their longitudinal modelling within a Social Network Analysis framework. Specifying Stochastic Actor-Oriented Models, we assess the capability of this approach to detect the social dynamics underlying online interaction. Application is on a virtual community of practice of Italian oncologists who collaborate in resolving diagnoses. Using repository and field data, we reconstruct a network, whose nodes are the clinicians and ties the emails sent to each other, and then model their cooperation behaviour longitudinally. The suitability of this approach is shown and its advantages and drawbacks are discussed.

A Multidisciplinary Approach to Validate a Friendship Psychometric Scale: a Study in an Italian Sample of Students

Emma Zavarrone, IULM University
Sandra De Francisci Epifani, IULM University

Abstract

This paper aims at creating and validating a Physiological Emotions Friendship Scale (PhEFS) combining psychometric characteristics, social network analysis and emotional/physical internal states. The friendship construct has been long investigated by these macro-disciplines. In the psychometric literature the scales on friendship have reached good results in terms of validity and reliability measurement (e.g. Friendship Qualities Scale, Friendship Intimate Scale) as well as in the social network literature the different aspects of friendship (influence, connection or cohesion analysis) have been largely explored (Scott, 2000; Wellman, 2008; Snijders, 2009). However, it has been observed that the psychological approach permits to detect the items related to the latent construct but not the interpersonal aspects, whereas the social network analysis does not allow to estimate the latent construct, but only the different aspects of friendship. An emerging research area, based on the neurophysiological discipline, states the possibility of encoding emotional reactions to stimuli: Rainville et al. (2006) assume that patterns of peripheral physiological activity are correlated to different emotions. The psychophysiology focuses on two emotional components: the qualitative component expressed by means of the word used to describe the emotion (valence), and the quantitative component defined by means of words of magnitude of the emotion (arousal) (Picard, 1995). The classification of specific arousal, can be detected by electronic devices (FlexComp Infiniti™ System, Thought), and offers a substantial explanation of human attitudes towards friendship, although the affective computing cannot identify the latent constructs. The key point could be represented from the study of the neurophysiological signals detected on emotion stimulation linked both the latent construct and the interaction among individuals. In order to take advantage from multidisciplinary efforts, we are in process of creating an experimental biofeedback scaling, able to detect the items related to the friendship construct. We assume that a set of items, exploring friendship construct, administered in a network of students, can generate specific emotions and internal states variation. In this perspective, items become the stimuli aimed at measuring the construct in the network. As a result, detected neurophysiological parameters will confirm the presence/absence of the friendship construct while parameters variations will allow us to determine the bond friendship intensity within the network itself. This experiment will be carried out with the support of Iulm Behavior & Brain Lab on four random sample of students. Each group will fill in the specific questionnaire under neurophysiologic stimuli detection. In particular, a set of sensors, connected to a PC and to each student, will record physiological activations caused by questionnaire submission. The collected data refer to neurophysiological parameters (e.g. skin conductance, heart rate variability, electromyography, breathing frequency, EEG, indices derived from eye-tracking), psychometric measures on friendship construct and network information. The use of multiway Anova approach on neurophysiological parameters, psychometric responses (treatment 1) and network measures (treatment 2) will allow to discriminate items and subjects. Finally, PhEFS will be created on the basis of the most items-related neurophysiological parameters and will be validated on a new sample of students.

Comparative Analysis on Cultural Changes in East Asian Values Based on Survey Data

Yuejun Zheng, Doshisha University

Abstract

Culture is a general term for behavioral patterns or lifestyles mastered, shared and transmitted by people who constitute a society, and refers to the system of customs and behaviors which humans have created over many years. It is important to mutually understand and respect the customs, behavioral patterns and social values of one's own or other's through cross-cultural exchanges, not to force one's own culture upon other individuals or groups. Chinese morality is the spirit of the whole culture, whose core has come from the philosophy of Confucius with 2500 years or more of history. Around the beginning of the fifth century A.D., it began to spread to Oriental countries, such as Korea and Japan, and has had remarkable impacts on various aspects of their cultures, politics, and morality. However, since the Meiji restoration in Japan, with various social changes in each country, the contents of culture and morality have been transformed into unique ones in spite of absorbing other cultures Nowadays, these three countries have some aspects of culture and values which are too different to be mutually understood, though they located in the same Oriental cultural sphere. In other words, in the East Asian cultural sphere, Confucianism has had major impacts historically and played the role of a cultural principal axis. On the other hand, it is also a fact that the changes of time and the modernization of each country have tinged them with their peculiar features, leading to many cultural differences. With the progress of globalization, the international exchanges in various fields such as culture and academics as well as politics and economics, and the mutual understanding of different cultures become more and more necessary for the East Asian countries. In order to encourage further mutual exchanges in the modern East Asia, it is important to objectively analyze how much understanding of the peculiar cultures and values are shared by China, Japan and South Korea, and what the differences among them are. Therefore, it will be essential to analyze the social survey data properly collected based on the statistical science in each country, and to extract the information which can serve as an aid to promote mutual exchanges. This paper, focusing on the traditional and modern cultures in East Asia, quantitatively analyzes the similarities and differences in the modern East Asia based on several survey data sets collected in the past decade, featuring such themes of interest to modern people as national culture, honor, purpose of raising a child, religion, qualities of leader, nationalism and international orientation etc.. The author mainly gropes for the information which serves as aid to create the mutual understanding and the cooperative relationship in the East Asia cultural sphere through a sequence of analyses.

Teenagers and Mass Media in Messina: an Estimation of the Cumulative Proportional Odds Model

Agata Zirilli, University of Messina
Angela Alibrandi, University of Messina
Massimiliano Giacalone, University of Calabria
Rosaria Barbara Ucchino, University of Messina

Abstract

The aim of this paper is to analyze the relationship between teenagers and media, in order to try to better understand the habits and to conduct analysis on social interactions with young people. From the methodological point of view, we estimated ordinal logistic regression model, to test the dependence of mass media influence and inspiration with respect to the preference of kind of television programs, the time spent on TV and on computer and the most used social network.

On the Stability of Blockmodels Based on Structural and Regular Equivalence Given Measurement Errors

Anja Žnidaršič, University of Maribor
Anuška Ferligoj, University of Ljubljana
Patrick Doreian, University of Pittsburgh, University of Ljubljana

Abstract

A widely used technique for finding such structural patterns is generalized blockmodeling. The result of a blockmodeling procedure is a partition of actors and an image matrix with established block types, where a block determines the relation between two clusters of actors. Most often, the data are assumed to be measured correctly. Social network data are gathered by using different techniques with surveys prominent among them. Yet survey research designs are known to produce data with measurement errors. Our concern with measurement error in social network data focuses on mis-specified network boundaries due to actor non-response and errors due to item non-response. Here we consider random measurement error on ties and its impact on the resulting blockmodels. Following Holland and Leinhardt (1973), we assume measurement errors take the form of missing ties (no choices are recorded in the network for relational ties that exist) and extra ties (choices are recorded in the network for which there are no corresponding relational ties). To study the impact of measurement error on identified blockmodel structures, we: i) start with a whole (or known) network and its blockmodel; ii) impose different amounts of random measurement errors on its ties; iii) establish the blockmodel of the resulting network; and iv) compare both blockmodel structures. The comparisons are made using the Adjusted Rand Index which measures the differences between two obtained partitions and the proportion of incorrectly identified blocks. We confine our attention to structural equivalence and regular equivalence. We use real networks and simulated networks to assess the extent to which a known (true) blockmodel is reproduced in the blockmodels of the data with missing data and different treatments of measurement error. We present results as to which type of equivalence produces 'correct blockmodels' in the face of measurement error and an assessment of the amount of measurement error that permits the delineation of acceptable blockmodels and how much error prevents this.

List of authors