



International Classification Conference 2011

Incorporating the
Symposium on Bourdieu and Geometric Analysis of Data

Conference venue
**School of Medical & Biological Sciences
University of St Andrews
1411 - 2011**



Sponsored by



British Classification Society

11 - 15 July, 2011

International Classification Conference

Incorporating the
Symposium on Bourdieu and Geometric Analysis of Data

Hosted by
**School of Management
University of St Andrews**

Conference venue
**School of Medical & Biological Sciences
University of St Andrews**

11 – 15 July, 2011

Sponsored by



British Classification Society

credits

ICC Scientific Programme Committee

Fionn Murtagh • Nema Dean • David Wishart

Session organizers **Christian Hennig • François Husson • Brian Mac Namee
Fred Stentiford**

Bourdieu Symposium Committee

Mike Grenfell • Brigitte Le Roux

Led by **Philippe Bonnet • Brigitte Le Roux • Frédéric Lebaron**

Local Organizing Committee

David Wishart • Martin Dowling • Paloma Salmeron Planells

Violoncello Concert

Gabriela and Patrick Bradley

Conference Book of Abstracts

Pedro Contreras

ICC-2011 Programme

Incorporating the

Symposium on Bourdieu and Geometric Analysis of Data

Monday 11th July 2011

Pre-Conference Workshop

10.00–18.00 Bourdieu and Geometric Data Analysis Workshop

(Organisers: Brigitte Le Roux, Michael Grenfell and Frédéric Lebaron)

Tuesday 12th July 2011

Session 1, Tuesday 12 July 2011

09:00–09:15 David Wishart (President, British Classification Society) Welcome

09:15–10:10 Adrian Raftery (University of Washington), “Model-Based Classification and Clustering in High Dimensions”

09:30 Social Programme – Coach tour to Edinburgh departs from Medical & Biological Sciences (NB: Entry to Edinburgh Castle is £13 payable by cash or credit card at the Castle). Return to New Hall at 18:00.

10:10–10:30 Coffee/tea

Session 2a, Tuesday 12 July 2011

10:30–12:00 Mixture Analysis, Model-Based Methods

Anderlucci Laura (University of Bologna) and Christian Hennig (University College London), “How well does mixture clustering do according to distance-based criteria?”

Cláudia Silvestre, Margarida Cardoso and Mário Figueiredo (IPL, ISCTE-IUL, IST, Portugal), “Simultaneously selecting categorical features and the number of clusters in model-based clustering”

Jan Schepers (Maastricht University), “Mixture semiparametric regression”

Markus Eichhoff and Claus Weihs (TU Dortmund), “Detection of musical instruments in intervals and chords”

Session 2b, Tuesday 12 July 2011

10:30–12:00 Learning with a Human-in-the-Loop (Organiser: Brian Mac Namee)

Sarah Jane Delany, Alexey Tarasov, Charlie Cullen and John Snel (Dublin Institute of Technology), “Using crowdsourcing in the rating of emotional speech assets”

Simon Fischer (Rapid Miner, Dortmund), “Lost in transformation: Data Mining beyond CRISP”

Nicolas Cebron, Fabian Richter and Rainer Lienhart (University of Augsburg), “Towards learning with complementary labels”

Brian Mac Namee (Dublin Institute of Technology), “Show me a picture: Visualising interactive machine learning algorithms”

Session 2c, Tuesday 12 July 2011

10:30–12:00 Symposium 1: Bourdieu and Geometric Analysis of Data: Introduction & Cultural Distinctions (Organisers: Brigitte Le Roux, Michael Grenfell and Frédéric Lebaron)

Michael Grenfell (Trinity College, University of Dublin), “Introduction: Bourdieu and the geometric analysis of data”

Henk Roose (Ghent University), Laurie Hanquiet (University of York) and Mike Savage (University of York), “Beauty and the eye of the beholder: aesthetic dispositions and museum audiences in Flanders (Belgium)”

Johs Hjellbrekke and Olav Korsnes (University of Bergen, Norway), “Cultural distinctions: a geometric data analysis”

12:00-13:00 Lunch (New Hall Restaurant – Location 4 Map Ref D3/E3) (All participants)

Session 3a, Tuesday 12 July 2011

13:30-15:00, Coffee/tea, 15:30-17:30 Aspects of Computational Perception in Vision and Text (Organiser: Fred Stentiford)

Virginia Fernandez Arguedas and Ebroul Izquierdo (Queen Mary, University of London), “Behaviour-based surveillance video analysis”

Fred Stentiford and Ade Bamidele (University College London and Nokia UK), Text detection in natural scenes using cliques of interest points for mobile visual search”

Dalia Chakrabarty (Warwick University), “Supervised learning of shapes of galaxy clusters”

Fred Stentiford (University College London), “A method of pattern recognition applied to the Poggendorff illusion”

Robert Shilston and Fred Stentiford (University College London), “Auto-focus algorithm selection: A methodology for comparing blur perception between observers”

Daniel Baier and Ines Daniel (TU Cottbus), “Using image clustering algorithms for marketing purposes”

Session 3b, Tuesday 12 July 2011

13:30-15:00, Coffee/tea, 15:30-17:30 Symposium 2: Bourdieu and Geometric Analysis of Data: Cultural Distinctions (Organisers: Brigitte Le Roux, Michael Grenfell and Frédéric Lebaron)

Daniel Laurison (University of California, Berkley), “Positioning political producers: the field of American campaign professionals”

Ylva Bergstrom and Tobias Dalberg (Uppsala University), “Political space of young Swedish upper secondary pupils

Frédéric Lebaron (CURAPP/CNRS, Université de Picardie), “Is there a universal model of central bankers?”

Bourdieu Symposium Round Table

18:00-19:00 Dinner at New Hall (Residents)

19:30-21:30 Conference Reception (School of Medical & Biological Sciences)

Wednesday 13th July 2011

Session 4, Wednesday 13 July 2011

09:00-10:00 Keith van Rijsbergen (University of Glasgow), "Revisiting clustering and classification in information retrieval"

10:00-10:30 Coffee/tea

Session 5a, Wednesday 13 July 2011

10:30-12:00 K-Means and Optimization

David Wishart (University of St Andrews), "Optimization in k-means analysis: some new developments"

Renato Cordeiro de Amorim and Boris Mirkin (Birkbeck University of London), "Minkowski metric k-means for feature weighting"

Daniel Aloise and Arthur Araújo (Universidade Federal do Rio Grande do Norte), "A k-means inspired heuristic for microdata protection"

Session 5b, Wednesday 13 July 2011

10:30-12:00 Supervised Learning and Diagnostic Modelling in Clinical Research

Berthold Lausen (University of Essex), "Meta-analysis methods for gene expression profiles"

Sivarit Sultornsanee, Ibrahim Zeid, and Sagar Kamarthi (Northeastern University, Boston), "A recurrence quantification analysis based approach to lung sounds classification"

Casper Albers (University of Groningen) and John Gower (Open University), "Between group metrics, and their use in canonical variate analysis"

Maria Paula Brito (Universidade do Porto) and A. Pedro Duarte Silva (Universidade Católica Portuguesa at Porto), "Parametric discriminant analysis of interval data"

Session 5c, Wednesday 13 July 2011

10:30-12:00 Symposium 3: Bourdieu and Geometric Analysis of Data: Power and Economy (Organisers: Brigitte Le Roux, Michael Grenfell and Frédéric Lebaron)

Carla DiGiorgio (University of Prince Edward Island, Canada), "Applying Bourdieusian social theory to the interaction between identity, power and inclusive practice in a minority language school"

Jacque Widin (University of Technology, Sydney), "Mapping an illegitimate field: power relations in international education"

Monica Becue-Bertaut (Universitat Politecnica de Catalunya), Belchin Kostov (CAP Les Corts – Hospitał Clinic) and Anne Morin (Irisa-Universite de Rennes), "Structure and vocabulary flow in chronological corpora: contribution of correspondence analysis"

Bourdieu Symposium Round Table

12:00-13:00 Lunch (New Hall)

13:15-18:00 Social Programme – Coach tour to Glamis Castle (Paloma Planells)

13:30-15:00 Social Programme – Historical walking tour of St Andrews (Dr Bill Knox)

Both tours depart from Medical & Biological Sciences – please do not be late

18:00-19:00 Dinner at New Hall (Residents)

Social Programme Wednesday 13 July 2011 (Medical & Biological Sciences)

19:30-21:30 David Wishart (University of St Andrews), "We'll tak' a cup o' kindness yet - the story of Scotch whisky"
Illustrated talk and tasting with 40 fine single malt whiskies

Thursday 14th July 2011

Session 6a, Thursday 14 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 **Machine Learning with Confidence, Conformal Prediction: Session in Honour of Chris Wallace and Ray Solomonoff**

Grace Solomonoff (Oxbridge Research), “Ray Solomonoff and the new kind of probability”

Valentina Fedorova, Ilia Nouretdinov and Alex Gammerman (Computer Learning Research Centre, Royal Holloway, University of London), “Conformal predictors and testing exchangeability assumption”

Ilia Nouretdinov, Matilde Santos and Alex Gammerman (Royal Holloway, University of London), “Non-conformity measures in multi-class prediction”

S. González, J. Vega, A. Pereira and I. Pastor (Asociación EURATOM/CIEMAT para Fusion, Madrid), “Advanced analysis and classification of images using conformal predictors”

Olga Ivina, Ilia Nouretdinov (University of Gerona) and Alex Gammerman (Royal Holloway, University of London), “Valid predictions with confidence estimation in air pollution problem”

Session 6b, Thursday 14 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 **Correspondence Analysis and Applications (Organiser: François Husson)**

M. Ono, Manabu Kano and Toshio Sugiman (University College London), “Understanding the immune system by Correspondence Analysis”

Belchin Kostov, Mónica Bécue-Bertaut, Jérôme Pagès, Marine Cadoret, Jordi Torrens and Pilar Urpi (Les Corts, Hospital Clinic; Universitat Politècnica de Catalunya; Agrocampus-Ouest, Rennes; Freixenet S.A.), “Verbalisation tasks in Hall test sessions”

Zhiyuan Luo (Royal Holloway, University of London), “Methods for reliable classification of network traffic”

Julie Josse, Marie Chavent, Benoît Liquet and François Husson (Agrocampus-Ouest, Rennes), “Handling missing values with regularized iterative multiple Correspondence Analysis”

Marie Chavent, Vanessa Kuentz, Benoît Liquet and Jérôme Saracco (Université de Bordeaux), “Clustering of a variable via the PCAMIX method”

Solène Bienaise and Mireille Gettler Summa (Université de Paris-Dauphine), “Validation of trajectories on factorial planes after a tandem clustering approach”

Session 6c, Thursday 14 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 **Symposium 4: Bourdieu and Geometric Analysis of Data: Education and Language (Organisers: Brigitte Le Roux, Michael Grenfell and Frédéric Lebaron)**

Andrea Tribess (CURAPP, France) “In which social context will working class students obtain a university diploma”

Jo Watson (University of Southampton), “Widening participation in higher education: capital that counts”

Jan Thorhauge Frederiksen (Roskilde University, Denmark), “Trawling for students: how do educational institutions regulate for students”

Bourdieu Symposium Round Table

12:00-13:00 **Lunch (New Hall)**

Session 7a, Thursday 14 July 2011

13:30-15:00, Coffee/tea, 15:30-16:30 Spatial Clustering, Geoscience and Environmental Science

Vladimir Batagelj, Anuska Ferligoj and Patrick Doreian (University of Ljubljana), “The nine nations of North America”

Adam Butler (Biomathematics and Statistics Scotland) and Ellie Owen (Royal Society for the Protection of Birds), “The use of GPS tracking data to infer foraging behaviour in seabirds”

Ayale Daher (Université de Bretagne Sud), “Clustering constrained spatially”

Simona Korenjak Černe, Vladimir.Batagelj and Natasa Kejzar (University of Ljubljana), “Clustering data described with discrete distributions: an application on population pyramids”

Session 7b, Thursday 14 July 2011

13:30-15:00, Coffee/tea, 15:30-17:00 The Philosophy of Clustering (Organiser: Christian Hennig)

Stephen Reid (University of Stellenbosch), “Leading indicators of currency crises in emerging economies”

Joachim M Buhmann (ETH Zurich), “Selecting clustering models by maximizing approximation capacity”

Nicholas T. Longford (SNTL and UPF, Barcelona), “Clusters defined by distinct correlation structures”

Christian Hennig (University College London), “How to find the best cluster analysis method (for social stratification based on mixed data)?”

Pascal Pr ea and Monique Rolbert (Laboratoire d'Informatique Fondamentale de Marseille), “Distinguishing and classifying from n-ary relations”

Session 7c, Thursday 14 July 2011

14:00-15:00, Coffee/tea, 15:30-17:00 Symposium 5: Bourdieu and Geometric Analysis of Data: Methodological Points and Applications (Organisers: Brigitte Le Roux, Michael Grenfell, Fr ed eric Lebaron)

Fionn Murtagh (Royal Holloway, University of London), Adam Ganz (Royal Holloway, University of London) and Joe Reddington (University of Aberdeen), “A metric and ultrametric platform for analysis of semantics and change”

Cheryl Hardy (Liverpool John Moores University), “Re-presenting the social world: Bourdieu and graphic illustrations of fields”

Vincent Berry (EXPERICE, University Paris XIII), Manuel Boutet (CESAER, INFRA, AgroSup Dijon) and Samuel Coavoux (Centre Max Weber, ENS Lyon), “Bourdieu in the playfield: rethinking video game studies”

17:15-18:15 British Classification Society AGM

17:15-18:15 International Federation of Classification Societies Council Meeting

18:00-19:00 Dinner at New Hall (Residents, not attending the ICC-2011 Conference Banquet)

19:00-23:00 ICC-2011 Conference Banquet, St Salvator’s College (location 44 map ref K2)

19:00 Welcome by the Highland Piper, Duncan Soutar, St Salvator’s College Cloisters

19:30 ICC-2011 Conference Dinner, Lower College Hall

21:30 Violoncello Concert, Gabriela and Patrick Bradley (Karlsruhe), Upper College Hall

Friday 15th July 2011

Session 8a, Friday 15 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 Evaluation and Visualization of Clustering and Classification – Education, Psychology, Economics, Politics

John C Gower (Open University), Niel J Le Roux (Stellenbosch University) and Sugnet Lubbe (University of Cape Town), “The canonical analysis of distance”

Maria José P. C. Amorim and Margarida G. M. S. Cardoso (ISEL; ISCTE-IUL Portugal), “Measuring the agreement between partitions: the use of thresholds values”

Tim Brennan and Markus Brietenbach (Northpointe Institute, Colorado), “A taxonomic analysis of female pathways to crime: From qualitative descriptions to replicated quantitative patterns”

Rebecca Nugent and Nema Dean (Carnegie Mellon University; Glasgow University), “Clustering students by their skill set profiles on the unit hypercube”

Session 8b Friday 15 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 High Dimensional Bioinformatics and Hierarchical Data and Computation

Charles Bouveyron (Université Paris 1 Panthéon-Sorbonne), "Parsimonious and sparse Gaussian models for the clustering of high-dimensional data"

Nema Dean and Adrian E. Raftery (Glasgow University; University of Washington), “Variable selection for latent class analysis applied to HapMap SNP data”

Ka Yee Yeung (University of Washington; Carnegie Mellon University), “Application of Bayesian Model Averaging to the construction of regulatory networks”

Patrick Erik Bradley and Andreas Christian Braun (Karlsruhe Institute of Technology), “Baire distance and feature ranking in classification”

Fionn Murtagh and Pedro Contreras (Science Foundation Ireland and Royal Holloway University of London; Thinking Safe), “Linear time hierarchical clustering using the Baire metric, and alternative perspectives: p-adic clustering, generalized ultrametrics, hashing and more”

Session 8c, Friday 15 July 2011

09:00-10:00, Coffee/tea, 10:30-12:00 Symposium 6: Bourdieu and Geometric Analysis of Data: Methodological points and applications Cont’d (Organisers: Brigitte Le Roux, Michael Grenfell, Frédéric Lebaron)

Brigitte LeRoux (MAP5, Université Paris Descartes and CEVIPOF/CNRS, Sciences-Po), “Class specific analysis”

Frédéric Lebaron (CURAPP/CNRS, Université de Picardie), Philippe Bonnet (Laboratoire de Psychologie et Neuropsychologie Cognitive, FRE 3292, CNRS et Paris-Descartes) and Brigitte Le Roux (MAP5, Université Paris Descartes and CEVIPOF/CNRS, Sciences-Po), “Geometric data analysis of French cultural practices.

Final Bourdieu Symposium Plenary

12:00-12:15 Closing Session

12:15-13:00 Lunch (New Hall)

14:00-15:00 Farewell Party (Medical & Biological Sciences)

Poster Session (11-15 July 2011)

Jules J.S. de Tibeiro and Duncan J. Murdoch (Université de Moncton, University of Western Ontario), “Correspondence analysis with incomplete paired data using Bayesian imputation”

Bourdieu and Geometric Data Analysis Workshop

Philippe Bonnet (Université Paris Descartes), Brigitte Le Roux (Université Paris Descartes & CEVIPOF/Sciences-Po),
Frédéric Lebaron (Université de Picardie)

The objective of this workshop is to introduce researchers and PhD students to Geometric Data Analysis (GDA). In this approach to Multivariate Statistics, data sets are represented as clouds of points and the interpretation is based on these clouds.

In this workshop we will address the following issues:

- GDA methods will be reviewed with a short historical overview and their uses in social sciences. We will introduce Multiple Correspondence Analysis (MCA) as one of the three main paradigms of GDA. Then we will present our leading example and review some methodological issues.
- MCA will be applied to the leading example using SPAD software. First, MCA principles will be presented, starting with the definition of the distance between individuals followed by a discussion of the properties of the clouds of individuals and categories. Then principal axes, contributions and the different steps in the analysis of a data set will be reviewed. Finally, we will present extensive analysis of a case study.

References

- Le Roux B. & Rouanet H. (2010) *Multiple Correspondence Analysis*, Series Quantitative Applications in the Social Sciences, n° 163, SAGE: Thousand Oaks (Contents: <http://www.mi.parisdescartes.fr/~lerb/livres/MCA/Overview.html>).
- Le Roux B. & Rouanet H. (2004) *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*, Kluwer, Dordrecht.
- Le Roux B., Rouanet H., Savage M., Warde A. (2008) Class and Cultural Division in the UK, *Sociology* 2008; 42; 1049-1071. (<http://soc.sagepub.com/cgi/reprint/42/6/1049>) (abstract: <http://www.mi.parisdescartes.fr/~lerb/publications/1050.pdf>).
- Hjellbrekke J., Le Roux B. Korsnes O., Lebaron F., Lennart R., Rouanet H. (2007) The Norwegian field of Power Anno 2000, *European Societies* , 9(2), p.245-273.

Contacts

Philippe.Bonnet@parisdescartes.fr

Brigitte.LeRoux@mi.parisdescartes.fr

flebaron@yahoo.fr

SPAD Software

A free copy of SPAD software will be available for each participant.

www.coheris.com - for the version that includes recent developments

Between group metrics, and their use in canonical variate analysis

Albers C. (1) and Gower J. (2)

(1) Department of Psychometrics & Statistical Methods, University of Groningen, the Netherlands

(2) Department of Mathematics & Statistics, The Open University, Milton Keynes UK

Canonical (variate) analysis deals with measurements on p variables for n samples in k groups. In canonical variate analysis we aim to represent the data matrix X in an r -dimensional space. This can be done by optimising the ratio form, or by constrained optimisation. In classical situations, these methods coincide. When there are more variables than samples, they generally do not. A method to generalise canonical variate analysis to this case will be presented. Also, as well as the usual canonical means in the range-space of the within-groups dispersion matrix, canonical means may be defined in its null space. In the range space we have the usual Mahalanobis metric; in the null space explicit expressions are given and interpreted for a new metric.

References

J.C. Gower and C.J. Albers (2011), Between-group metrics, submitted

C.J. Albers and J.C. Gower (2011), Canonical Analysis: Ranks, Ratios and Fits, in preparation

A k -means inspired heuristic for microdata protection

Aloise D. and Araújo A.

Department of Computer Engineering and Automation
Universidade Federal do Rio Grande do Norte, Campus Universitário s/n,
Natal-RN, Brazil, 59072-970

The increasing amount of data generated from filling online questionnaires and forms in our society demands thinking about how public and private organizations would have access to all this information. The biggest challenge when disclosing private data is to share information contained in databases while protecting people from being individually identified. Microaggregation is a family of methods for statistical disclosure control. The principle of microaggregation is that confidentiality rules permit the publication of individual records if they are partitioned into groups of g or more data, where none is more representative than the others in the same group. The application of such rules leads to replacing individual values by those computed from small groups (microaggregates), before data publication. The microaggregation procedure should be developed to reduce as much as possible the information loss caused by this replacement process, so that valuable and accurate information could still be extracted from the microaggregated data. This work proposes a microaggregation method based on the popular k -means heuristic. Computational experiments show that the proposed method finds the best results for a number of benchmark instances in the literature.

How well does mixture clustering do according to distance-based criteria?

Anderlucci L. (1) and Hennig, C. (2)

(1) University of Bologna (Italy)

(2) University College London (UK)

There are different ways to cluster categorical data in the literature. The choice among them is strongly related to the aim of the researcher, if we do not take into account time and economical constraints. Main approaches for clustering are, among others, model-based and distance-based methods: the former assume that objects belonging to the same class are similar in the sense that their observed values come from the same probability distribution, whose parameters are unknown and need to be estimated; the latter evaluates distances among objects by a defined dissimilarity measure and, basing on it, allocates units to the closest group.

As clustering is defined as the classification of similar objects into groups, we wonder if observations coming from the same probability distribution are as similar as they would be if a distance-based method were applied and how good are those latter methods in finding the 'true' model-based clustering, if it exists. In order to answer, two approaches, namely a latent class model (mixture of multinomial distributions) and "partition around medoids"-clustering, are evaluated and compared by several indexes in a comprehensive simulation study.

References

Kaufman, L. and Rouseeuw, P.J. (1990), *Finding Groups in Data*, Wiley, New York.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* 61, 215-231.

Hennig, C. and Liao, T. (2010) Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Research report no. 308, Department of Statistical Science, UCL.
<http://www.ucl.ac.uk/statistics/research/reports>

Measuring the agreement between partitions: the use of thresholds values

Amorim M. J. P. C. (1) and Cardoso M. G. M. S. (2)

(1) Mathematics Department, ISEL, Portugal
mjamorim@deq.isel.pt

(2) Department of Quantitative Methods, ISCTE-IUL, Portugal
margarida.cardoso@iscte.pt

The property of stability is often considered when evaluating the quality of a clustering solution. In particular, the solution's reproducibility in diverse data sets drawn from the same source may be considered as an indicator of stability. In order to measure stability one can use indices of agreement (IA) between the alternative partitions obtained from the diverse data sets. In fact, there are countless IA which can be used for this end. However, there has been few investment concerning the determination of IA thresholds values which can help deciding how much agreement is enough to derive stability (the well known Hubert and Arabie adjusted Rand index is an important exception).

In the present work, we propose using simulated IA values, corresponding to cross-classification tables generated under the hypothesis of restricted independence (table with fixed marginal totals), to obtain IA thresholds. The Rand index, Mutual Information and the Variation of Information are used as IA examples. The R software is used to implement the proposed approach. Four simulated data sets (Gaussian, mixture model based, with different degrees of separation) are used to obtain experimental results. The stability of alternative clustering solutions provided by different clustering algorithms (K-Means and EM type) is evaluated and discussed using a cross-validation approach. In addition, the agreement with the real partition is also discussed.

Using image clustering algorithms for marketing purposes

Baier D. and Daniel I.

Institute of Business Administration and Economics, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany

The popularity of classification methods in marketing research can be easily seen from their detailed description in nearly all available marketing and marketing research textbooks as well as their heavy usage in statistical software packages like SAS or SPSS, both by practitioners and academic researchers. So, e.g., in market segmentation, hierarchical or partitioning algorithms are applied to sociodemographic, psychographic, preference, or usage descriptions of potential customers in order to derive homogeneous groups of them. However, recently, through the success of social network services (e.g. facebook, flickr) available resources for this purpose have extended. So, in social networks, potential customers meet and share their interests, activities, and opinions with other network members by providing, e.g., self-describing profiles, contact lists, videos, music, or images. Mostly, these images and videos reflect their individual activities, opinions, and interests, e.g., w.r.t. their living conditions as well as their spare time and holiday experiences or their favorite stars, songs, films, or pictures. Consequently these different media types could be used as a basis for market segmentation where personal activities, interests, and opinions play a major role. In this paper we discuss, how the application of clustering algorithms to such uploaded image collections can be used for deriving market segments. Software prototypes are discussed and applied.

References

- Baier, D., Daniel, I. (2011): Image Clustering for Marketing Purposes. To appear in *Studies in Classification, Data Analysis, and Knowledge Organization*.
- Figueiredo, M. (2007): Semi-Supervised Clustering: Application to Image Segmentation. In: *Studies in Classification, Data Analysis, and Knowledge Organization*, 34, 39-50.
- Van House, N.A. (2009): Collocated Photo Sharing, Story-Telling, and the Performance of Self. In: *International Journal of Human-Computer Studies*, 67(12), 1073-1086.

The nine nations of North America

Batagelj V. (1), Ferligoj A. (2), and Doreian P. (3)

(1) University of Ljubljana, Faculty of Mathematics and Physics, vladimir.batagelj@fmf.uni-lj.si

(2) University of Ljubljana, Faculty of Social Sciences, anuska.ferligoj@fdv.uni-lj.si

(3) University of Pittsburgh, pitpat@pitt.edu

In his book *The Nine Nations of North America* [1, 3], written in 1981, Joel Garreau suggests that North America can be divided into nine regions or “nations” (The Empty Quarter, Quebec, New England, The Foundry, Dixie, The Islands, Mexamerica, Breadbasket and Ecotopia) that represent the true nature of North American society.

In the paper we present an attempt to check the Garreau’s classification for the area of United States on the basis of the data on US counties [2]. Besides standard clustering methods also clustering with relational constraints and different network analysis methods are used.

References

[1] Joel Garreau. (1981). *The Nine Nations of North America* Houghton Mifflin. ISBN 0395291240.

[2] USA Counties data files, U.S. Census Bureau. <http://www.census.gov/support/USACdataDownloads.html>

[3] Wikipedia: *The Nine Nations of North America*. http://en.wikipedia.org/wiki/The_Nine_Nations_of_North_America

Structure and vocabulary flow in chronological corpora. Contribution of correspondence analysis.

Bécue-Bertaut M. (1), Kostov B. (2), Morin A. (3)

(1) Universitat Politècnica de Catalunya

(2) CAP Les Corts - Hospital Clínic

(3) Irista-Université de Rennes-I

When analyzing a chronological corpus through statistical methods, our aims are both to uncover its structure and identify the vocabulary flow. Correspondence analysis brings an important contribution to these purposes. However, both require different segmentation levels, that is, different corpus_segments×words tables. Structure arises when the segments are long enough so that the ever important local variability is smoothed. On the contrary, the second task requires operating at sentence/paragraph level (more generally, context unit) to identify the different metakeys - or groups of words that, together, constitute a topic- which leads to build the paragraph×words table. The joint analysis of these two tables leads to replace the topics into the chronological structure and then to study their flow and recurrences. As a complement, chronological clustering allows for dating the changes and identify the isolated passages or singletons. We will present applications to such different corpora as a closing speech for the prosecution in a trial for murder, a collection of sentences pronounced by the Spanish Supreme Court during the so-called legal transition (1979-1996) and the series of articles published in Le Monde in the international section from 1987 to 2003.

Political space of young Swedish upper secondary pupils

Bergström Y. and Dalberg T.

Department of Education, Uppsala University

The overall aim of our project is to study differences in political opinions, values and participation within groups of young pupils of upper secondary education from two different municipalities in Sweden.

After decades of silence, the research field on political socialization is undergoing a somewhat revitalization. Of much concern is young citizens' decline in joining political parties and involvement in political communities. These trends of decline in political party membership and party activism, have been accompanied by another development; the rise of relatively new forms of political participation, having their origins in wider social and technological changes. In many Western democracies, so called, consumer participation has become an increasingly important feature of politics, changing the political landscape; offering new arenas addressing alternative political issues. In much research, however, youth is considered more or less as a homogenous group, differentiated only by age, whereas the comparisons mainly focus on differences between adolescents, youth and adults. Using geometric data analysis our paper focus on differences in political opinions, values and participation within a group of young upper secondary pupils and how differences in political opinion and values correspond to different educational positions.

The data set stems from a survey carried out in 2008 and 2009 among third grade pupils in upper secondary schools in two different municipalities – the university town of Uppsala (n=1093) and the mining districts (n=400) – in Sweden.

Bourdieu in the playfield. Rethinking video game studies

Berry V. (1), Boutet M. (2), and Coavoux S. (3)

(1) EXPERICE – Université Paris XIII

(2) CESAER – INRA – AgroSup Dijon

(3) Centre Max Weber – ENS Lyon

Discussions among video games scholars are largely focused on two ideas: Huizinga's magic circle and De Certeau's tactics. In a nutshell, the regularities of practices are seen as a consequence of the designers' creation of rules, and their meaningfulness as an outcome of the local agencies of players - for instance in online games "guilds" and "raids".

Pierre Bourdieu's thought and the tools he used open new directions. First, it allows us to take into account the various ways the game is played, and more to the point, how those practices are structured in the form of a field. Indeed, these practices are ordered. Furthermore, our investigations demonstrate that players do not arbitrarily choose a practice, but rely on tastes built - outside the game in the social space, as well as during the game experience.

This paper relies on two quantitative surveys of the players of the online role-playing game World of Warcraft, and uses multiple correspondence analysis to show how this community of online video game players is structured by a field-specific logic, namely the amount of game-specific capital players own. It then relates the hierarchical structure of the game to the social properties of the players and shows that there is no strict homology between the game social space and the space of social positions, although the two are not altogether unrelated. It appears necessary, then, to take into account game-related variables in order to understand the social structure of the game.

Validation of trajectories on factorial plans after a tandem clustering approach

Bienaise S. and Gettler Summa M.

Université Paris Dauphine, CEREMADE
1 pl. du Ml. de Lattre de Tassigny - Paris - France
bienaise@ceremade.dauphine.fr
summa@ceremade.dauphine.fr

In order to find clusters in a set a multiple multidimensional time series, a tandem clustering approach can be performed [1]: first a principal component analysis (for continuous data) or a multiple correspondence analysis (for categorical or binary data), then a clustering phase. By projecting the clusters on the factorial mappings, one may find patterns that can be interpreted according to the interpretation of the factorial axes.

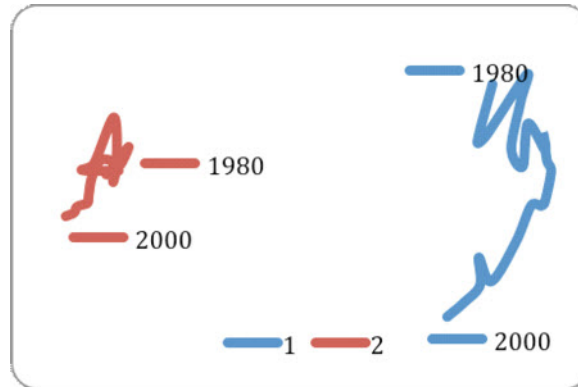


Figure 1

Figure 1 shows two trajectories that have been projected on the two first principal components (first axis is horizontal, second one orthogonal to the first one). Each of them belongs to a different cluster [2]. It could be concluded that in class 1 evolution are less important than in class 2 along the second axis. It could also be observed that class 2 crosses the plan from the upper (positive ordinates) right (positive abscissas) quarter, to the lower (negative ordinates) right quarter, crossing the first axis. This paper aims at validating such conclusions (e.g.: is the difference between the ranges of the two classes changes significant?) in the framework of inferential factorial analysis [3]. Bootstrap approaches are taken into consideration and some statistical tests are proposed for further improvements.

References

- [1] Arabie, P., Hubert, L.J., and de Soete, G. (1996). Clustering and Classification. Word Scientific
- [2] Clustering Trajectories of a Three-Way Longitudinal Data Set, in M. Gettler Summa et al., Statistical Learning and Data science, Chapman & Hall, in press.
- [3] Le Roux, B., and Rouanet, H. (2004), Geometrical Data Analysis, Kluwer Academic Publishers

Baire distance and feature ranking in classification

Bradley P. E. and Braun A. C.

Karlsruhe Institute of Technology (KIT)

Various methods of employing the Baire metric in the context of supervised and unsupervised classification are described. This metric is used to obtain a fast feature ranking or, vice versa, an ordering of features is shown to define a natural Baire distance. This is applied to supervised and unsupervised classification, where a ranking is empirically found to give the most characteristic features of a cluster, and the inverse ranking yields the most discriminative features in supervised classification. Baire metrics are parametrised by their basis, and any Baire distance yields a fast hierarchical clustering algorithm. The question of optimal orderings can be solved by Dijkstra's algorithm on the graph defined by the power set of features. The performance in terms of complexity and quality of results are discussed, as well as the dependence of results on the basis.

A taxonomic analysis of female pathways to crime: From qualitative descriptions to replicated quantitative patterns

Brennan T. and Brietenbach M.

Northpointe Institute, Golden, Colorado

In the last decade the classification of women's pathways to crime has been dominated by qualitative research. This has identified several typified female "pathways" to criminal behavior and proposed new gender-specific features for women's offending. We report on a quantitative classification study of a sample of women prisoners ($N = 718$) assessed on the new feminist features and on standard psycho-social, personality and criminal history features. Using several cluster analysis methods a hierarchical classification model ranging from 4 to 8 clusters at successive K levels was developed. Cross method convergence was examined using bagged K-Means, Semi-supervised clustering and Ward's method. Cross sample stability was examined at each K level using McIntyre-Blashfield's validation approach. Class separation was examined using the Minimax Probability Machine. New case assignment was tested using the Support Vector Machine method. The identified female pathways are described and contrasted to the prior qualitative pathways research.

Parametric discriminant analysis of Interval data

Brito P. (1) and Duarte Silva A. P. (2)

(1) Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal

(2) Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal

Symbolic Data Analysis (Billard and Diday (2005); Diday and Noirhomme (2008); Noirhomme and Brito (2011), provided a framework that allows taking directly into account variability and/or uncertainty which might be associated to each single observed “individual”. New variable types have been introduced, which may assume multiple, possibly weighted, values. We focus on the analysis of interval data, i.e., where elements are described by variables whose values are intervals of IR.

In Duarte Silva and Brito (2005), we have studied and compared different methods for linear discriminant analysis of interval data. These, however, rely on non-parametric approaches, therefore not allowing for inferential studies.

Parametric inference methodologies based on probabilistic models for interval-valued variables are developed in Brito, Duarte Silva (2011, in press) where each interval is represented by its midpoint and log-range, for which Normal and Skew-Normal (Azzalini and Dalla Valle (1996)) distributions are assumed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, which are represented by five different possible configurations.

In this work we adopt this parametric modeling in linear and quadratic discriminant analysis of data described by interval-valued variables. The performance of the new approach is then compared with previous proposals.

References

Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution. *Biometrika* 83 (4), 715-726.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.

Brito, P. and Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, (in press).

Duarte Silva, A.P. and Brito, P. (2006). Linear discriminant analysis for interval data, *Computational Statistics* 21, (2), 289-308.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, Chichester.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining*, (in press).

Parsimonious and sparse Gaussian models for the clustering of high-dimensional data

Bouveyron C.

Maître de Conférences en Mathématiques Appliquées
Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne

Clustering of high-dimensional data has become a recurrent problem in scientific fields such as Biology, Chemometry or Image analysis. The most popular approaches for clustering are based on the mixture model which unfortunately suffers from the curse of dimensionality in high-dimensional spaces. After a short overview of usual ways for dealing with high-dimensional data, the talk will focus on the use of parsimonious and sparse Gaussian models for the clustering of high-dimensional data. In particular, subspace classification methods will be studied in detail and compared to their challengers. The special case " $n \ll p$ " and variable selection will be also considered since it concerns nowadays several scientific applications such as biology or chemometry.

References

- C. Bouveyron, S. Girard and C. Schmid, High-Dimensional Data Clustering, Computational Statistics and Data Analysis, vol. 52 (1), pp. 502-519, 2007.
- C. Bouveyron, G. Celeux and S. Girard, Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA, Preprint HAL n°00440372, Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, 2010.
- C. Bouveyron and C. Brunet, Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, Statistics and Computing, in press, 2011.

Selecting clustering models by maximizing approximation capacity

Buhmann J. M.

ETH Zurich

Exploratory data analysis by data clustering requires (i) to specify a suitable clustering principle and (ii) to choose an appropriate number of clusters. We propose maximum approximation capacity as a principle for model selection and model-order selection. Our principle ranks competing clustering cost functions according to their ability to extract context sensitive information from noisy data with respect to a hypothesis class. Sets of approximate solutions serve as a basis for a communication protocol. We select the clustering model that maximizes the so-called approximation capacity. We demonstrate our framework on three different problem domains: model-order selection for mixtures of Gaussians, learnability of hard problems in high dimensions, and clustering binary data in a security application.

This is joint work with M. H. Chehreghani, A. P. Streich, M. Frank.

The use of GPS tracking data to infer foraging behaviour in seabirds

Butler A. (1), and Owen E. (2)

(1) Biomathematics and Statistics Scotland

(2) Royal Society for the Protection of Birds

The populations of many seabird species are in decline. In order to conserve these species it is important to understand the areas in which birds forage (feed), and the environmental conditions that are associated with foraging. Direct data on the foraging behavior of seabirds are extremely difficult to collect, but GPS tracking data - which monitor the location of individual birds at regular intervals, and thereby provide information on direction and speed - are now becoming widely available. Direction and speed are, in turn, known to be closely related to behavior - foraging behavior tends, for example, to be associated with relatively rapid changes in direction and relatively low speeds - although the precise nature of this relationship is not well understood. The use of GPS tracking data to draw inferences about the behavioral state of each bird at each location and point in time is a classification problem, which is complicated by the existence of missing data and by the presence of spatial and temporal autocorrelation within tracking data (Aarts et al., 2008). We outline possible approaches for drawing such inferences, including Bayesian state space modelling (Jonsen et al., 2003; Patterson et al., 2007), and highlight some of the methodological challenges that are involved in using GPS tracking data for this purpose. The ideas and methodologies are illustrated using data from the Future of the Atlantic Marine Environment (FAME) project.

References

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M. and Matthiopoulos, J. (2008) Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, 31, 140-160.

Jonsen, I. D., Myers, R. A. and Flemming, J. M. (2003) Meta-analysis of animal movement using state-space models. *Ecology*, 84(11), 3055-3063.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O. and Matthiopoulos, J. (2007) State space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2), 88-94.

Towards learning with complementary labels

Cebron N., Richter F., and Lienhart R.

Multimedia Computing Lab, University of
Augsburg, Universitätsstr. 6a, 86156 Augsburg, Germany

The goal in supervised classification is to infer a function from training examples that maps from the examples to the class labels. While a plethora of algorithms for supervised classification has been developed, only a few works deviate from this classical setting.

Finding the correct class label for an example can be a tedious process especially when there are a large number of classes. In the work of [1], it has been shown that the human error rate and the time needed to find the correct label grows with the number of classes; at the same time the user distress increases. In some situations, it might not even be possible for the human expert to determine the correct class label out of many possible class labels. In a normal classification setting, we would have to ignore this example.

In this work, we want to introduce a new paradigm in supervised classification: we do not obtain the label information itself, but the labels of the classes that this example does not belong. We call these labels \bar{C} (not-labels¹). It is easier to obtain this kind of information, especially in classification problems with many classes. But there is of course a loss of information induced from this setting, as we might observe only partial not-label information for an example. We will present a classification framework that incorporates not-labels seamlessly with normal class labels. We evaluate this framework with different amounts of supervision with not-labels.

References

[1] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback. In CVPR, pages 2995-3002. IEEE, 2010.

Supervised Learning of Shapes of Galaxy Clusters

Chakrabarty D.

Department of Statistics, University of Warwick, Coventry CV4 7AL

Shape classification is regularly handled within Computer Vision, Artificial Intelligence, etc. However, the basic difference between the classification of morphology in n -D systems, using n -D images, and the counterpart in astronomy is that images of clusters of galaxies is in 2-D while the classification of their 3-D shapes is sought, along with an estimate of the unmeasurable orientation angles i_1, i_2 along which the density is projected, to form the image I . Here $I := \{ I(x_1^{(i)}, x_2^{(i)}) \}_{i=1}^N$, i.e. the image I comprises N pixels, with 2-D radiation density recorded at each grid point on the X_1 - X_2 plane. Also, I is piecewise continuous and *assumed* piecewise smooth, is non-negative and bounded in $[0, \infty)$. In general, $I(x_1, x_2) = \int_{-\infty}^{\infty} \rho(x_1, x_2, x_3) h(x_1^{(i)}, x_2^{(i)}) dx_3$, where $h(x_1, x_2)$ is the known blurring function, convolution and $h(x_1^{(i)}, x_2^{(i)})$ is the maximal extent along the X_3 -axis, for any point $(x_1^{(i)}, x_2^{(i)})$ on the image.

Had there existed a measurable cluster parameter S , the shape dependence of which was a known functional $f[S(x, i_1, i_2)]$, of the morphology parameter S , such that $S = f(S)$, we could have inverted this equation to estimate morphology. As no such S exist, {we resort to solving the forward problem} of estimating a cluster property - profile of S along X -axis, i.e. S , assuming S - the used proxy for which is $h(x_1^{(i)}, x_2^{(i)})$ - and then classifying. In our model, the global system geometry is *assumed* to be triaxial as inspired by gravitational theory.

The classifier is achieved via intercomparison of estimated profiles that emanate from inverting I under 4 benchmark choices of the function $h(x_1, x_2)$, abbreviated as $h_k, k = 1, \dots, 4$ (Chakrabarty, de Filippis & Russell, 2008). We record 3 parameters of I, h_k , namely $\{ \mu, \sigma, \tau \}$. The estimation of I, h_k is performed with the Bayesian non-parametric inversion methodology DOPING (Chakrabarty 2010), $k = 1, \dots, 4$. We then compare with $h_l, l = 1, \dots, 4, j = 1, \dots, 4$ and the result is -1, 0 or 1 depending on if $\mu_j < \mu_l, =$ or $>$ respectively, where these relations are inferred within the inferred 90% credible regions within which the learnt I lies, which also have the measurement uncertainties incorporated. Thus, the comparison results in the discrete, trichotomous comparison parameter $c_{kl},$ where $m=1, \dots, 6$, since 6 independent pairs of k, l can be chosen from a set of 4. Thus, the matrix of c_{kl} is a 3×6 matrix - the "comparison matrix". A fiduciary 3-D system morphology maps to a unique c_{kl} -matrix. The mapping is established as injective, where S is the space of fiduciary triaxial 3-D shapes and A that of the c_{kl} -matrices. In another viewpoint, a given system is a point in an 18-D shape manifold; we train our classifier to identify the location of this point with the cluster shape.

We train our classifier on data drawn from simulated galaxy clusters and then apply the classifier to test data comprising sample of I from 24 observed galaxy clusters. The resolution of a system as prolate, oblate or triaxial is reported and broad classification of the range of i is also achieved. Subsequently, we invoke the ancillary measurements that offer $\max \{ h_k(x_1^{(i)}, x_2^{(i)}) \}_{i=1}^N$, relative to the extent along a principal axes of the (elliptical) image. This measurement is very noisy, but taking the noise into account, we perform interval estimates for shape parameters and orientations.

Dalia Chakrabarty, "Non-parametric Deprojection of Surface Brightness Profiles of Galaxies", 2010, *Astronomy & Astrophysics*, 510, 45.

Chakrabarty, D., de Filippis, B. & Russell, H., 2008, "Shapes and Inclinations of Galaxy Clusters from Deprojection Uncertainties", *Astronomy & Astrophysics*, 487, 75.

Clustering of a variables via the PCAMIX method

Chavent M. (1,2), Kuentz V. (3), Liquet B. (4), and Saracco J. (1,2)

- (1) IMB, University of Bordeaux, France
- (2) CQFD team, INRIA Bordeaux Sud-Ouest, France
- (3) CEMAGREF, UR ADBX, France
- (4) ISPED, University of Bordeaux, France

Clustering of variables is as a way to arrange variables into homogeneous clusters i.e. groups of variables which are strongly related to each other and thus bring the same information. Clustering of variables can then be useful for dimension reduction and variable selection. Several specific methods have been developed for the clustering of numerical variables. However concerning qualitative variables or mixtures of quantitative and qualitative variables, much less methods have been proposed.

The ClustOfVar package has then been developed specifically for that purpose. The homogeneity criterion of a cluster is the sum of correlation ratios (for qualitative variables) and squared correlations (for quantitative variables) to a synthetic variable, summarizing “as good as possible” the variables in the cluster. This synthetic variable is the first principal component obtained with the PCAMIX method. Two algorithms for the clustering of variables are proposed: iterative relocation algorithm, ascendant hierarchical clustering. We also propose a bootstrap approach to determine suitable numbers of clusters. The proposed methodologies are illustrated on real datasets.

Minkowski metric k -means for feature weighting

Cordeiro de Amorim R.,(1) and Mirkin B.(1,2)

(1) Department of Computer Science and Information Systems, Birkbeck University of London, Malet Street, London WC1E 7HX, UK

(2) Department of Data Analysis and Machine Intelligence, Higher School of Economics, Moscow, RF
E-mail: {renato, mirkin}@dcs.bbk.ac.uk

This paper represents another step in overcoming a drawback of K-Means, its lack of defense against noisy features, by using feature weights in the criterion. We extend the Weighted K-Means method by Huang et al. to the corresponding Minkowski metric for measuring distances. Under Minkowski metric the feature weights become intuitively appealing feature rescaling factors in a conventional K-Means criterion. To see how this can be used in overcoming two other drawbacks of K-Means: the lack of advice on the number of clusters and initial setting, we adapt Mirkin's method of anomalous clusters to initialize K-Means. We experimentally validate our method on datasets from Irvine repository and generated sets of Gaussian clusters, both as they are and with additional uniform random noise features, and demonstrate its competitiveness in comparison with other K-Means based feature weighting algorithms.

Clustering constrained spatially

Daher A.

Lab-STICC UMR 3192 / UBS

Spatial clustering is intended to form classes that contain sites with similar characteristics and are geographically neighbors with some spatial coherence.

Classic clustering algorithms are used to partition the data space, the classification obtained is generally very fragmented. To avoid this fragmentation, the spatial structure of the data must be considered.

The article proposes a model that explicitly takes into account the spatial structuring intra-class and/or inter-class, as well as introduces a spatial clustering algorithm.

This approach is validated on real data.

Correspondance analysis with incomplete paired data using bayesian imputation

de Tibeiro J. J. S. (1) and Murdoch D. J. (2)

(1) Université de Moncton, Moncton, N.-B., Canada

(2) The University of Western Ontario, London, ON, Canada

In this paper we consider the analysis of incomplete tables using Correspondence Analysis (CA). We focus on a dataset concerning congenital heart disease (Fraser and Hunter 1975), in which the data forms a square table, but only a symmetrized version of the off-diagonal entries was reported. We use Markov chain Monte Carlo (MCMC) on a hierarchical Bayes model to estimate the underlying rates, and use CA to study the relationships in the completed table.

References

Benzécri, J. P. (1992). Correspondence Analysis Handbook. Marcel Dekker.

de Tibeiro, J. J. S. and Murdoch, D. J. (2010). "Correspondence Analysis with Incomplete Paired Data using Bayesian Imputation." *Bayesian Analysis*, 5:3 1-14.

de Tibeiro, J. J. S. (1996). "Sur les traits associés par paires: malformations cardiaques congénitales chez des enfants ayant même parents." *Les Cahiers de l'Analyse des Données*, 21: 45-52.

Fraser, F. C. and Hunter, A. D. W. (1975). "Etiologic Relations Among Categories of Congenital Heart Malformations." *The American Journal of Cardiology*, 36: 793-796.

MacGibbon, B. (1983). "A Log-linear Model of a Paired Sibling Study." In Chaubey, Y. and Dwivedi, T. D. (eds.), *Proceedings of Statistics '81 Canada Conference*, 193-197.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 Users Manual*. MRC Biostatistics Unit, Cambridge.

Variable selection for latent class analysis applied to HapMap SNP data

Dean N. (1) and Raftery A. E. (2)

(1) School of Mathematics & Statistics, University of Glasgow;
(2) Department of Statistics, University of Washington

We propose a method for selecting variables in latent class analysis, which is the most common model-based clustering method for discrete data. The method assesses a variable's usefulness for clustering by comparing two models, given the clustering variables already selected. In one model the variable contributes information about cluster allocation beyond that contained in the already selected variables, and in the other model it does not. A headlong search algorithm is used to explore the model space and select clustering variables. In simulated datasets we found that the method selected the correct clustering variables, and also led to improvements in classification performance and in accuracy of the choice of the number of classes. In a dataset from the International HapMap Project consisting of 639 single nucleotide polymorphisms (SNPs) from 210 members of different groups, our method discovered the same group structure with a much smaller number of SNPs.

Using Crowdsourcing in the rating of Emotional Speech Assets

Delany S. J., Tarasov A., Cullen C., and Snel J.

Digital Media Centre, Dublin Institute of Technology, Dublin, Ireland.

The automatic recognition of emotion from speech recordings uses supervised machine learning techniques which requires labeled training data in order to operate effectively. The performance of these supervised learning techniques depends on the quality of the training data and therefore on the quality of the labels or ratings. In this domain the ratings are typically estimated from the subjective opinion of a small number of experts.

Recently with the availability of crowdsourcing services it has become inexpensive to acquire labels from multiple non-expert annotators which has led to the use of crowdsourcing for labelling training data in a variety of domains. It can be argued that emotional expertise does not necessarily correlate with emotional experience suggesting that wider non-expert raters can provide equally valid ratings in the domain of emotion recognition from speech also.

There are a number of challenges with using crowdsourcing to label speech assets, including how to select which assets are presented for rating, how to estimate the reliability or bias of the annotators, how to derive the ground truth for the asset and maintaining the balance between data coverage and data quality. Our work in this area is considering these issues for crowdsourcing ratings for a high quality emotional speech corpus which has been generated using Mood Induction Procedures. We are developing an online rating tool which will use active learning techniques to select assets to present for ratings to the raters that have been identified as the best performing raters up to that point in the process.

Applying Bourdieuan social theory to the interaction between identity, power and inclusive practice in a minority language school

DiGiorgio C.

University of Prince Edward Island, Canada

This study aimed to study the process of inclusion as it developed alongside identity and power relationships in a relatively new school with a strong cultural mandate. Students with special needs have been found to have additional difficulties due to cultural, linguistic, and economic challenges at the school level (Manyak, 2002; Hanson and Gutierrez, 1997; McCray and Garcia, 2002; Gay, 2002). The school in this study faced these challenges as well, due to their mandate to develop minority culture in a hegemonic English environment.

This study is an ethnographic case study of one school as it developed over the period of a year, in its early years of implementing an inclusion policy as set out by the provincial government. Through interviews, participant observation, and document analysis, data were gathered, and analyzed continuously using the “constant comparative method” of Glaser and Strauss (1967). This grounded theory approach led to a theory elaboration of sociologist Pierre Bourdieu’s theory of social organization (1996, 1992, 1986, 1982). Bourdieu’s notions of habitus, capital, and field were applied to inclusion, and the results brought new insight into the relationship between individual and group experiences of school for all stakeholders involved in inclusion

Detection of musical instruments in intervals and chords

Eichhoff M. and Weihs C.

TU Dortmund, Chair of Computational Statistics

To determine the portion of a musical instrument type in a piece of music, the classification of musical instruments is useful. For this, a few spectral and temporal (see [5]) features are used to classify single tones (see [2], [3]). This ansatz has been applied to the classification of intervals consisting of two tones and chords containing three or four tones. Eight instrument classes of frequently played musical instruments are considered: “violin”, “viola”, “cello”, “flute”, “trumpet”, “acoustic guitar”, “e-guitar”, “piano”.

A prefiltering method based on the AIC-criterion (see [1]) is used to select variables before starting the statistical classification methods such as LDA, SVM, random forest, decision trees or boosting methods. Evaluation is carried out by using the Hamming Loss (see [4]) and the selected variables are analyzed.

References

- [1] Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716-723 (1974) doi:10.1109/TAC.1974.1100705
- [2] Eichhoff, M., Vatulkin, I. and Weihs, C. (2010): Musical Instrument Detection based on Extended Feature Analysis, GfKI-Cladag 2010, 08.-10. September, Florence, Italy
- [3] Eichhoff, M. and Weihs, C. (2010): Musical Instrument Recognition by High-Level Features, 34th annual conference of the Gesellschaft für Klassifikation (GfKI), 21.-23. Juli, Karlsruhe, Deutschland
- [4] Li, T., Zhang C. and Zhu, S. (2006): Empirical Studies on Multi-label Classification, Proceedings ICTAI '06 Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, Washington, DC, USA, 2006
- [5] Weihs, C. and Ligges, U. (2003): Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: R. Kopiez, A. C. Lehmann, I. Wolther and C. Wolf (Eds.): Proceedings of the 5th Triennial ESCOM Conference, Hanover University of Music and Drama. Germany, September 8-13, 332-335.

Conformal predictors and testing exchangeability assumption

Fedorova V., Nouretdinov I., and Gammerman A.

Computer Learning Research Centre
Royal Holloway, University of London

It's traditional in machine learning to assume that our data follow the *i.i.d.* assumption: the data are independent and identically distributed (or a slightly weaker assumption of exchangeability). If this assumption is satisfied we can prove the validity of the results in many algorithms, for example Conformal Predictors [1, 2] and others but a violation of the assumption can give incorrect results. Therefore it is important to test the assumption. In [1] a technique has been developed to check the assumption. The main idea is to calculate a value of martingale. If the final value of martingale is large it means that our assumption about data set has not been satisfied. To test assumption of exchangeability we calculate a martingale value for a sequence of p-values is generating by Conformal Predictors in on-line mode.

The paper presents the results of exchangeability testing using several known and real-life data sets: the Iris data set, the USPS handwritten digits data set, the abdominal pain [3]. The martingale's values for each of these data sets were calculated for the original sequence of data and then after a permutation of examples: if the original data distribution was not exchangeable, it would become one after the permutation or the opposite way. Naturally, this idea of using martingales can be extended to check various assumptions and the paper outlines a plan for future research that would involve testing of several popular assumptions such as Gaussian.

References

- [1] V.Vovk, A.Gammerman, G.Shafer. "Algorithmic Learning in a Random World", Springer, 2005.
- [2] A.Gammerman and V.Vovk. "Hedging Prediction in Machine Learning, The Computer Journal, v.50, No.2, pp.151-163, 2007.
- [3] A.Gammerman and A. R.Thatcher. "Bayesian Diagnostic probabilities without assuming independence of symptoms", Meth. Inf. Med., v. 30, pp.15-22 ,1991.

Behaviour - based surveillance videos analysis

Fernandez Arguedas V. and Izquierdo E.

Multimedia Vision Research Group, School of Electronic Engineering and Computer Science, Queen Mary, University of London, London, E1 4NS.

{virginia.fernandez, ebroul.izquierdo}@eecs.qmul.ac.uk

The recent exponential popularity of surveillance systems could be largely attributed to increased safety concerns of the public. Recently, Norris and McCahill [1] estimated that UK accommodates more than 4.2 million surveillance cameras. The CCTV ubiquitousness generates everyday a large amount of video data which requires constant human supervision. In efforts to mitigate the dependency of constant supervision of surveillance footage, several object and event detection and recognition algorithms has been developed in recent times. Broadly, these techniques could be classified into two categories (i) based on low-level visual features and (ii) based on psychological studies.

Psychological studies have shown that human beings can routinely discriminate and recognize the type of object using motion pattern, even in large viewing distances or poor visibility conditions where other familiarity cues such as appearance tend to disappear [2]. Inspired by these studies, an object behaviour-based surveillance video analysis framework is proposed in this paper. The motion components extracted from the surveillance footage are exploited to construct behaviour modelling of objects. The behaviour features extracted include, velocity, trajectory and shape proportion. (i) Velocity aims to discriminate objects from different classes in extreme situations. (ii) Trajectory exploits object directionality and is the keystone of the approach. (iii) Shape proportion exhibits the object shape ratio, real size and proportion. The behaviour features thus extracted are further processed using a fuzzy rule based classifier for indexing different objects.

References

- [1] McCahill, M. and Norris, C.: "Estimating the extent, sophistication and legality of CCTV in London". In: CCTV, pp. 51-66, 2003.
- [2] Johansson, G., "Visual perception of biological motion and a model for its analysis". In: Attention, Perception, & Psychophysics, vol. 14, no. 2, pp. 201-211, 1973, Springer.

Lost in transformation: Data mining beyond CRISP

Fischer S.

Rapid-I. Dortmund, Germany

Today's data mining solutions offer a vast variety of functionality of which the actual "learning" component is only one step in many. Any real-world data mining project involves various sub-problems like data access, integration and transformation, result visualization etc., all of which can typically be implemented in manifold ways. All of this makes the successful application of data mining techniques a challenge even for experienced data analysts.

On the other hand, data mining is becoming an important tool for biologists, physicists, and other expert scientists to whom this richness of possible operations may be irritating and even intimidating. It is therefore necessary to find approaches that assist users of data mining software to find solutions to their problems easily and quickly while at the same time retaining the possibility to use the full-grown palette of operators like they exist, e.g. in RapidMiner.

In this talk, we present various approaches to this challenge. We show how the community of scientists can be supported by sharing knowledge and experience and turning them into operational systems to assist other analysts pursuing similar tasks. We show how we can guide them through the jungle of operators, based on hierarchical workflow planning, meta mining, and using recommender system techniques. And we show how they can develop their prototype into a productive system using RapidAnalytics.

Trawling for students – How do educational institutions regulate competing for students?

Frederiksen J. T.

Department of Psychology and Educational Studies,
Roskilde University
janjaf@ruc.dk

Educational institutions are increasing economically dependent on recruitment for their survival. Educational institutions, in Denmark as elsewhere, are thus competing for advantageous position within the field of education, as well as competing for student recruitment. Dominant positions within the field of education does not in itself translate to a secure position in terms of recruitment, for number of reasons, educational policies being the most obvious one. The relative positions of educational institutions are thus competitively related – or possibly allied - on several levels. This paper examines the relations between the Danish National Institutes of Social Education and their recruitment. Geometric data analysis shows the population of students is structured by different forms of capital related to two fields: the field of education and the field of welfare work (wherein the social educator profession resides). The relations between institutions are shown to be structured by these two forms of capital, but distinct, separate subsets of competitive relations relate to each of the two fields. In the first plane geography and the nomos of the field of education structure the competition, yet the second plane is structured by tacit alliances on differentiated strategies for attracting students.

Advanced analysis and classification of images using conformal predictors

González S., Vega J., Pereira A., and Pastor I.

Asociación EURATOM/CIEMAT para Fusión, Madrid 28040, Spain

A novel technique for image analysis and classification based on conformal predictors has been developed. To perform a detailed analysis of images, they have been divided into regions and, for each region, a multiclass conformal predictor (one vs. rest) has been created. The values of confidence and credibility are used, on the one hand, to hedge the classifier outputs and, on the other hand, to choose the most relevant regions in the image.

Since there are different classifiers working on the classification of a single image, it is possible to obtain a disagreement in their responses. Therefore, a combination of the responses generated by the most important classifiers is given as the image class.

This technique has been applied to the classification of images in the Thomson Scattering diagnostic of a nuclear fusion device: the TJ-II stellarator. There are 5 different types of images. A database of more than 1200 TJ-II Thomson Scattering images has been analyzed. Results will be given for different region sizes and will be compared to a previous classification work.

The canonical analysis of distance

Gower J. C. (1), le Roux N. J. (2), and Lubbe S. (3)

- (1) The Open University, Milton Keynes, UK
- (2) Stellenbosch University, Stellenbosch, South Africa
- (3) University of Cape Town, Cape Town, South Africa

Canonical Variate Analysis (CVA) is one of the most useful of multivariate methods. It is concerned with separating between and within-group variation among N samples from K populations with respect to p measured variables. Mahalanobis distance between the K group means can be represented as points in a $(K - 1)$ dimensional space and approximated in a smaller space, with the variables shown as calibrated biplot axes. Within group variation may also be shown, together with circular confidence regions and convex prediction regions, which may be used to discriminate new samples.

This type of representation extends to what we term Analysis of Distance (AoD) whenever a Euclidean inter-sample distance is defined. Although the $N \times N$ matrix, which may be large, is required, eigenvalue calculations are needed only for the much smaller $K \times K$ matrix of distances between group means. All the ancillary information that is attached to a CVA analysis is available in an AoD analysis.

We outline the theory and the R programs we developed to implement AoD by presenting two examples: (a) using Pythagorean distance and (b) using Clark's distance. In our first example we consider measurements made on eight digitized variables for five predetermined groups of ore froth. The overlap and separation among the five groups of data is shown in an AoD biplot with linear biplot axes. Our second example involves measurements on eight wood and pulp variables for samples of five species of South African pine trees. The pine tree data set allows an AoD even though the within groups covariance matrices are significantly different or even singular. Using Clark's distance the associated AoD biplot has nonlinear biplot axes.

Bourdieu and the geometric analysis of data

Grenfell M.

Trinity College, University of Dublin, Ireland

This paper will introduce the Bourdieu strand to the conference. It will set the context for papers across the four days.

The paper will begin with raising the essential epistemological principles which underlie Bourdieu's theory of practice; namely his attempt to synthesis objective and subjective approaches. It will address this issue with special focus on the way 'structure' featured across his oeuvre. Key concepts of *Field*, *Habitus* and *Capital* will be introduced.

The paper will then consider the challenge of '*field theory*'; what it sets out to do, and the key questions that emerge from it. This discussion will lead to an exploration of the 'key' stages and aspects in his methodology; namely, the construction of the research object, 3-stage *field* analysis (structural relationships between *fields* and the field of power, the structure of the *field*); and the *habitus* of *field* participants); and participant objectivation.

The paper will address the challenges of depicting and representing the outcome of such analyses graphically. I shall consider Multiple Correspondence Analysis, and what Bourdieu has to say about it. This approach will be set along side other graphic representations of data as a way of considering the pros and cons of their respective forms.

The paper will highlight some of the issues arising from the above; in particular, it will raise questions pertinent to both the MCA and non-MCA applications of the conference strand.

The whole is offered by way of showing how far MCA can answer Bourdieu's need to depict the structure of *fields*, and indeed their operations. The potential limitations of the approach will be considered alongside other methods of graphic representation in order to develop an agenda for extending future applications.

Re-presenting the social world: Bourdieu and graphic illustration of fields

Hardy C.

Liverpool John Moores University

Bourdieu employed a number of data analysis techniques, and often published his findings in graphic forms (See amongst others, Bourdieu (1979/1984) and Bourdieu (1992/1996)). This paper begins by considering a range of the diagrams used by Bourdieu, and the underlying rationale for their use. Some advantages and disadvantages of each will be discussed briefly.

A series of small-scale empirical 'field analyses' are presented in which a variety of graphic illustrations of data have been utilised. The paper will demonstrate how a three-level relational analysis can be undertaken by using a methodological approach based on that advocated by Bourdieu (See, for example, Bourdieu and Wacquant 1992, and Bourdieu 1993/1999) and will consider how such an approach 'mimics' the relational analyses of larger scale methods from MCA and GMA.

I shall show how empirical data is collected and analysed, and how this is then used to generate structural features of fields, which can be used as part of the representation of the research topic. These representations will be graphically illustrated. I chose my examples from a range of areas, but will concentrate primarily on art and education. Various diagrams and schemes will be referred to for contrast and comparison.

My objective is to tease out the essential aspects of such an approach and explore what it offers as a small scale alternative or for preliminary investigations for larger scale Geometric Analyses. This paper is offered by way of contrast to more formal statistical procedures, all whilst respecting the conventions of Bourdieusian field analyses.

How to find the best cluster analysis method (for social stratification based on mixed data)?

Hennig C.

Department of Statistical Science, University College London

This presentation will treat two issues. 1. A general approach for the decision about the ``best" cluster analysis method will be sketched. This is based on the idea that a subjective decision is needed about the cluster concept required in a particular application, and that there is no such thing as an objectively true clustering determined by the data alone. For example, even if data can be perfectly fitted by a Gaussian mixture, this does not necessarily mean that the mixture components correspond to the ``true" clusters, because in a given application unimodal mixtures of more than one Gaussian may be considered as a single cluster (see Hennig 2010).

On the other hand, large within-cluster dissimilarities may not be admissible, in which case certain Gaussian components need to be split up.

2. The general philosophy is applied to the problem of defining social strata from data with mixed continuous, ordinal and categorical variables. Clusterings as obtained from a mixture/latent class approach (as for example fitted by the LatentGOLD software, Vermunt and Magidson, 2005) are compared with clusterings obtained from applying k -medoids (Kaufman and Rousseeuw, 1990) to dissimilarities balancing the different types of variables in a sensible user-specified way.

References

Hennig, C. (2010): Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4, 3-34.

Kaufman, L. and Rousseeuw, P. J. (1990): *Finding Groups in Data*. Wiley, New York.

Vermunt, J. K. and Magidson, J. (2005): *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont Massachusetts.

Cultural Distinctions: A Geometric Data Analysis.

Hjellbrekke J. and Korsnes O.

Department of Sociology,
University of Bergen. Norway
johs.hjellbrekke@sos.uib.no

In the ISSP-surveys “Social Inequality III” and “Social Inequality IV”, the Norwegian respondents stood out with by far the most egalitarian views of their respective country. Against this background, the relevance of a class analytical approach in general, and the applicability of Bourdieu’s theory of fields and capitals have been subject to multiple debates (Skarpenes & Sakslind 2010, Chan, Birkelund, Aas & Wiborg 2010). One of the claims has been that the middle class has internalized the egalitarian structures deeply embedded in Norwegian society and are mobilized against cultural hierarchies. Also for this reason, Bourdieu’s notion of cultural capital is problematic (Skarpenes 2007).

Inspired by Bourdieu’s classic study “Distinction” (Bourdieu 1979) and by Le Roux and Rouanet’s “Geometric Data Analysis” (Le Roux & Rouanet 2010), this paper offers an analysis of the structures of taste and cultural preferences in Norway, which often is perceived as an *egalitarian* society (Hjellbrekke & Korsnes 2006). The data originate from “The Culture and Media Survey 2008”, and we analyze a representative sample of Norwegians 24 yrs and older (N=1194).

44 questions on 6 different topics have been included in the analysis. These are variables on

- TV-preferences,
- participation in cultural activities
 - music preferences
 - newspaper and magazine readership
 - interest in books and literature
 - radio listening preferences

By way of specific multiple correspondence analysis, hierarchical cluster analysis, concentration and confidence ellipses and class specific analysis (see Le Roux & Rouanet 2010), a 3-dimensional space with 6 clusters is identified and examined in greater detail.

- Axis 1 separates between engaged/active vs. disengaged/inactive respondents.
- Axis 2 is a primarily media/TV-axis
- Axis 3 separates between traditional vs. new/emerging media preferences

When class, age educational level and sex are introduced as structuring factors in this space, the results contradict the claims by Chan & Goldthorpe regarding the social stratification of cultural preferences (2005, 2007a,b), but are more accordance with the analyses of Le Roux, Rouanet, Savage and Warde (2008) and Bennett & al. (2009) of the UK-case. Differences in cultural preferences and practices are strongly related to class inequalities and inequalities with respect to educational levels, i.e. institutionalized cultural capital, along axis 1. Axis 2 is less clear, but describes a minor opposition between men and women. Axis 3 is an age-axis.

The 6 cluster within the fulldimensional space can be described as follows:

- Cluster 1, 26% of the respondents, is a cluster of “actives”; they frequent cultural arrangements relatively frequently
- Cluster 2, 10,5%, is a cluster of “hyperengaged”.
- Cluster 3, 4,7%, is a cluster of respondents with traditional TV-preferences.
- Cluster 4, 8%, is primarily av TV- and media-cluster
- Cluster 5, 25%, is a cluster of music listeners
- Cluster 6, 26%, is a cluster of inactives.

In the next step of the analysis, internal oppositions in cluster 2 is studied in greater detail by way of Class Specific MCA (Le Roux & Rouanet 2010). The first results indicate an internal between two types of “univores” (Peterson & Simkus 1992); the ones who are high-frequenters of the theater vs. the high-frequenters of art exhibitions and museums. The paper ends with an analysis of the internal oppositions in the EGP 1– Class, the upper Service Class 1.

Valid predictions with confidence estimation in air pollution problem

Ivina O. (1), Nouretdinov I. (2), and Gammerman A. (2)

(1) GRECS research group, University of Girona

(2) CLRC Royal Holloway, University of London

The present study is aimed to evaluate the levels of air pollution for the Barcelona Metropolitan Region. A newly developed approach called Conformal Predictors is considered here, and particular use is made of the Ridge Regression Confidence Machine. This technique allows to compute confidence intervals for regression estimation without making any prior assumption on data distribution - except that the data should be independently and identically distributed (iid). The hallmark of this method is that it gives valid interval estimates. Validity here means that for each observation, the fitted interval encloses the actual value of the parameter (label) with a given confidence level.

The data for this study has been provided by the XVPCA (Network for Monitoring and Forecasting of Air Pollution) of the Generalitat of Catalonia. Three pollutants are taken up in this paper: PM10, CO and NO2. The levels have been observed annually, and the timeline embraces the period from 1998 to 2009. Unfortunately, the observations are not available for every year and station, and that sufficiently reduces the scope of this study.

In the first iteration of modelling, only the geographical (mercator) coordinates have been used as regressors for given year and pollutant. It has yielded good results, as for the given confidence level of 80 per cent, the number of errors have not exceeded the number of allowed errors. Also, the prediction intervals are narrow: lesser than the minimum observed value in the set.

The following step towards increasing the precision of the prediction implies adding more regressors to the model. Later on the model shall be transformed from the linear to a non-linear specification by taking up an appropriate kernel function, like polynomial and the RBF kernel.

References

- Vladimir Vovk, Alex Gammerman and Glenn Shafer (2005), *Algorithmic learning in a random world*. New York: Springer.
- Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman. (2009) On-line predictive linear regression. *Annals of Statistics*, 37:1566 - 1590.
- Barceló M.A., Saez M., Saurina C. (2009) Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. *Science of the Total Environment*, 407 (21): 5501 - 5523.

Handling missing values with regularized iterative multiple correspondence analysis

Josse J. (1), Chavent M. (2), Liquet B. (3), and Husson F. (1)

(1) Agrocampus, 65 rue de St-Brieuc, 35042 Rennes, France
(2) Université V. Segalen Bordeaux 2, 146 rue L. Saignat, 33076 Bordeaux, France
(3) Equipe Biostatistique de l'U897 INSERM, ISPED

A common approach to deal with missing values in Exploratory Data Analysis consists in minimizing the loss function over all non-missing elements. This can be achieved by EM-type algorithms where an iterative imputation of the missing values is performed during the estimation of the axes and components. This presentation proposes such an algorithm, named iterative MCA, to handle missing values in Multiple Correspondence Analysis (MCA). This algorithm, based on an iterative PCA algorithm, is described and its properties are studied. We point out the overfitting problem and propose a regularized version of the algorithm to overcome this major issue. Performances of the regularized iterative MCA algorithm are assessed from both simulations and a real dataset. Results are promising for MAR and MCAR values (Little and Rubin, 1987, 2002) with respect to other methods such as missing-data passive modified margin, an adaptation of missing passive method used in Gifi's Homogeneity analysis framework.

References

- M. Greenacre & R. Pardo (2006). Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological methods and research*, 35 (2):193-218.
- J. Josse, J. Pagès & F. Husson (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique*, 150: 28-51.
- R. J. A. Little & D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New-York.
- Y. Takane & H. Hwang (2006). Regularized multiple correspondence analysis. In J Blasius and M J Greenacre, editors, *Multiple Correspondence Analysis and Related Methods*, pages 259-279. Chapman & Hall.

Clustering data described with discrete distributions: an application on population pyramids

Korenjak-Černe S. (1), Kežzar N. (2), and Batagelj V. (3)

- (1) University of Ljubljana, Faculty of Economics, simona.cerne@ef.uni-lj.si
- (2) University of Ljubljana, Faculty of Medicine, Institute for Biostatistics and Medical Informatics, natasa.kejzar@mf.uni-lj.si
- (3) University of Ljubljana, Faculty of Mathematics and Physics, vladimir.batagelj@fmf.uni-lj.si

Population pyramid presents an age-sex distribution of the human population of a particular region. Its shape is mainly influenced by main demographical indicators as fertility, mortality and migrations. Besides them, also many other social and political policies and events affect it. Clusters of countries with similar pyramidal shapes can offer additional insight into the study of countries for field-related researchers.

The 'centroid' pyramids of the clusters of countries, obtained with classical clustering methods, do not represent the real population pyramids describing the whole population in the clusters. We present an alternative approach, where the population pyramids are clustered based on the description with two vectors (one for each gender) of age group frequencies. Such a description of variables with discrete distributions (i.e. histograms) represent a special kind of symbolic data. For clustering so represented population pyramids, the adapted Ward's hierarchical clustering method and the adapted leaders clustering methods were developed. The implementation was done in R. The main advantages of these methods compared to the classical approaches are: such a representation of the population pyramid with two vectors of age distributions preserves information about population size for each gender, the information about population size for each gender can be included into a clustering process, and the adapted clustering methods produce meaningful cluster representatives – real population pyramids.

The results of the adapted hierarchical clustering for the population pyramids of the world countries (International Data Base 2008), and the results obtained with the combination of the methods for the data of 3219 US Counties (US Census 2000) will be presented.

Verbalisation tasks in hall test sessions

Kostov B. (1), Bécue-Bertaut M. (2), Pagès J. (3), Cadoret M. (3), Torrens J. (4), and Urpi P. (4)

- (1) CAP Les Corts - Hospital Clínic
- (2) Universitat Politècnica de Catalunya
- (3) Agrocampus-Ouest
- (4) Freixenet S.A.

In hall test sessions organized in food and wine industries for disclosing the sensory attributes of products, free-text comments are more and more used. This kind of session tackles the valuation of 8 to 12 products by a panel of 12 to 20 judges and thus the whole of the comments constitute a very short corpus, quite different of those usually analysed by text-mining methods.

In this work, supported by a particular application to a series of cavas, we insist on the acquisition of the textual data and their pre-processing, taking into account the very specific field of wine industry. The way of giving rise to verbalisation is essential as it structures the textual data and that is just what makes their analysis possible.

We also present an original methodology to tackle the repeatability of the panel and panellists, through different sessions, as measured from their free-text comments. The starting point of the methodology consists in building the global and partial cava's words table. The global table contains the frequency with which every word is used (by all the panellists along both sessions) to qualify every cava. The partial tables correspond to only one session or only one session/ panellist. Then, global and partial points of view are introduced in correspondence analysis (CA), in a multiple factor analysis way (MFA).

Positioning political producers: The field of American campaign professionals

Laurison D.

University of California, Berkley

Political professionals shape nearly every aspect of Americans' experience of politics, and yet we know next to nothing about the ways agents enter into or gain status within this field. Taking inspiration from Bourdieu's work on the field of politics (primarily in *Language and Symbolic Power* [1991]), this paper uses Multiple Correspondence Analysis to understand the types of capital and oppositions at play among these key participants in the political field. It analyzes an original dataset of career histories for over 3000 staffers, consultants, and advisors involved in the 2008 United States Presidential & Senate campaigns, plus data drawn from in-depth interviews with 60 of these politicians.

I show that the field is structured along two primary axes: the first describes the opposition between incumbents and challengers, or those with more and less field-specific capital (which is primarily indicated by the number and quality of positions held in politics); the second describes the opposition between heteronomous and autonomous poles, with those who move in and out of politics, who work primarily in consulting shops, and who do media work on the more heteronomous side.

I then examine more closely the positions and position-takings of my 60 interview subjects. There are many interesting ways in which the position-takings of my interviewees are structured by their positions in the field. For example, the challengers – newer entrants to the field, with smaller quantities of field-specific capital – consistently critique “tired” ways of doing things, while the incumbents disparage “kids” who are always going after some “shiny new toy” rather than sticking to “tried and true” methods of running campaigns and generating political products. A key opposition between the heteronomous and autonomous poles of the field has to do with views of “success” – while fully political politicians define success for themselves as winning campaigns or (often secondarily) improving their reputation, those on the more heteronomous side talked more about money, or to a lesser extent, the thrill of the competition.

Meta-analysis methods for gene expression profiles

Lausen B.

Department of Mathematical Sciences, University of Essex, Colchester, UK

A fast increasing amount of public available gene expression data sets allows the use of meta analysis techniques to validate and to identify molecular signatures.

I discuss several recent approaches. Evaluating predictive and diagnostic marker sets an important condition for pre-processing methods is that the data analysis work flow should not be influenced by properties of other data sets included in the meta analysis. For example a pre-processing method of one Affymetrix cel file should be invariant under different sets of cel files included in the meta analysis. Proposing a simple data analysis approach I illustrate the talk with the evaluation of predictive and diagnostic rules based on gene expression data sets of colorectal and breast cancer.

References

- Buffa, F.M., Harris, A.L., West, C.M., Miller, C.J. (2010): Large Meta-analysis of Multiple Cancers Reveals a Common, Compact and Highly Prognostic Hypoxia Metagene. *British Journal of Cancer*, 102, 428-35.
- Gorlov, I.P., Sircar, K., Zhao, H., et al. (2010): Prioritizing genes associated with prostate cancer development. *BMC Cancer* 10:599.
- McCall, M.N., Bolstad, B.M., Irizarry, R.A. (2010): Frozen robust multiarray analysis (fRMA). *Biostatistics* 11, 2, 242-253.
- Mpindi, J.P., Sara, H., Haapa-paananen, S. et al. (2011): GTI: A Novel Algorithm for Identifying Outlier Gene Expression Profiles from Integrated Microarray Datasets. *PLoSone* 6, 2, e17259.
- Shi, F., Abraham, G., Leckie, C., Haviv, I., Kowalczyk, A. (2011): Meta-analysis of gene expression microarrays with missing replicates. *BMC Bioinformatics* 12,84.

Geometric data analysis of French cultural practices.

Lebaron F. (1), Bonnet P. (2), and Le Roux B. (3)

(1) Centre Universitaire de Recherche sur l'Action Publique et le Politique (CURAPP), UMR 6054, Université de Picardie – Jules Verne et CNRS.

(2) Laboratoire de Psychologie et Neuropsychologie Cognitive (LPNCog), FRE 3292, CNRS et Paris-Descartes.

(3) CEVIPOF/CNRS, Centre de Recherches Politiques, Sciences Po Paris et MAP5/CNRS, Université Paris Descartes.

Our proposal is to make an assessment of the latest breakthroughs in geometric data analysis. This will be illustrated with cultural practices data from the permanent INSEE survey (2003) about cultural and sport participation. The aim is to prolong the theoretical and methodological approach Pierre Bourdieu and Monique de Saint-Martin initiated in “L’anatomie du goût”. This approach can be enriched with the new possibilities of geometric data analysis.

Different kinds of problems will be examined at different steps of analysis:

- preparation of the data table: choose active individuals and active variables and encode categories;
- choose the method (MCA, specific MCA);
- interpretation of axes ;
- supplementary elements: individuals and variables;
- deep investigation of the cloud of individuals (structuring factors, concentration ellipses, between-within variance,...);
- euclidean clustering and interpretation of the clusters in the space of individuals;
- statistical inference.

These problems will be presented and illustrated within the data analysis of cultural practices and lifestyles.

References

Bourdieu, P. (1979). *La distinction. Critique sociale du jugement*. Paris: Minuit.

Le Roux, B. & Rouanet, H. (2010). *Multiple Correspondence Analysis (QASS Series)*. Thousand Oaks, CA: Sage.

Rouanet, H., Ackermann, W., Le Roux, B. (2000). The geometric analysis of questionnaires: The lesson of Bourdieu's *La Distinction*. *Bulletin de Méthodologie Sociologique*, 65, 5-18.

Clusters defined by distinct correlation structures

Longford N. T.

SNTL and UPF, Barcelona, Spain

In the established view, clusters are subsets (or subpopulations) that have well separated locations (centroids) in relation to their dispersions. This paper discusses a largely neglected way how meaningful subpopulations can be defined for multivariate outcomes -- the clusters may arise as having different covariance or correlation structures. Mixture modelling (of multivariate data) is well set for studying such clusters, without requiring any adaptation.

Experience from mixture analysis of income in the European Community Household Panel (ECHP) and the longitudinal part of the European Union Statistics on Income and Living Conditions (EU-SILC) will be discussed and the relation of correlations structures to patterns of annual household income, and its stability in particular, highlighted.

An improper component is included in the analysis to take care of observations with aberrant patterns of income, often with a very small income in one year. It helps to reduce the number of components and to distill the patterns in each proper component.

References

Longford, N.T., and Pittau, G.P. (2006). Stability of household income in the European countries in the 1990's. *Computational Statistics and Data Analysis* 51, 1364-1383.

Longford, N.T., and Nicodemo, C. (2011). A mixture analysis of income in European countries. Unpublished.

Methods for reliable classification of network traffic

Luo Z.

Department of Computer Science.
Royal Holloway, University of London

Accurate classification of network traffic can offer substantial benefits to service differentiation, enforcement of security policies and traffic engineering for many network operators and service providers [1]. Reliable classification requires algorithmic capability, in particular, machine learning algorithms. The idea of using machine learning techniques for network traffic classification is not new. Many machine learning algorithms have been successfully used to classify network traffic with good results [2].

However, not knowing the confidence of these classifications makes it difficult to measure and control the risk of error using a decision rule. Modern network resource management systems are becoming increasingly complex and as such require high quality, reliable predictions to optimise performance. Therefore, introducing more reliable machine learning algorithms for network resource management will result in higher quality of service for network users.

In this paper, we consider the problem of reliable network traffic classification. We present two recently developed machine learning algorithms [3], namely the Conformal Predictor and Venn Probability Machine, for making reliable decisions under uncertainty and achieving performance guarantees.

These two algorithms are based on the i.i.d. assumptions and do not depend on the probability distribution of examples.

Experiments on publicly available real network traffic datasets show these two algorithms can perform well and produce reliable classifications [4]. Comparison is also made between these two algorithms.

References

- [1] M. Mellia, A. Pescapé and L. Salgarelli. Traffic classification and its applications to modern networks. *Computer Networks*, Vol. 53, pp. 759-760, 2009.
- [2] N. Williams, S. Zander and G. Armitage. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *SIGCOMM Comput. Commun. Rev.* Vol. 36, No. 5, pp. 5-16, 2006.
- [3] V. Vovk, A. Gammerman and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.
- [4] A.W. Moore and D. Zuev. Internet traffic classification using Bayesian analysis techniques. *SIGMETRICS*, 2005.

Show me a picture: Visualising interactive machine learning algorithms

Mac Namee B.

Applied Intelligence Research Centre
Dublin Institute of Technology, Dublin, Ireland.

In interactive (or human-in-the-loop) machine learning [9] analysts interact with automated machine learning processes in order to guide them to better solutions. Examples of machine learning tasks that can benefit from interactive input include active learning [4], case base maintenance [2], and feature selection [9]. One of the key challenges in designing human-in-the-loop machine learning solutions is presenting all of the required information to analysts in a way that is easily understood, and so allows them to make the best decisions to guide the process. The use of visualisations, and in particular holistic visualisations of an entire dataset, can be extremely useful in this regard.

For example, a large set of documents could be visualised using a force directed graph in which each point represents a document and the points are arranged such that points that are most similar to one-another appear closest together. The colour and shape of each point represents the class of the underlying document. This presentation of the dataset (which when used by a machine learning algorithm would be represented as a term-document matrix of thousands of numbers indicating the frequency of words in each document) is easily interpreted by a human user and can allow easy interpretation.

While the creation of visualisations such as that described is not new [6, 7, 5] there are many open questions as to how to use such visualisations most effectively for interactive machine learning tasks. This paper will consider which dataset visualisation approaches are best suited to particular interactive machine learning tasks: namely active learning [4], case base maintenance [2], and similarity measure design and selection [6].

References

- [1] Chen, K., Liu, L.: ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Trans. Inf. Syst.* 24(2), 245–294 (2006)
- [2] Delany, S.J., Cunningham, P.: An analysis of case-base editing in a spam filtering system. In: *ECCBR*. pp. 128–141 (2004)
- [3] Hu, R., Delany, S.J., Mac Namee, B.: Sampling with confidence: Using k-nn confidence measures in active learning. In: *Proceedings of the UKDS Workshop at 8th International Conference on Case-based Reasoning (ICCBR 09)*. pp. 181–192 (2009)
- [4] Hu, R., Delany, S.J., Mac Namee, B.: Egal: Exploration guided active learning for tcbr. In: *Proceedings of International Conference on Case-based Reasoning (ICCBR) 2010* (2010)
- [5] van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* 9, 2579–2605 (2008)
- [6] Mac Namee, B., Delany, S.J.: Cbtv: Visualising case bases for similarity measure design and selection. In: *Proceedings of the International Conference on Case-based Reasoning (ICCBR) 2010* (2010)

Linear time hierarchical clustering using the Baire metric, and alternative Perspectives: p-Adic clustering, generalized ultrametrics, hashing, and More

Murtagh F. (1, 2) and Contreras P. (2, 3)

(1) Science Foundation Ireland, Dublin, Ireland.

(2) Department of Computer Science, Royal Holloway, University
of London, Egham, England.

(3) Thinking Safe Ltd., Egham, England.

Email: fmurtagh@acm.org

Through use of the Baire, or longest common prefix substring, metric, we have a powerful means of inducing a hierarchical clustering on a data set in linear computation time. This is due to the fact that the Baire metric is simultaneously an ultrametric.

There are various vantage points on this new hierarchical clustering method. The clusters themselves can be determined by a hashing algorithm. The ultrametricity induced by reading off the Baire distances can be used to enhance inherent hierarchical or ultrametric properties in the data, through modifying the data precision of measurement. The generalized ultrametric also arises naturally in this framework and, through it, lattice representation of the data.

References

F. Murtagh, G. Downs and P. Contreras, "Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding", *SIAM Journal on Scientific Computing* 30, 707-730, 2008.

P. Contreras and F. Murtagh. "Fast hierarchical clustering from the Baire distance", *Classification as a Tool for Research*. Eds. H Locarek-Junge and C Weihs. Springer, pp 235-243, 2010.

P. Contreras and F. Murtagh. "Fast, linear time hierarchical clustering using the Baire metric". *Journal of Classification*, submitted, 2011.

F. Murtagh and P. Contreras, "Fast redshift clustering with the Baire (ultra) metric", in *Proceedings of Science: Image in Action*, 7th International "Livio Scarsi and Vito De Gesu" Data Analysis in Astronomy, Erice, Italy, 2011, World Scientific, in press, 2011.

A metric and ultrametric platform for analysis of semantics and change

Murtagh F. (1,2), Ganz A. (3), and Reddington J. (4)

(1) Science Foundation Ireland, Dublin, Ireland.

(2) Department of Computer Science, Royal Holloway, University of London

(3) Department of Media Arts, Royal Holloway, University of London

(4) School of Computing, University of Teesside.

Email: fmurtagh@acm.org

We study two aspects of information semantics: the collection of all relationships, and tracking and spotting anomaly and change.

Correspondence analysis, which maps data in a range of input formats, including frequencies of occurrence, into a Euclidean space is a useful tool for exploring all pairwise relationships between observables and/or attributes. As such it is a tool for exploration of semantics of the domain being investigated. Additional insight into the data is provided by hierarchical clustering, which is conveniently constructed from the Euclidean embedding provided by correspondence analysis. When the hierarchical clustering takes account of a sequencing of the observables (for example, on a time-line or successive sections in a document), then hierarchical clustering can capture change or anomaly, as we will show. An ultrametric is more restrictive than a metric (e.g. Euclidean). An ultrametric is defined by a hierarchical clustering.

This data analysis "platform" will be exemplified through a wide range of case studies. These include the Colombian social violence (I explore the 1990-2004 period here); and document or multimedia object analysis and/or synthesis. I will list out some of the many interesting avenues opened up for further investigation in this work.

References

F. Murtagh, *Correspondence Analysis and Data Coding with R and Java*, Chapman and Hall/CRC Press, 2005.

F. Murtagh, M. Spagat and J.A. Restrepo, "Ultrametric wavelet regression of multivariate time series: Application to Colombian conflict analysis", *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 41, 254-263, 2011.

F. Murtagh, A. Ganz and J. Reddington, "New methods of analysis of narrative and semantics in support of interactivity", *Entertainment Computing*, in press, 2011.

F. Murtagh, A. Ganz and S. McKie, "The structure of narrative: the case of film scripts", *Pattern Recognition*, 42, 302--312, 2009. (See discussion in Z. Merali, "Here's looking at you, kid. Software promises to identify blockbuster scripts.", *Nature*, 453, p. 708, 4 June 2008.)

F. Murtagh, A. Ganz, S. McKie, J. Mothe and K. Englmeier, "Tag Clouds for Displaying Semantics: The Case of Filmscripts", *Information Visualization Journal*, 9, 253-262, 2010

F. Murtagh, "The Correspondence Analysis platform for uncovering deep structure in data and information", *Sixth Boole Lecture, Computer Journal*, 53 (3), 304-315, 2010.

Non-conformity measures in multi-class prediction

Nouretdinov I. (1), Santos M. (2), and Gammerman A. (1)

(1) CLRC Royal Holloway, University of London

(2) Complutense University of Madrid

To apply conformal predictors [1] for a classification or regression estimation problems we need to develop a suitable measure of non-conformity (or strangeness). Here we present three such measures based on nearest neighbour algorithms that are useful in multi-class prediction. For an illustration purpose we consider the following problem: given certain spectral reflections in atmosphere, identify the corresponding type of material that makes this reflection. The number of possible classes (materials) is, usually, high (in the region of 100 classes) and the number of examples (pixels) various in different datasets between a thousand and twenty thousand with noise.

Conformal predictor outputs a list of materials. The method guarantees that it covers the true classification with probability corresponding to confidence level.

Several non-conformity measures for this task are derived in accordance with complex structure of classes. Measure A is based on nearest neighbours without limitation of its number. Measure B exploits the two-level hierarchical structure: type (group of classes with similar names) and class. Measure C does not rely on external information, structure is based on empirical separability of classes. Efficiency of measures is understood in terms of small size of predictive sets: the less is the choice of possible materials the more informative the prediction is. The results are presented and discussed.

Reference

[1] Alexander Gammerman and Vladimir Vovk. Hedging Prediction in Machine Learning. *The Computer Journal* 2007 50: 173-177.

Clustering Students by their Skill Set Profiles on the Unit Hypercube

Nugent R. (1) and Dean N. (2)

(1) Department of Statistics. Carnegie Mellon University

(2) Department of Statistics. University of Glasgow

In cognitive diagnosis models, mastery of a skill is often measured from zero (no mastery) to one (complete mastery). Grouping students into similar skill set profiles then requires clustering in a restricted unit hypercube space with as many dimensions as skills in the profile. The standard methods for clustering often have implicit spherical/elliptical shape constraints on the clusters they can discover. Examples of such methods include: k-means, complete hierarchical clustering, ward's method, and model-based clustering (with finite Gaussian mixtures). While these methods should perform well for discovering groups in the center of the hypercube, they may perform poorly when searching for groups of students at or near the corners. This work compares the performance of commonly used, but potentially problematic methods (e.g. k-means) with alternative clustering methods tailored to the shape of the space. We discuss the performance of a finite mixture of beta distributions (estimated by the EM algorithm) as well as k-means and model-based clustering applied to arcsine transformed data. Results from both simulated data from various models and an educational data mining data set from an intelligent online tutor will be presented.

Understanding the immune system by Correspondence Analysis

Ono M. (1), Kano M. (2), and Sugiman T. (3)

1) Institute of Child Health, University College London, London, UK.

2) Department of Chemical Engineering, Kyoto University, Kyoto, Japan.

3) Department of Human Coexistence, Graduate School of Human Environmental Studies, Kyoto University, Kyoto, Japan.

Background: The immune system protects the human body against the invasion of pathogens, and its abnormalities result in various diseases. Immunologists have classified cells in the immune system (immunocytes) into various cell populations to understand how the system works. However, the relationships between cell populations are still ambiguous, as immunology lacks the insight of multidimensional analysis. Nowadays, tens of thousands of gene expressions (concentrations of gene products, namely, ribonucleic acids [RNAs]) in each cell population can be simultaneously measured by microarray technology. Thus, it is possible to understand immunocytes by analysing the patterns of gene expressions.

Objectives: To visualise the relationships between immunocytes and genes, thereby provide big pictures of the immune system and ways to explore their relationships.

Results: It was essential to decrease the noise of microarray measurements. Here we propose to “polish” data by using principal component analysis (PCA) with considering between-group variance. The patterns of gene expressions and cell populations in the polished data were visualized by Correspondence Analysis (CA). The proposed method not only confirmed the current immunological knowledge but also revealed unexpected associations between cell populations and gene expressions.

Conclusion: CA is a powerful tool for understanding the qualitative data such as cell population and gene expression in molecular biology/ genome biology. PCA can be used for preprocessing for CA in microarray datasets.

Distinguishing and Classifying from n-ary Relations

Préa P. and Rolbert M.

Laboratoire d'Informatique Fondamentale de Marseille LIF, UMR CNRS 6166, Marseille, France

In this talk, we will present an application of computational linguistics to classification.

Given a discourse and its entities (i.e. the objects/persons that are mentioned in the discourse), a referring expression of an entity A is an expression which is true for A and no other entity [1]. An entity which admits a referring expression is said to be distinguishable. In [2], an incremental definition of distinguishable entities is given. An entity A is "0-distinguishable" from an entity B if A has a property that B does not have; an entity A is "k-distinguishable" ($k > 0$) from an entity B if A is in relation (via a property P) with an entity C such that, if B is in relation with an entity D via P, C is (k-1)-distinguishable from D. Consequently, and from a computational linguistics point of view, an entity A is k-distinguishable from an entity B if there exists an expression of length k which is true for A but not for B. This definition yields a hierarchical classification of the entities: If we call $K(A,B)$, the smallest k such that the entities A is k-distinguishable from B or B is k-distinguishable from A, $1/(K+1)$ is an ultrametric. In addition, the values $K(A,B)$ can be computed in optimal time.

This notion of k-distinguishability can be applied to other fields than computational linguistics, such as classification or data analysis. Actually, a discourse is considered as a set of objects (the entities) with a set of relations (unary, binary,...) between objects. Starting just from this set of relations (which can be seen as the internal structure of the data base), we are able to give, in optimal time, a natural hierarchy on the objects.

References

[1] R. Dale, Cooking up Referring Expressions, Proceedings of the 27th Annual Meeting of the ACL, Vancouver, 1989.

[2] M. Rolbert & P. Préa, Distinguishable Entities: Definition and Properties, Proceedings of the 12th ENLG, Athens, 2009

Model-Based Classification and Clustering in High Dimensions

Raftery A. E.

University of Washington

I will describe model-based classification and clustering methods for high-dimensional data in two contexts. The first is a model-based classification method for determining food authenticity from high-dimensional Near Infrared Spectroscopy data. The model includes variable selection and is fitted in a semi-supervised manner using both labeled and unlabeled data.

The second is a computationally fast method for clustering social network data using the latent position model-based clustering model of Hoff et al (2002) and Handcock et al (2007).

The computational cost of likelihood-based inference for this model is of order $O(N^2)$, where N is the number of nodes, making it infeasible for large networks. Borrowing the case-control idea from epidemiology, we construct a case-control likelihood which is an unbiased estimator of the full likelihood. This reduces the computational time to $O(N)$, making it feasible for large networks.

This is joint work with Brendan Murphy, Nema Dean, Xiaoyue Niu, Peter Hoff and Ka Yee Yeung.

Leading indicators of currency crises in emerging economies

Reid S.

University of Stellenbosch, Stellenbosch, South Africa

Currency crises have erupted with surprising regularity in emerging markets over the last few decades. Emerging market governments, funding any number of spending programmes and encouraging sometimes dubious private sector dealings, found foreign debt (both private and public) reaching unsustainable levels. A sharp currency depreciation was all that was required for debilitating balance sheet effects to take hold and plunge large chunks of their economies into insolvency. The aim of this paper is to use a variety of economic indicators to predict such currency crises 12 months before they occur. Crises are identified as historically large currency depreciations and/or large decreases in foreign currency reserves. Once crises are identified, the preceding 12 months are flagged as “pre-crisis” months. A classifier is used to determine whether a country finds itself in a “pre-crisis” or “tranquil” state in any given month.

The paper builds on an often neglected insight of the Signals model of currency crises developed by Kaminsky et al. (1997): that of casting the problem of detecting crises as one of statistical binary classification. Once regarded as such, many powerful modern statistical classification techniques can be used to aid our quest.

A version of the Signals model was fit to approximately 18 years of South African experience (from 1991 to 2008). Cross validation was used to determine important model tuning parameters. Linear discriminant (with novel variable selection technique) and boosted tree classifiers were also fit to the data over this period. The latter produced truly encouraging results, providing a perfect fit to historical data and superior out of sample prediction performance (consistently detecting between 8 and 11 of pre-crisis months correctly).

The boosted tree model was also fit to two other major emerging economies with considerable outward orientation: Brazil and South Korea. Similarly encouraging results were obtained.

References

Kaminsky, G., Lizondo, S. & Reinhart, C. M. 1997. Leading Indicators of Currency Crises. IMF Working Paper no. WP/97/79, July 1997, <<http://www.imf.org/external/pubs/ft/wp/wp9779.pdf>>

Revisiting clustering and classification in information retrieval.

Rijsbergen K. V.

University of Glasgow

I will present some thoughts about the theoretical underpinnings of hierarchical clustering. The clusterings will be derived from dissimilarity measures, these also have a theoretical foundation. The question of the evaluation of cluster methods will be addressed. And, finally, an attempt will be made to present an approach to the logical analysis of clustering. On the way I will relate this to IR where appropriate.

Beauty and the eye of the beholder: aesthetic dispositions and museum audiences in Flanders (Belgium)

Roose H. (1), Hanquinet L. (2), and Savage M. (3)

(1) Department of Sociology. Ghent University (Belgium) henk.roose@ugent.be

(2) Department of Sociology. The University of York (United Kingdom) laurie.hanquinet@york.ac.uk

(3) Department of Sociology. The University of York (United Kingdom)

Mike.Savage@york.ac.uk

Views on what the fine arts are/should be and how they should be appropriated have changed during the last century—in line with Bourdieu's conception of the emergence and development of a field. The 'traditional' view grounded in 19th century bourgeois society with its emphasis on beauty, elegance, balance, and emotional aesthetics has been challenged by the emergence of artistic movements based on transgression, experimentation, conceptual forms, sensuousness. At present—anno 2011—these different forms of art and ditto forms of reception coexist and are more or less similarly institutionalized. One of the key findings of *La Distinction* is exactly showing that aesthetic judgment and taste are socially anchored. In this paper we want to contribute two things using Multiple Correspondence Analysis (MCA):

- (1) we use an item battery specifically designed to measure the different aesthetic dispositions and conceptions. We want to construct the space of aesthetic perception by means of about twenty items describing how/what art should be or incorporate: e.g. 'Art should be beautiful' versus 'One colour or one line are enough to create a piece of art';
- (2) we analyze the relationship between these different aesthetic conceptions and the social origin and environment of spectators, i.e. besides including the traditional socio-demographic variables, we insert items that indicate perceptions of someone's social position, and perceptions of what members of his/her social network prefer.

The use of MCA has a number of advantages. First, it allows for the inclusion of a lot of variables simultaneously. Second, it explicitly visualizes a space in terms of a relational logic. Third, it enables the inclusion of a number of structuring—supplementary—factors showing the homology or association between aesthetic perception and their social embeddedness. We rely on an audience survey of about 1,000 visitors from two art museums in Ghent (Belgium), namely the Museum of Fine Arts and the City Museum for Contemporary Art (also known as S.M.A.K.).

Mixture semiparametric regression

Schepers J.

Maastricht University, The Netherlands

When studying the predictive relation between a set of predictor variables and a criterion variable, mixture (or latent class) regression methods can be used to cope with population heterogeneity. The key assumption is that the observations are sampled from a population that consists of a small number of subgroups or clusters (with unknown cluster sizes) where a different regression model holds within each subgroup. This talk will present an extension of the mixture regression model that allows the class-specific effect of (a subset of) the predictor variables to be described by an unspecified function, the only assumption being smoothness. An approach to estimate this novel model, called mixture semiparametric regression, as well as a model selection tool, are proposed and evaluated using a simulation study. Finally, an application of the novel method is presented.

Auto-focus algorithm selection: A methodology for comparing blur perception between observers.

Shilston R. and Stentiford F.

Department of Computer Science, University College London

Since the middle of the 20th century the technological development of conventional photographic cameras has taken advantage of the advances in electronics and signal processing. One specific area that has benefited from these developments is that of auto-focus, the ability for a camera's optical arrangement to be altered so as to ensure the subject of the scene is in focus. However, whilst the precise focus point can be known for a single point in a scene, the method for selecting a best focus for the entire scene is an unsolved problem. Many focus algorithms have been proposed and compared (eg [1, 2, 3]), though no overall comparison between all algorithms has been made, nor have the results been compared with human observers. This work describes a methodology that was developed to benchmark focus algorithms against human results. Experiments that capture quantitative metrics about human observers were developed and conducted with a large set of observers on a diverse range of equipment. From these experiments, it was found that humans were highly consensual in their experimental responses. The human results were then used as a benchmark, against which equivalent experiments were performed by each of the candidate focus algorithms. A second set of experiments, conducted in a controlled environment, captured the underlying human psychophysical blur discrimination thresholds in natural scenes (extending [4]). The resultant thresholds were then characterised and compared against equivalent discrimination thresholds obtained by using the candidate focus algorithms as automated observers. The results of this comparison and how this should guide the selection of an auto-focus algorithm are discussed, with comment being passed on how focus algorithms may need to change to cope with future imaging techniques.

References

- [1] F. C. Groen, I. T. Young, and G. Ligthart, "A comparison of different focus functions for use in autofocus algorithms," *Cytometry*, vol. 6, no. 2, pp. 81–91, 1985.
- [2] A. Santos, C. Ortiz de Solorzano, J. J. Vaquero, J. M. Pena, N. Malpica, and F. del Pozo, "Evaluation of autofocus functions in molecular cytogenetic analysis," *Journal of Microscopy*, vol. 188, no. 3, pp. 264–272, 1997.
- [3] Y. Sun, S. Duthaler, and B. Nelson, "Autofocusing in computer microscopy: Selecting the optimal focus algorithm," *Microscopy Research and Technique*, vol. 65, no. 3, pp. 139–149, 2004.
- [4] S. M. Wuerger, H. Owens, and S. Westland, "Blur tolerance for luminance and chromatic stimuli," *Journal of the Optical Society of America A*, vol. 18, no. 6, pp. 1231–1239, June 2000.

A method of pattern recognition applied to the poggendorff illusion

Stentiford F.

University College London

This paper applies a recognition mechanism using the properties of interest points to the problem of modelling the Poggendorff illusion. Recognition is based upon matching the intensity gradient at interest points and the angular relationship between pairs of such points. The strength of the recognition is determined by the size of matching cliques of interest points [6]. The recognition mechanism is shown to exhibit the same effects as the human visual system on the standard illusion and reduced effects are modelled on a variant without parallels. The model shows that the effect can be explained as a perceptual compromise between the alignment of the elements in the oblique axis and their displacement from each other in the vertical. In addition an explanation is offered how obtuse angled variants of the Poggendorff figures yield stronger effects than the acute angled variants. The results lend support to the approach to pattern recognition in which clusters of patterns are identified according to the strength of properties that are reflected between them, and not on pre-selected measurements.

References

- [1] B. Spehar and B. Gillam, "Modal completion in the Poggendorff illusion: support for the depth-processing theory," *Psychological Science*, vol. 13, no. 4, pp. 306-312, 2002.
- [2] M. J. Morgan, "The Poggendorff illusion: a bias in the estimation of the orientation of virtual lines by second-stage filters," *Vision Research*, vol. 39, pp. 2361-2389, 1999.
- [3] Y. Yu and Y. Choe, "Angular disinhibition effect in a modified Poggendorff illusion," *Proc. 26th Ann. Conf. of the Cognitive Soc.*, pp. 1500-1505, 2004.
- [4] G. Westheimer and C. Wehrhahn, "Real and virtual borders in the Poggendorff illusion," *Perception*, vol. 26, pp. 1495-1501, 1997.
- [5] R. H. Day, "The Poggendorff illusion and apparent interparallel extents," *Perception*, vol. 21, pp. 599-610, 1992.
- [6] F. W. M. Stentiford and A. Bamidele, "Image recognition using maximal cliques of interest points," *ICIP*, Hong Kong, 2010.

Text detection in natural scenes using cliques of interest points for mobile visual search

Stentiford F. (1) and Bamidele A. (2)

(1) University College London

(2) Nokia UK

Text location in natural scenes is important for automatically identifying and describing images. Many approaches rely upon extracting a number of features thought to characterise text and follow this with a series of constraints to improve performance [1-5]. This paper describes a method of detecting text that makes use of the structure present in groups of interest points [6]. Pairs of interest points and their local gradients are extracted and are matched against those present in a generic character. Those pairs that bear the same angular relationship become members of a partial clique of interest points whose size and degree of connectivity determines whether text is present. A large and highly connected partial clique signifies that the structure of the image object is strongly reflected in the generic character. This approach avoids the pre-selection of features and any subsequent training and does not require that all text items are subjected to the same set of measurements for detection.

References

- [1] J. Liang, D. Doermann, H. Li, "Camera-based analysis of text and documents: a survey", *International Journal on Document Analysis and Recognition*, vol. 7, no 2-3, pp. 83-200, 2005.
- [2] X. Chen, A. Yuille, "Detecting and Reading Text in Natural Scenes", *Computer Vision and Pattern Recognition (CVPR)*, pp. 366-373, 2004.
- [3] Q. Liu, C. Jung, S. Kim, Y. Moon and J. Kim, "Stroke filter for text localization in video images," in *Proc. Int. Conf. Image Process.*, Atlanta, GA, USA, pp. 1473-1476, 2006.
- [4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," *CVPR*, 2010.
- [5] K. Subramanian, P. Natarajan, M. Decerbo and D. Castanon, "Character-stroke detection for text-localization and extraction," *Int. Conf. on Document Analysis and Recognition*, 2007.
- [6] F. W. M. Stentiford and A. Bamidele, "Image recognition using maximal cliques of interest points," *ICIP*, Hong Kong, 2010.

A recurrence quantification analysis based approach to lung sounds classification

Sultornsanee S., Zeid I., and Kamarthi S.

Department of Mechanical and Industrial Engineering,
Northeastern University, Boston, MA 02115, USA

Respiratory diseases are a major cause of sickness throughout the world. Therefore, the study and classification of lung sounds have attracted attention over the years. There exist multiple classification methods such as multivariate linear autoregressive model, wavelet coefficients, and combined neural network and genetic algorithm. However, many of these current methods of time series analysis do not rely on the characteristics of lung sound signals because of their complexity, nonlinearity, and non-stationarity. In this paper, we introduce a novel method of analysis of lung sound signals using recurrence quantification analysis and classification using support vector machines. Lung sound signals are transformed into recurrence plots and a set of statistical features are extracted using recurrence quantification analysis. Support vector machine employing radial basis function is implemented to classify the normality and abnormality of lung sounds in respiratory system. Examining the acoustic patterns in lung sounds, they are classified into one of the five categories: normal sounds of inspiratory, normal sounds of expiratory, crackles, rhonchus, and wheezes. The results show that the proposed method is successful with 100% classification accuracy. The accurate results indicate that the proposed lung sounds classification method is very effective for real world lung sound classification applications.

Simultaneously selecting categorical features and the number of clusters in model-based clustering

Silvestre C., Cardoso M., and Figueiredo M.

School of Communication and Media Studies, IPL - Portugal
Dep. of Quantitative Methods, ISCTE-IUL, Portugal
Department of Electrical and Computer Engineering, IST, Portugal

The determination of the number of clusters and the selection of the subset of relevant features, from data, have been active research areas in clustering, with most methods devoted to numerical data. In this work, a finite mixture of multinomials is used for clustering categorical data. The proposed approach simultaneously addresses the problems of estimating the number of clusters and the subset of relevant categorical features, which are usually strongly inter-related. We adopt a concept of feature saliency and a minimum message length (MML) criterion is used to estimate both the number of clusters and the relevant features. In this setting, the MML-type criterion encourages the saliencies of the irrelevant features to be null, thus allowing the selection of the subset of relevant features. The resulting criterion is addressed by a variant of the expectation-maximization (EM) algorithm, which estimates all the parameters of the mixture, including the number of clusters and the features saliencies. Results obtained with synthetic data evidence a promising performance.

Extracting Lexical Chains for Text Discrimination and Categorisation

Thurlow I. (1) and Stentiford F. (2)

(1) BT

(2) UCL

Each year BT submits a huge number of project reports to its clients. From BT's perspective, the purpose of the project report is to satisfy all the requirements of the client and to enter into a continuing dialogue. To ensure the continued quality of its documentation, BT completed a quality assessment study to try to find out where improvements could be made. Accordingly, domain experts judged a representative sample of BT's report documents against a number of quality indicators, including compliance against the clients' requirements and the quality of writing; the executive summaries were also judged separately against the same indicators.

This work uses the categorisation given by the domain experts to identify the features of the text that best characterise a set of previously judged executive summaries; overall, the intention is to use those features in a self-help tool that will help authors to improve future reports in terms of their quality and effectiveness. To this end, this work attempts to reproduce the human characterisation of the executive summaries through an analysis of the texts; in particular, through the identification of lexical chains that discriminate between executive summaries that, according to the quality assessment indicators, have been classed as being fit for purpose from those that show some weaknesses.

Commonly, the use of lexical chains as features in textual analysis is restricted to non-function words such as nouns and verbs and is effective on those documents that suit the selected vocabulary [1,2,3]. Categorisation, however, is more difficult if a more diverse vocabulary is employed or if the document categories are more subjective. This paper describes a method of extracting lexical chains that provide optimal discrimination between classes of documents. The chains may consist of contiguous or non-contiguous sequences of any words in the document vocabularies. Some comparative results are reported on an analysis of a range of executive summaries.

References

- [1] N Stokes, "Spoken and written news story segmentation using lexical chains," Proc HLT-NAACL, pp 49-54, 2003.
- [2] R Barzilay and M Elhadad, "Using lexical chains for text summarization," Proc. Workshop on Intelligent Scalable Text Summarization, 1997.
- [3] M Galley and K McKeown, "Improving word sense disambiguation in lexical chaining," Proc. IJCAI, 2003.

In which social context will working class students obtain a university diploma?

Tribess A.

Université de Picardie Jules Verne, Amiens

While *Bourdieu*-orientated researches usually tend to reveal social inequalities in the access to the best university studies, this paper seeks to examine what kind of social context will best help children of poorer families to obtain university qualifications. By “social context” I mean the mixture of social origins. My hypothesis is to say that most of the children from the working class will suffer in a group of students mainly composed of children from the upper class; not because they aren’t able to follow the program, but because they will suffer by cultural and social exclusion.

Researching actually on the data of *Picardian* Students from 1998 to 2008, I’d like to confront the obtainment of university qualifications in different geographical and disciplinary contexts of the University of Picardie (North of France). I will classify those contexts by the percentages of different social classes in the first and further years of the study. Would the results of the *Multiple Correspondence Analysis* be transformed by introducing the new variable of the “social context”? What can we conclude about the importance of the social context for the success of the working class students in University Studies?

Widening participation in higher education: Capital that counts

Watson J.

University of Southampton

The under-representation in higher education of those from less privileged social backgrounds is an enduring problem in the UK. While on an individual basis there are examples of productive participation, the pattern of collective trajectories of this group differs sharply from that of traditional entrants (Reay, 2006). Predictably, the onus falls on students to adapt to the established practices of the field which remain very much oriented towards its traditional white middle-class population and effectively resists inclusivity (Layer, 2002; Read et al., 2003; Burke, 2005), regardless of governmental policy objectives.

Analysis of qualitative data emerging from a three-year longitudinal case study exploring the educational experiences of students with non-traditional academic backgrounds studying in one of the UK's research-intensive universities was underpinned by Bourdieu's theory of practice. The findings highlight the role of academic, linguistic, social and practice-oriented capital in developing a feel for and learning to play 'the game' in this sub-field of higher education and the positional tendencies and trajectories of the study's thirteen volunteer participants. This paper will outline the nature and illustrate the role played by these key forms of capital in the 'affinities, convergences and divergences' (Grenfell, 2007 p.137) experienced by participants.

References

- Burke, P. J. (2005) "Access and widening participation." *British Journal of Sociology in Education* 26(4): 555-562.
- Grenfell, M. (2007) *Pierre Bourdieu: Education and training*. London, Continuum International Publishing Group.
- Layer, G. (2002) "Developing inclusivity." *International Journal of Lifelong Education* 21(1): 3-12.
- Read, B., Archer, L. and Leathwood, C. (2003) "Challenging cultures? Student conceptions of 'belonging' and 'isolation' in a post-1992 university." *Studies in Higher Education* 28(3): 261-277.
- Reay, D. (2006) "The zombie stalking English schools: Social class and educational inequality." *British Journal of Educational Studies* 54(3): 288-397.

Mapping an illegitimate field: power relations in international education

Widin J.

Faculty of Arts and Social Sciences, University of Technology Sydney

The field of university led international English language education projects (IELEPs) is a relatively recent development within the field of Australian higher education. I draw on Bourdieu's 'thinking tools' to analyse the way IELEPs work within the broader social context, the field of power. In this paper I present my interpretation of Bourdieu's approach to examine the relationships between the different agents, the positions they occupy in the field and the capital which they have accumulated.

Australian higher education is a particularly volatile field: government funding and education policies have changed dramatically in the last decade, student demographics are changing in line with the general trend of an ageing population in industrialised societies and the role of the university as the prime supplier of tertiary education is contested by training providers in the private sector. The international education field adds another level of complexity in that it crosses national boundaries highlighting inequities in distribution of resources and the uneven playing field of the market for cultural capital.

In this presentation I demonstrate how I use Bourdieu to map out three distinct though not hierarchical levels of the IELEPs: (i) the field vis-à-vis the meta-field of power; (ii) the map of relations between the agents in the field and; (iii) analysis of the habitus of agents. The three levels of analysis must be considered interdependently however here I focus on Level Two to investigate the way capital and consequently, benefits are distributed in the IELEP field.

Optimization in k-means: some new development

Wishart D.

School of Management
University of St Andrews

The Euclidean Sum of Square (ESS) criterion function to be optimized is $E_p = \sum_{i \in p} c_i \sum_j w_j (x_{ij} - \mu_{pj})^2 / \sum_j w_j$ where x_{ij} are the data, c_i are differential case weights, w_j are differential variable weights, and μ_{pj} are the means in a partition of k clusters. Total ESS over all k clusters is $E = \sum_p E_p$.

We apply an exact relocation test as follows: move a case i currently in cluster p to another cluster q when $E_{p-i} + E_{q+i} < E_p + E_q$, that is to say, make all moves that reduce the criterion function ESS. Since ESS is a positive sum of squares, which by definition is reduced by each such relocation, the algorithm must converge to a stable k-means solution from which no further relocations can be made. This may be the optimum (minimum) least squares solution, or one of several higher-order sub-optimal stable solutions.

We contrast the exact k-means relocation test with a naïve k-means implementation, where a case i is relocated when $d_{iq}^2 < d_{ip}^2$, but it will be shown that this is not guaranteed to converge. Such implementations necessarily resort to ad hoc termination rules to prevent the algorithm oscillating indefinitely, and their solutions can be sub-optimal.

Random starting criteria are applied in a series of k-means trial replications, to seek the optimum ESS solution. Because exact k-means must converge, the frequency of each stable solution can be computed and all solutions are ordered by ascending ESS values. The solution with the least ESS may be the global optimum solution, and we propose a stability statistic to test its replicability.

A dendrogram is constructed for the optimum k-means solution, by compiling a within-cluster sub-tree for each cluster and a between-cluster super-tree between the k clusters. The dendrogram can then be ordered by serialization of the distance matrix, to obtain a representation of the data that has a meaningful linear interpretation of cases and clusters, as illustrated by the author's "Whisky Classified" application.

References

Wishart, D (2006), ClustanGraphics Primer, Clustan Ltd, Edinburgh.

Wishart, D (2006), Whisky Classified: Choosing Single Malts by Flavour, Pavilion Books, London.

Sources

www.clustan.com

www.whiskyclassified.com

We'll tak' a cup o' kindness yet *The Story of Scotch Whisky*

Wishart D.

School of Management
University of St Andrews

Discover the flavours of Scotch, Irish and Welsh malt whiskies at this talk and tasting by David Wishart, Fellow of Management and author of *Whisky Classified: Choosing Single Malts by Flavour*, second edition, now published in eight languages.

Dubbed the "Carl Linnæus" of whisky by fellow writer Charles MacLean, David Wishart was the first to categorise single malt whiskies by flavour. He will guide you through the history and romance of Scotch whisky, from the *aqua vitae* of the early monasteries, the alchemist's art of turning barley into liquid gold, and the hedonistic *uisge beatha* of remote Scottish crofts, to smuggling into Royal Mile taverns, royal parties at Holyrood, and hot toddies in Edinburgh's New Town.

Whisky is evoked in the poetry of Burns, in the travelogues of Stevenson, and in the art of Landseer and Wilkie. London toasted with brandy during the Regency period, but when a tiny phylloxera beetle devastated Cognac in 1863 the upper classes turned to whisky and the famous "Scotch" brands were born.

Today, the flavour of malt whisky is more diverse than ever, due to the influence of variable peating, cask preparation, extended maturation, and special finishing. David describes his unique scientific classification by flavour based on sensory analysis and profiling, with a selection of fine single malt whiskies to taste.

He has chosen some well-known favourites for his tasting, plus several malts and pot still Irish whiskeys that are harder to find. They span the whisky "flavour spectrum", the complete range of flavours of single malt whiskies as described in David's book "Whisky Classified".

For tonight's tasting he will be featuring Aberfeldy, Aberlour, Ardmore, Balvenie, Ben Nevis, Benromach, Bunnahabhain, Connemara, Dalmore, Glencadam, Glenfarclas, Glenfiddich, Glengoyne, Glenlivet, Glenrothes, Isle of Jura, Laphroaig, Ledaig, Longmorn, Macallan Fine Oak, Macallan Sherry Oak, Old Pulteney, Penderyn, Singleton of Dufftown, Smokehead, Speyburn, Talisker, Tobermory, Tomatin, Tomintoul, Tullibardine and Tyrconnell single malt whiskies, Green Spot and Redbreast single pot still whiskeys, and Cutty Sark blended Scotch whisky.

ICC-2011 The International Conference on Classification
with a Special Symposium on Bourdieu and Geometric Analysis of Data
Hosted by The British Classification Society
We'll tak' a cup o' kindness yet - the story of Scotch whisky
School of Medical Sciences, University of St Andrews
19.30 Wednesday July 13, 2011 **Dr David Wishart**
An illustrated talk and tasting of fine malt whiskies **School of Management**
University of St Andrews

References

Wishart, D (2006), *Whisky Classified: Choosing Single Malts by Flavour*, Pavilion Books, London.

Sources

www.whiskyclassified.com

Application of Bayesian Model Averaging to the construction of regulatory networks

Yeung K. Y.

Department of Biostatistics, University of Washington

The inference of regulatory and biochemical networks from large-scale genomics data is a basic problem in molecular biology. The goal is to generate testable hypotheses of gene-to-gene influences and subsequently to design bench experiments to confirm these network predictions. We generated microarray data measuring time-dependent gene expression levels in 95 genotyped yeast segregants subjected to a drug perturbation, and developed a Bayesian model averaging regression algorithm that incorporates external information from diverse data types. We showed that our inferred network recovers existing and novel regulatory relationships. Following network construction, we generated independent microarray data on selected deletion mutants to prospectively test network predictions. Applying our construction method to previously published data demonstrates that our method is competitive with leading network construction algorithms in the literature.

Index of Authors

- Albers C., 9
Aloise D., 10
Amorim M. J. P. C., 12
Anderlucci L., 11
Araújo A., 10
- Baier D., 13
Bamidele A., 68
Batagelj V., 14, 47
Bécue-Bertaut M., 15, 48
Bergström Y., 16
Berry V., 17
Bienaise S., 18
Bonnet P., 51
Boutet M., 17
Bouveyron C., 22
Bradley P. E., 19
Braun A. C., 19
Brennan T., 20
Brietenbach M., 20
Brito P., 21
Buhmann J. M., 23
Butler A., 24
- Cadoret M., 48
Cardoso M., 70
Cardoso M. G. M. S., 12
Cebon N., 25
Chakrabarty D., 26
Chavent M., 27, 46
Coavoux S., 17
Contreras P., 55
Cordeiro de Amorim R., 28
Cullen C., 32
- Daher A., 29
Dalberg T., 16
Daniel I., 13
de Tibeiro J. J. S., 30
Dean N., 31, 58
Delany S. J., 32
DiGiorgio C., 33
Doreian P., 14
Duarte Silva A. P., 21
- Eichhoff M., 34
- Fedorova V., 35
Ferligoj A., 14
Fernandez Arguedas V., 36
Figueiredo M., 70
Fischer S., 37
Frederiksen J. T., 38
- Gammerman A., 35, 57
Ganz A., 56
Gettler Summa M., 18
González S., 39
Gower J., 9
Gower J. C., 40
- Hanquinet L., 64
Hardy C., 42
Hennig C., 43
Hennig, C., 11
Hjellbrekke J., 44, 45
Husson F., 46
- Ivina O., 45
Izquierdo E., 36
- Josse J., 46
- Kamarthi S., 69
Kano M., 59
Kejžar N., 47
Korenjak-Černe S., 47
Korsnes O., 44
Kostov B., 15, 48
Kuentz V., 27
- Laurison D., 49, 50
Le Roux B., 51
le Roux N. J., 40
Lebaron F., 51
Lienhart R., 25
Liquet B., 27, 46
Longford N. T., 52
Lubbe S., 40, 41
Luo Z., 53
- Mac Namee B., 54
Mirkin B., 28
Morin A., 15
Murdoch D. J., 30, 31
Murtagh F., 55, 56
- Nouretdinov I., 35, 45, 57
Nugent R., 58
- Ono M., 59, 60
Owen E., 24
- Pagès J., 48
Pastor I., 39
Pereira A., 39
Préa P., 60
- Rafferty A. E., 61
Reddington J., 56
Reid S., 62
Richter F., 25
Rijsbergen K. V., 63
Roose H., 64
- Santos M., 57
Saracco J., 27
Savage M., 64
Schepers J., 65
Shilston R., 66
Silvestre C., 70

Snel J., 32
Stentiford F., 66, 67, 68, 71
Sugiman T., 59
Sultornsanee S., 69

Tarasov A., 32
Thurlow I., 71
Torrens J., 48
Tribess A., 72

Urpi P., 48

Vega J., 39

Watson J., 73
Weihs C., 34
Widin J., 74
Wishart D., 75, 76

Yeung K. Y., 77

Zeid I., 69

ICC-2011 Conference Banquet

St Salvator's College Hall

Founded in 1450

Thursday July 14, 2011



The ICC-2011 “Old Course” Banquet
Highland Welcome (Duncan Souter, Bagpipes)
Five Course Dinner, with Wine and Coffee
Robert Burns’ Address To The Haggis (David Wishart)
Violincello Concert (Gabriela and Patrick Bradley)