IFCS 2015

Conference of the International Federation of Classification Societies

6-8 July 2015,
Bologna, Italy

CONFERENCE PROGRAM and BOOK of ABSTRACTS

Conference Program

Organizing Team

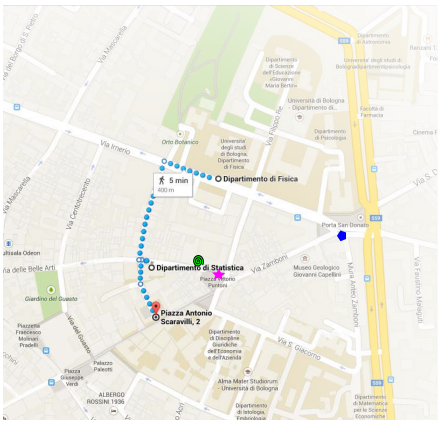
Scientific Program Committee

Francesco Palumbo, Chair SPC
Angela Montanari, Chair LOC
Maurizio Vichi, IFCS President
Iven Van Mechelen, Past-President IFCS
Akinori Okada, President-Elect IFCS
Christian Hennig, Secretary IFCS, BCS
Paul McNicholas, Pubblication Officer IFCS, Canada
Nema Dean, Treasurer IFCS
Andrea Cerioli, CLADAG
Daewoo Choi, KCS
Carlos Cuevas-Covarrubias, SOLCAD
Anuška Ferligoj, SSS
Paolo Giudici, CLADAG
Tadashi Imaizumi, JCS
Berthold Lausen, GFKL
Sugnet Lubbe, SUSE
Mohamed Nadif, SFC
Iannis Papadimitriou, GSDA
Józef Pociecha, SKAD
Abderrahmane Sbihi, MCA
Bryan Scotney, IPRCS
Fernanda Sousa, CLAD
Doug Steinley, CS
José Fernando Vera, SEIO
Jeroen Kornelis Vermunt, VOC
Vincenzo Esposito Vinzi, President ISBIS
Patrick J.F. Groenen, President IASC

Local Organizing Committee

Angela Montanari
Laura Anderlucci
Silvia Bianconcini
Silvia Cagnone
Luca De Angelis
Giuliano Galimberti
Alessandro Lubisco
Paola Monari
Luisa Stracqualursi
Cinzia Viroli

Conference Venues



The conference will be hosted at the Department of Physics and Astronomy (via Irnerio 46) and at the School of Economics, Management and Statistics (via Belle Arti 41, Piazza Scaravilli 2).
The venues are all within a walking distance of a few minutes.

Specifically:

- Sessions scheduled on July 6th in the morning will be held at the Aula[1] Magna of the Department of Physics and Astronomy (via Irnerio 46), ground floor.
- Sessions scheduled on July 6th, in the afternoon, will be held at
  - Room I (ground floor) and Room III (second floor), via Belle Arti 41;
  - Room 1 (ground floor), Room 21 (second floor) and Room 31 (third floor), piazza Scaravilli 2.
- Sessions scheduled on July 7th-8th will be held at
  - Room I (ground floor) and Room III (second floor), via Belle Arti 41;
  - Room 12 (first floor floor), Room 21 (second floor) and Room 31 (third floor), piazza Scaravilli 2.

The coffee break on July 6th in the morning will be served in the garden of the Department of Physics and Astronomy.
All the other coffee breaks will be served in the Statistics library, first floor, via Belle Arti 41.

For those who have bought lunch vouchers, meals will be served at Bononia University Restaurant, Piazza Vittorio Puntoni 1 (see fuchsia star on the map).

[1] "Aula", which you may find written in many places, is the Italian word for "Room"

Social Program

The conference welcome reception will be held on July 6th (7.15 p.m.) at the Accademia di Belle Arti ("Academy of Fine Arts of Bologna"), that is a public tertiary academy of fine art in Bologna (in via delle Belle Arti, 54 - see green spiral on the map at page 4).

The conference dinner will be held on July 7th at Palazzo Albergati, a beautiful country dwelling named after the Bolognese family. It is one of the most significant examples of 17th-century Emilian architecture.

A guided tour of the Palazzo has been planned. A bus service between the conference venue and the villa will be provided. Buses will leave at 6.45 p.m. at Porta S. Donato, in front of the Department of Mathematics (see blue pentagon on the map at page 4).

Monday, 6th July 2015

from 8.30 Start registration

9.00 - 9.30 **Opening ceremony** (Room: Aula Magna, Department of Physics and Astronomy, via Irnerio 46)

9.30 - 10.30 **IFCS 30th birthday** - Chair: *Francesco Palumbo* (Room: Aula Magna, Department of Physics and Astronomy, via Irnerio 46)

10.30 - 11.00 Coffee Break

11.00 - 13.00 **Plenary invited: Generalized Additive Models - 25 years** - Chair: *Angela Montanari* (Room: Aula Magna, Department of Physics and Astronomy, via Irnerio 46)

Generalized additive models: a retrospective, Robert Tibshirani
GAM selection via convex optimization, Trevor Hastie

13.00 - 14.00 Lunch Break

14.00 - 15.20 Parallel Sessions

SP1: Benchmarking in cluster analysis I, organized by I. Van Mechelen - Chair: *N. Dean* (Room I)

Benchmarking in cluster analysis: preview of a white paper, Douglas L. Steinley, Iven Van Mechelen, IFCS Task Force on Benchmarking.
What are the true clusters?, Christian Hennig.
Using principles of optimal treatment regime estimation in simulation studies for benchmarking, Lisa Doove, Iven Van Mechelen, Tom Wilderjans, Antonio Calcagnì.
Discussant: *Willem Heiser*

14.00 - 15.20 **SP2: Classification Models for Forecasting of Economic Processes** - Chair: *J. Pociecha* (Room III)

Comparison of Classification Accuracy for Corporate Bankruptcy Prediction Models Including Changes in Economic Environment, Mateusz Baryla, Barbara Pawelek, Jozef Pociecha.
Student Value Systems, Andreas Geyer-Schulz, Thomas Hummel, Claire Roederer, Victoria-Anne Schweigert.
Application of Extreme Value Theory in the measurement of household investment risk, Krzysztof Jajuga.
Application of ordered multinomial logit model to identification of determinants of LAU2's financial condition, Andrzej Pawel Woloszyn, Romana Glowicka-Woloszyn, Feliks Wysocki.

14.00 - 15.20 **CONTR1: Questionnaire based surveys**, - Chair: *S. Mignani* (Room 1)

IRT-based conjoint analysis in the optimization of banking products, Justyna Brzezińska-Grabowska, Aneta Rybicka, Adam Sagan.
Accounting for attribute non-attendance in subset-conjunctive choice models, Michel Meulders
An adaptive item selection method for curtailment, Niels Smits
A Random Matrix Theory study of Covariance Matrices of Continuous and Categorical Variables, Graziano Vernizzi, Miki Nakai.

14.00 - 15.20 **CONTR2: Methods for the analysis of large data sets**, - Chair: *B. Mirkin* (Room 21)

Large covariance matrix estimation by composite minimization, Matteo Farnè.

Robust regularized discriminant analysis based on implicit weighting, Jan Kalina.
Revisit on Big Data Analysis Framework - How to Make Proper 'Mini data'?, Hiroyuki Minami, Masahiro Mizuta.
A computationally fast variable importance test for random forests for high-dimensional data, Silke Janitza, Anne-Laure Boulesteix.

14.00 - 15.20 **CONTR3: Clustering Agreement** - Chair: *F. Murtagh* (Room 31)

Determining the number of clusters from decompositions of a Rand index-based measure of partitioning stability, Patrice Bertrand, Lassad El Moubarki, Ghazi Bel Mufti.
On comparing partitions, Marjan Cugmas, Anuška Ferligoj.
Set valued prototypes through the consensus analysis, Mario Fordellone, Francesco Palumbo.
Correction for chance and correction for maximum value, Matthijs J. Warrens.

15.30 - 16.30 Parallel Sessions

SP3: Measuring systemic risk to classify financial institution - Chair: *P. Giudici* (Room I)

Estimating Binary Spatial Autoregressive Models for Rare Events, Raffaella Calabrese, Johan A. Elkink.
Stability Transmission and Risk of Islamic Banking Networks in the MENA Region, Shatha Qamhieh Hashem, Paolo Giudici.
Discovering SIFIs in interbank community, Alessandro Spelta.

15.30 - 16.30 **SP4: Mixture models: Recent Developments and Applications for Unsupervised and Supervised Classification** - Chair: *G. McLachlan* (Room III)

On the choice of the number of groups in the context of model-based clustering, Laura Anderlucci, Geoff McLachlan.
The joint role of trimming and constraints in robust estimation for mixtures of skew normal, Francesca Greselin, Agustin Mayo-Iscar, Luis-Angel Garcia-Escudero.
Using mixture models with random effects to test the differences between classes for supervised classification, Shu-Kay (Angus) Ng.

15.30 - 16.30 **SP5: Data Streams Clustering and Classification**, - Chair: *R. Verde* (Room 1)

A new spatial prediction method for georeferenced data streams, Antonio Balzanella, Antonio Irpino, Rosanna Verde.
Functional data analysis for optimizing strategies of cash flow management, Francesca Di Salvo, Marcello Chiodi, Pietro Patricola, Fabrizio Mineo, Claudio Lo Piccolo.
Fuzzy clustering of distribution-valued data usingan adaptive L2 Wasserstein distance, Antonio Irpino, Francisco De Assis Tenhorio De Carvalho, Rosanna Verde.

15.30 - 16.30 **CONTR4: Missing Data** - Chair: *P. Monari* (Room 21)

Missing data imputation by Multitree, Agostino Di Ciaccio.
Missing Data Imputation and Its Effect on the Accuracy of Classification, Lynette Hunt.
Handling missing data in observational clinical studies concerning cardiovascular risk: an evaluation of alternative approaches, Nadia Solaro, Daniela Lucini, Massimo Pagani.

15.30 - 16.30 **CONTR5: Functional Data I** - Chair: *D. Calò* (Room 31)

Correlation analysis for multivariate functional data, Tomasz Górecki, Waldemar Wolynski.
Covariance based classification in multivariate and functional data analysis, Francesca Ieva, Anna Maria Paganoni, Nicholas Tarabelloni.
Generalization, Combination and Extension of Functional Clustering Algorithms, Christina Yassouridis, Friedrich Leisch.

16.30 - 17.00 Coffee Break

17.00 - 18.40 Parallel Sessions

CONTR6: Clustering Methods and Data Fusion - Chair: *F. Palumbo* (Room I)

Fuzzy bi-clustering, with application to open-ended questionnaires, François Bavaud, Pascale Deneulin, Laurent Gautier, Yves Le Fur.
A Self-tuning Region-Growing Algorithm for Deriving Upwelling Areas on Sea Surface Temperature Images, Susana Nascimento, Boris Mirkin, Sérgio Casca.
Marked Point Processes for MicroArray Data Clustering, Khadidja Henni, Olivier Alata, Abdellatif El Idrissi, Brigitte Vannier, Lynda Zaoui, Ahmed Moussa.
A pairwise likelihood approach to simultaneous clustering and dimensional reduction of ordinal data, Monia Ranalli, Roberto Rocci.
Some Statistical paradigm to approach data fusion and high-level information fusion in the Big Data era, Maurizio Salusti.

17.00 - 18.40 **CONTR7: Supervised Learning I** - Chair: *C. Hennig* (Room III)

Machine learning diagnoses on patients presenting abdominal pain, Hong Gu.
Credibility Classification with Missing Data, Toby Kenney.
Clinical Decision Support System for HCC using Classification Models, Taerim Lee.
Machine-learning classification methods with the ability to hesitate, Michal Trzesiok.
Improving Predictions for Tree Ensembles using Distributions of Estimated Probabilities with Applications in Record Linkage, Samuel L. Ventura, Rebecca Nugent.

17.00 - 18.40 **CONTR8: Advances in k-means clustering** - Chair: *H. H. Boch* (Room 1)

T-sharper images and T-level cuts of fuzzy partitions, Slavka Bodjanova.
Minkowski weighted k-means clustering with a median-based consensus rule, Renato Cordeiro de Amorim, Vladimir Makarenkov.
A Note on Spherical k-Means Clustering, Yasunori Endo.
Clustering based on adaptive Mahalanobis kernels, Marcelo Ferreira, Francisco de Carvalho.

17.00 - 18.40 **CONTR9: Regression Models** - Chair: *F. Greselin* (Room 21)

Unit level small area model with covariates perturbed for disclosure limitation, Serena Arima.
SparseStep: Approximating the Counting Norm for Sparse Regularization, Gertjan van den Burg, Patrick Groenen.

Sparse Principal Covariates Regression for high-dimensional data, Katrijn Van Deun, Eva Ceulemans.
Dual model selection for principal covariates regression, Marlies Vervloet, Katrijn Van Deun, Wim Van den Noortgate, Eva Ceulemans.
On Associative Confounder Bias, Priyantha Wijayatunga.

17.00 - 18.40 **CONTR10: Network Analysis** - Chair: *A. Ferligoj* (Room 31)

Clustering of links in networks, Jernej Bodlaj, Vladimir Batagelj.
Two-stage agglomerative hierarchical clustering using medoids for network clustering, Sadaaki Miyamoto.
Network Tools and Homophily Measures for Brand Image Analysis, Agnieszka Stawinoga, Simona Balbi, Germana Scepi.
Predicting the evolution of a constrained network: a beta regression model, Luisa Stracqualursi, Patrizia Agati.

19.15 **Welcome Reception** (at **Accademia di Belle Arti**, via Belle Arti 54)

Tuesday, 7th July 2015

9.00 - 10.00 Parallel Sessions

SP6: Density-based clustering - Chair: *G. Galimberti* (Room I)

Clustering via Mixture Models with Flexible Components, Geoff McLachlan, Sharon Lee.
Density-based clustering multiplex networks, Giovanna Menardi, Domenico De Stefano.
Modal Clustering and cluster inference, Surajit Ray.

9.00 - 10.00 **SP7: Accuracy and validation in clustering and scaling models I**, - Chair: *J. F. Vera* (Room III)

Prediction error in distance-based generalized linear models, Eva Boj del Val, Teresa Costa Cor, Josep Fortiana Gregori.
Prediction Accuracy in Logistic Biplots for categorical data, Jose Luis Vicente-Villardon, Julio Cesar Hernandez-Sanchez.
Cluster Analysis and Distance Stability in Multidimensional Scaling, José Fernando Vera.

9.00 - 10.00 **SP8: Multi-Dimensional Scaling for Sparse Association Matrices**, - Chair: *T. Imaizumi* (Room 12)

Multi-Ddimensional Scaling of Sparse Block Diagonal Similarity Matrix, Tadashi Imaizumi.
The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation, Atsuho Nakayama.
A study on the overlapping cluster analysis for the large data, Satoru Yokoyama.

9.00 - 10.00 **SP9: Data science in Biomedical Research** - Chair: *B. Lausen* (Room 21)

Exhaustive biomarker selection techniques, Hans A. Kestler.
AUC-based splitting criteria for random survival forests, Matthias Schmid, Marvin Wright, Andreas Ziegler.
Ensembles of selected classifiers applied to genomic data, Berthold Lausen, Asma Gul, Zardad Khan, Osama Mahmoud.

9.00 - 10.00 **CONTR11: Change-point detection** - Chair: *C. Conversano* (Room 31)

Comparing the performance of non-parametric change point detection methods for capturing response concordance, Jedelyn Cabrieto, Francis Tuerlinckx, Eva Ceulemans.
Time Series Rolling-Window Cluster Analysis on geological data, Carlo Drago, Fabio Matano, Germana Scepi.
Analysis of Influence Scores for Detecting a Change Point, Kuniyoshi Hayashi, Koji Kurihara.

10.15 - 11.00 **President's invited lecture: Focused graphical model estimation**, Gerda Claeskens - Chair: *Maurizio Vichi* (Room I)

11.00 - 11.30 Coffee Break

11.30 - 13.10 Parallel Sessions

SP10: Cluster Analysis of Asymmetric Relationship - Chair: *A. Okada* (Room I)

Seriation benchmarking environment for different permutations and measures of goodness, Innar Liiv.
Representable Hierarchical Clustering Methods for Asymmetric Networks, Facundo Memoli.
Social Differentiation of Cultural Taste and Practice in Contemporary Japan: Nonhierarchical Asymmetric Cluster Analysis, Miki Nakai.

An Algorithm of Nonhierarchical Asymmetric Cluster Analysis, Akinori Okada, Satoru Yokoyama.
Advances in clustering asymmetric proximity data, Donatella Vicari.

11.30 - 13.10 **CONTR12: Dissimilairity and Distance Measures** - Chair: *A. Irpino* (Room III)

Correcting Jaccard and other similarity indices for chance agreement in cluster analysis, Ahmed Najeeb Albatineh.
Application of spatial median of Weber and positional formulation of TOPSIS method to the assessment of financial condition of local government units, Agnieszka Bernadetta Kozera, Aleksandra Łuczak, Feliks Wysocki.
Distance measures based on the probabilistic information of the data with applications in classification problems, Gabriel Martos Venturini.
Tree-Robinsonian Dissimilarities, Pascal Préa, François Brucker.
Novel similarity measures for categorical data based on mutability and entropy, Zdenek Sulc, Hana Rezankova.

11.30 - 13.10 **CONTR13: Multiple Correspondence Analysis** - Chair: *A. Iodice D'Enza* (Room 12)

Correspondence Analysis in Identification of Structure of Performance Measurement Systems of Polish Firms, Barbara Batóg, Jacek Batóg, Wanda Skoczylas, Andrzej Niemiec, Piotr Waśniewski.
A principal component method to analyze frequency tables connected by contextual variables, Belchin Kostov, Mónica Bécue-Bertaut, François Husson.
A method for recoding ordinal variables, Odysseas Moschidis, Theodore Chadjipadelis.

Big Data Scaling through Metric Mapping: Correspondence Analysis in Very High Dimensional Spaces, Fionn Murtagh.
Biplot-based visualisations to explore relationships between pneumonia and nasopharyngeal pathogens, Johané Nienkemper-Swanepoel, Sugnet Lubbe, Niël Le Roux, Emilee Smith, Heather Zar, Mark Nicol.

11.30 - 13.10 **CONTR14: Methods for DNA data** - Chair: *A. Tenenhaus* (Room 21)

An efficient implementation of adjacency-constrained hierarchical clustering of a band similarity matrix, Alia Dehman, Guillem Rigaill, Pierre Neuvial, Christophe Ambroise.
Tissue-aware age prediction from DNA methylation data, Marcelo Ferreira, Ivan Costa.
Bootstrap test of ordered RIG for multiple testing in genomics of Quantitative Trait Loci in yeasts, Evgeny Mirkes, Thomas Walsh, Edward J. Louis, Alexander N. Gorban.
Discovering modulators of gene regulation by local energy statistics, Teppei Shimamura, Yusuke Matsui.
Haplotype estimation, Haplotype Block Identification and statistical analysis for DNA data, Makoto Tomita.

11.30 - 13.10 **CONTR15: Applied classification and clustering** - Chair: *T. Imaizumi* (Room 31)

Music classification from the user side, Nicolas Greffard, Pascale Kunt.
Transportation clustering, Guillaume Guex, Théophile Emmanouilidis, François Bavaud.
Unsupervised classification of perfusion imaging data using multiple equivalence tests, Fuchen Liu, Yves Rozenholc, Charles-André Cuénod.

Using taxonomies and aggregate rankings for measuring research impact, Boris G. Mirkin, Mikhail A. Orlov.
Cause Related Marketing: a qualitative and quantitative analysis on Pinkwashing, Gabriella Schoier, Patrizia De Luca.

13.10 - 14.10 Lunch Break

14.10 - 14.55 **Presidential Address: Hierarchical Disjoint Non-negative Factor Analysis**, Maurizio Vichi - Chair: *A. Okada* (Room I)

14.55 - 15.40 **Award Session** - Chair: *S. Ingrassia* (Room I)

15.40 - 16.10 Coffee Break

16.10 - 17.30 Parallel Sessions

SP11: Benchmarking in cluster analysis II - Chair: *I. Van Mechelen* (Room I)

The IFCS Cluster Benchmark Data Repository, Friedrich Leisch, IFCS Task Force on Benchmarking.
A Statistical Framework for Hypothesis Testing in Real Data Benchmark Experiments, Anne-Laure Boulesteix.
Benchmarking a New Bayesian Disease Mapping Cluster Detection Method, Nema Dean, Duncan Lee, Craig Anderson.
Discussant: *Rainer Dangl*.

16.10 - 17.30 **SP12: New trends and applications of alternating least squares in data analysis** - Chairs: *Y. Mori, K. Adashi* (Room III)

A general method for fuzzy partitioning and component analysis, Maria Brigida Ferraro, Paolo Giordani, Maurizio Vichi.
The multivariate power method: a fast iterative algorithm for repeated dimension reduction, Alfonso Iodice D'Enza, Michel van de Velden, Patrick J.F. Groenen.

On acceleration methods for Alternating Least Squares algorithm, Michio Sakakihara, Msahiro Kuroda, Yuichi Mori, Msaya IIzuka.
Canonical correlation analysis for three-mode three-way data, Jun Tsuchida, Hiroshi Yadohisa.

16.10 - 17.30 **SP13: Analysis of multivariate longitudinal data** - Chair: *P. Giordani* (Room 12)

From mixtures of SEMs to mixtures of Double-structure SEMs, Francesca Martella, Marco Alfò, Paolo Giordani.
Analysis of Multivariate Longitudinal Data Subject to Nonrandom Dropout, Mai Sherif Hafez, Irini Moustaki, Jouni Kuha.
Two-mode K-Spectral Centroid analysis for studying multivariate dynamical processes, Joke Heylen, Iven Van Mechelen, Eiko Fried, Eva Ceulemans.
Rotation involving subsets of variables to achieve an insight into structures of within-person variability, Marieke Timmerman, Eva Ceulemans, Henk Kiers.

16.10 - 17.30 **CONTR16: Symbolic Data** - Chair: *R. Verde* (Room 21)

Regime change analysis of interval-valued time series with an application to PM10, Carmela Cappelli, Pierpaolo D'Urso, Francesca Di Iorio.
Two-sample test with distributional data and detection of differential DNA methylation, Yusuke Matsui, Teppei Shimamura.
Detection of singularities in distribution valued data, Masahiro Mizuta, Hiroyuki Minami.
Interval-valued logistic regression ensemble vs noisy variables and outliers, Marcin Pelka, Aneta Rybicka, Justyna Brzezinska-Grabowska.

16.10 - 17.30 **CONTR17: Non gaussian mixture models and model selection** - Chair: *C. Viroli* (Room 31)

Mixtures of Hidden Truncation Hyperbolic Distributions, Ryan Browne.
Parsimonious multiple scaled mixtures, Brian C. Franczak, Ryan P. Browne, Paul D. McNicholas.
Mixtures of Coalesced Generalized Hyperbolic Distributions, Cristina Tortora.

17.30 - 19.00 **IFCS Council Meeting** (Statistics Library)

PRIN MISURA workshop - Chair: *P. Giudici* (Room III)

Modeling Contagion and Systemic Risk, Roberto Casarin.
Correlated stochastic processes for systemic risk, Laura Parisi.
Spillover Effects in the Banking Sector: CoVaR and Vine-CoVaR Approaches, Andrea Ugolini
Indeterminacy, misspecification and forecastability: Good luck in bad policy?, Luca Fanelli.
Reliable robust regression diagnostics, Silvia Salini.
Parsimonious representations of the prior in finite mixture regression models for multivariate mixed responses, Marco Alfò.

18.45 Bus Departure for **Conference Dinner** at **Palazzo Abergati**

Wednesday, 8th July 2015

9.00 - 10.00 Parallel Sessions

SP14: Mixture models in Biology - Chair: *S. Ingrassia* (Room I)

A simple approach to bi-clustering discrete data, Marco Alfò, Maria Francesca Marino, Francesca Martella.
Modeling mixtures in genomics, Jeanine Houwing-Duistermaat.
Mixture model with multiple allocations for clustering spatially correlated gene expression data, Saverio Ranciati, Cinzia Viroli, Ernst Wit.

9.00 - 10.00 **SP15: Statistical Models for ordinal data**, - Chair: *S. Cagnone* (Room III)

A finite mixture IRT model for ordinal responses with nonignorable missingness, Silvia Bacci, Francesco Bartolucci, Leonardo Grilli, Carla Rampichini.
Dealing with Large Heterogeneity in Sample Surveys, Stefania Capecchi, Domenico Piccolo.
Partial Possibilistic Regression Path Modeling, Rosaria Romano

9.00 - 10.00 **SP16: Joint Multivariate Data Reduction**, - Chairs: *A. Iodice D'Enza, M. Van de Velden* (Room 12)

Cluster Correspondence Analysis, Alfonso Iodice D'Enza, Michel van de Velden, Francesco Palumbo.
Clustering and Dimensional Reduction for mixed variables, Henk Kiers, Donatella Vicari, Maurizio Vichi.
A simultaneous analysis of dimension reduction and clustering with correlated error variables, Michio Yamamoto.

9.00 - 10.00 **SP17: Accuracy and validation in clustering and scaling models II**, organized by J.F.Vera - Chair: *M. Comas* (Room 21)

Stability analyses in human exposure to background air pollution in urban environments, Álvaro Gómez-Losada, José F. Vera-Vera.
An extension of the Adjusted Rand Index for fuzzy partitions, Sonia Amodio, Antonio D'Ambrosio, Carmela Iorio, Roberta Siciliano.
An integrated formulation for merging mixture components based on posterior probabilities, Marc Comas-Cufí, Josep Antoni Martń Fernández, Glòria Mateu Figueras.

9.00 - 10.00 **CONTR18: Clustering validation** - Chair: *T. Lee* (Room 31)

DESPOTA: a permutation test algoritm to detect a partition from a dendrogram, Dario Bruzzese, Domenico Vistocco.
On a Comprehensive Metadata Framework for Artificial Cluster Data Generation, Rainer Dangl, Friedrich Leisch.
Benchmarking cluster algorithms for ordinal survey data, Dominik Ernst, Friedrich Leisch.

10.15 - 11.00 **Plenary invited lecture: Modern multivariate data analysis through monitoring**, Marco Riani - Chair: *Andrea Cerioli* (Room I)

11.00 - 11.30 Coffee Break

11.30 - 12.50 Parallel Sessions

SP18: IASC Session on Bilinear models and regularization - Chair: *P. Groenen* (Room I)

Generalized Additive Models (GAMs) via Bayesian P-splines using INLA, Cajo ter Braak, María Xosé Rodríguez-Álvarez, Martin Boer, Paul Eilers, Havard Rue.

Multinomial correspondence analysis, Patrick J.F. Groenen, Julie Josse.
Three-way data analysis with clustered bilinear models, Pieter Schoonees.
Regularized Generalized Canonical Correlation analysis for Multiway data, Arthur Tenenhaus, Laurent Le Brusquet.

11.30 - 12.50 **SP19: Multivariate Visualization and Discrimination** - Chair: *S. Lubbe* (Room III)

Biplot-based visualizations in latent class modelling, Zsuzsa Bakk, Niël Le Roux, Jeroen Vermunt.
Small sample multi-label discriminant analysis, Nelmarie Louw.
Feature selection and kernel specification for support vector machines using multi-objective genetic algorithms, Martin Philip Kidd, Martin Kidd, Surette Bierman.
Measures of fit for nonlinear biplots, Karen Vines.

11.30 - 12.50 **CONTR19: Mixture Models** - Chair: *P. Schlattmann* (Room 12)

Mixture simultaneous factor analysis for modeling structural differences in multivariate multilevel data, Kim De Roover, Jeroen K. Vermunt, Marieke E. Timmerman, Eva Ceulemans.
A multiple clusterings model based on Gaussian mixture, Andrea Pastore.
Criteria for model selection in model-based clustering, Fumitake Sakaori.

Partially Supervised Biclustering of Gene Expression Data with Applications in Nutrigenomics Biomarker Discovery, Monica Hiu Tung Wong.

11.30 - 12.50 **CONTR20: Bayesian Clustering** - Chair: *J. Vermunt* (Room 21)

Bayes Clustering Operators for Random Labeled Point Processes, Lori Dalton.
Bayesian clustering: a novel nonparametric framework for borrowing strength across populations, Antonio Lijoi, Bernardo Nipoti, Igor Pruenster.
Dirichlet process Bayesian clustering: an application to survival data, Silvia Liverani.
Optimizing Mailing Decisions Based on a Mixture of Dirichlet Processes, Nadine Schröder, Harald Hruschka.

11.30 - 12.50 **CONTR21: Longitudinal and Spatial Data** - Chair: *J. G. Dias* (Room 31)

Latent class modeling of markers of day-specific fertility, Francesca Bassi, Bruno Scarpa.
Attitudes to maternity in Poland - a longitudinal analysis based on latent Markov models with covariates, Ewa Genge, Joanna Trzesiok.
Space-time clustering for radiation monitoring post data based on hierarchical structure, Fumio Ishioka, Koji Kurihara.
Regional Spatial Moving Average and new Spatial Correlation Coefficient, Andrzej Sokolowski, Malgorzata Markowska, Marek Sobolewski, Danuta Strahl, Sabina Denkowska.

12.50 - 14.00 Lunch Break

14.00 - 14.45 **Plenary invited lecture: Removing unwanted variation for classification and clustering**, Terry Speed - Chair: *Geoff McLachlan* (Room I)

14.45 - 15.30 **Poster Session**

Unimodal Logistic Discrimination, Joaquim Costa, A. Rita Gaio.
High-dimensional regression mixture models to perform clustering - application to electricity dataset, Emilie Devijver, Jean-Michel Poggi, Yannig Goude.
One class classification based on transvariation probability, Francesca Fortunato.
Quantified SWOT method and its use in assessing the financial situation of local administrative units, Romana Glowicka-Woloszyn, Aleksandra Luczak, Andrzej Woloszyn.
Football and the dark side of cluster analysis, Christian Hennig, Serhat Akhanli.
Segmentation of Online Consumer Behaviour of Polish Youth: a Means-End Approach, Anna Mirosława Myrda.
Classification Methods in the Research on the Financial Standing of Construction Enterprises after Bankruptcy in Poland, Barbara Pawełek, Jadwiga Kostrzewska, Artur Lipieta, Maciej Kostrzewski, Krzysztof Gałuszka.
Clustering and Classification methods for an experimental study on prion diseases, Giorgia Rocco.
Repeated measures analysis for functional data using two-cumulant approximation - with applications, Łukasz Smaga.
Coping with poor information for classifiers to support business decision-making. A case study, Annalisa Stacchini.
Analysis of Quality of Life among Hemophiliac Patients Using Scores of Medical Outcomes Study, Shinobu Tatsunami.

15.30 - 16.15 Coffee Break

16.15 - 17.35 Parallel Sessions

SP20: What's the best cluster analysis method? - Chair: *C. Hennig* (Room I)

Progress and Open Problems in Clustering: Beyond Algorithm Development, Margareta Ackerman.
Information-Theoretic Validation of Clustering Algorithms, Joachim M. Buhmann.
Averaging and Asymmetry in Cluster Analysis, Paul McNicholas.
Disentangling Continuous and Discrete Structure Within Data, Doug Steinley.

16.15 - 17.35 **CONTR22: Longitudinal Data** - Chair: *S. Bianconcini* (Room III)

Financial technical analysis using hidden Markov models, José G. Dias.
Power Analysis for the Likelihood Ratio Test in Latent Markov Models: Short-cutting the bootstrap p-value based method, Dereje W. Gudicha, Verena D. Schmittmann, Fetene B. Tekle, Jeroen K. Vermunt.
Clusterwise three-way component models to account for heterogeneity in three-way data, Tom F Wilderjans, Eva Ceulemans.

16.15 - 17.35 **CONTR23: Functional Data II** - Chair: *A. Luati* (Room 12)

Statistical data depth for clustering macroseismic fields, Claudio Agostinelli, Renata Rotondi, Elisa Varini.
Robust model-based functional clustering of satellite data, Carlo Gaetan, Paolo Girardi, Roberto Pastres.
Functional Data Analysis for the investigation of longitudinal response patterns in health care, Sugnet Lubbe, Felix Dube, Mark Nicol, Heather Zar.

Clustering word life-cycles in chronological corpora: what data transformation for differing clustering goals, Matilde Trevisani, Arjuna Tuzzi.

16.15 - 17.35 **SP21: New developments in Classification and Regression Trees** - Chair: *E. Dusseldorp* (Room 21)

Estimating the average treatment effect through Balancing Recursive Partitioninng, Claudio Conversano, Massimo Cannas, Francesco Mola.
A framework for measuring the stability of recursive partitioning results, Michel Philipp, Thomas Rusch, Kurt Hornik, Carolin Strobl.
Combining model-based recursive partitioning and random-effects estimation for the detection of treatment subgroups, Marjolein Fokkema.
Meta-Cart Integrating Classification and Regression Trees into Meta-analysis, Xinru Li, Elise Dusseldorp, Jacqueline Meulman.

16.15 - 17.35 **CONTR24: Supervised Learning II** - Chair: *S. Ingrassia* (Room 31)

Real-time classification of in-flight aircraft damage, Brenton S. Blair, Herbert K. H. Lee.
The δ-machine, Mark De Rooij.
The Five Factor Model of personality and evaluation of drug consumption risk, Elaine Fehrman, Awaz K. Muhammad, Evgeny Mirkes, Vincent Egan, Alexander N. Gorban.
K-NN controlled condensation: a new method for data preprocessing in classification tasks, Carmen Villar-Patiño, Carlos Cuevas-Covarrubias.

17.40 **Closing ceremony** (Room I)

Contents

Abstracts

Generalized additive models: a retrospective

Robert Tibshirani

Abstract I will give an informal talk on the history of generalized additive models. I will discuss how Trevor Hastie and I conceived of the idea, and important influences on our work. Finally I will touch on some connections to our more recent work on sparsity.

Robert Tibshirani
Stanford University, USA, e-mail: tibs@stanford.edu

GAM selection via convex optimization

Trevor Hastie and Alexandra Chouldechova

Abstract While smoothing and additive models were the rage in the 80s and 90s, convex optimization is one of the present-day tools of choice - the lasso and its relatives induce sparsity in models. In this talk we describe a family of penalties that induce the right kind of sparsity in generalized additive models: from zero, to linear, to nonlinear.

Trevor Hastie
Stanford University, USA, e-mail: hastie@stanford.edu

Alexandra Chouldechova
Carnegie Mellon University, USA, e-mail: achould@cmu.edu

Benchmarking in cluster analysis: Preview of a white paper

Douglas L. Steinley, Iven Van Mechelen, and IFCS Task Force on Benchmarking

Abstract To achieve scientific progress in terms of building a cumulative body of knowledge, careful attention to benchmarking is of the utmost importance. This means that proposals of new methods of data preprocessing, data-analytic techniques and algorithms, and heuristics for model selection should be extensively and carefully compared with existing alternatives. To date, benchmarking and recommendations for benchmarking have been frequently seen in the context of supervised learning. Yet, unfortunately, there has been a dearth of guidelines for the unsupervised setting. To address this problem, an IFCS Task Force is currently preparing a white paper on benchmarking in cluster analysis. In this paper discussion is given to fundamental conceptual underpinnings, and attention is paid to both the best and worst practices in the field, in the context of presenting and analyzing simulation data as well as empirical data; to conclude, foundational recommendations are made. In the present talk we will offer a sneak preview of this white paper.

Keywords

benchmarking; cluster analysis; white paper

Doug Steinley
University of Missouri, USA, e-mail: steinleyd@missouri.edu

Iven Van Mechelen
University of Leuven, Belgium, e-mail: iven.vanmechelen@ppw.kuleuven.be

What are the true clusters?

Christian Hennig

Abstract In much of the literature on cluster analysis there is the implicit assumption that in any situation in which cluster analysis is applied, there are some "true" clusters at which the analysis aims; and usually the "true" clustering is assumed to be unique. Benchmarking of clustering algorithms usually is based on datasets with some assumed truth, so that it can be seen how well this truth is recovered by the algorithms. I will argue that there are several legitimate clusterings on the same data and that defining "true" clusters is highly problematic. I will discuss a number of related issues: philosophical background, constructive and realist aims of clustering, and various ways to define "true clusters", namely based on the data alone, on an underlying true class variable, or on probability models. Implications for cluster benchmarking and variable selection in clustering are also mentioned.

Keywords

cluster benchmarking; clustering aims; cluster validation

References

HENNIG, C. (2015): What are the true clusters? arXiv:1502.02555

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

Using principles of optimal treatment regime estimation in simulation studies for benchmarking

Lisa Doove, Iven Van Mechelen, Tom Wilderjans, and Antonio Calcagnì

Abstract In benchmarking studies with simulated data sets in which two or more methods are compared, over and above the search of a universally winning method, one may investigate how the winning method may vary over patterns of characteristics of the data and the data-generating mechanism. Interestingly, this problem bears strong formal similarities to the problem of looking for optimal treatment regimes in biostatistics when two or more treatment alternatives are available for the same medical problem or disease. In that case, one may wish to induce from empirical data a rule that indicates which types of patients should preferably receive which treatment. The optimal rule or treatment regime then is the one that yields the highest expected (potential) outcome if the rule would be applied to the entire population of patients under study. In this talk, we will outline how rules for optimal calling in methods can be derived from benchmarking studies with simulated data by means of a classification tree method that is based on principles of optimal treatment regime estimation. We will illustrate by means of analyses of data from a benchmarking study to compare two different algorithms for the estimation of a two-mode additive clustering model.

Keywords

benchmarking; classification trees

Lisa Doove
University of Leuven, Belgium, e-mail: Lisa.Doove@ppw.kuleuven.be

Iven Van Mechelen
University of Leuven, Belgium, e-mail: Iven.VanMechelen@ppw.kuleuven.be

Tom Wilderjans
University of Leuven, Belgium, e-mail: Tom.Wilderjans@ppw.kuleuven.be

Antonio Calcagnì,e-mail: Antonio.Calcagni@unitn.it

Comparison of Classification Accuracy for Corporate Bankruptcy Prediction Models Including Changes in Economic Environment

Mateusz Baryla, Barbara Pawelek, and Jozef Pociecha

Abstract Many types of corporate bankruptcy prediction models have been formulated by the business theory and practice. Among them, a wide group is composed of classification models, which can divide the population of firms into two groups: bankrupts and non-bankrupts. The aim of the paper is to present the outcomes of the comparative analysis of classification accuracy for selected kinds of corporate bankruptcy prediction models. While building models, both the companys internal financial factors (financial ratios) and variables which reflect changes in economic environment are taken into account. The analysis is based on data concerning Polish manufacturing companies from 2005 to 2009. The following four types of bankruptcy prediction methods are employed: logistic regression, discriminant function, classification trees and neural networks. In order to assess the classification accuracy of a model for a training set and test set, among other things, three measures are used, i.e.: sensitivity, specificity and AUC. The bootstrap technique and simple random sampling without replacement are also implemented in the study.

Keywords

classification models; classification accuracy; corporate bankruptcy

Mateusz Baryla
Cracow University of Economics, Poland, e-mail: mateusz.baryla@uek.krakow.pl

Barbara Pawelek
Cracow University of Economics, Poland, e-mail: barbara.pawelek@uek.krakow.pl

Jozef Pociecha
Cracow University of Economics, Poland, e-mail: jozef.pociecha@uek.krakow.pl

References

BELLOVARY, J.L., GIACOMINO, D.E. and AKERS, M.D. (2007): A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education*, 33(4), 3–41.

DE LEONARDIS, D. and ROCCI, R. (2014): Default Risk Analysis via a Discrete-time Cure Rate Model. *Applied Stochastic Models in Business and Industry*, 30(5), 529–543.

PAWEŁEK, B., POCIECHA, J. and BARYŁA, M. (2015): Dynamic Aspects of Bankruptcy Prediction Logit Model for Manufacturing Firms in Poland. Proceedings of the *Second European Conference on Data Analysis* (ECDA 2014), Springer, tentatively accepted.

Student Value Systems

Andreas Geyer-Schulz, Thomas Hummel, Claire Roederer, and Victoria-Anne Schweigert

Abstract We report on a survey of student value systems conducted at the Karlsruhe Institute of Technology (Germany) and EM Strasbourg Business School (France) in the winter term 2014/15 which combines constructs from Rokeach's value survey, Mitchell's values and life-style (VALS) as well as Kahle's list of values (LOV). All constructs have been evaluated in a pre-test and are available in English, French, and German. Methodologically, we concentrate on a comparison of the different operationalizations of the same latent constructs in the instruments presented above: Rokeach's value survey is based on ranking, Mitchell's VALS survey on rating data, and Kahle experimented with both types of operationalization.

Keywords

value systems; latent constructs; partial order;

Andreas Geyer-Schulz
Information Services and Electronic Markets, Institute of Information Systems and Marketing, (IISM), Karlsruhe Institute of Technology (KIT), Germany, e-mail: andreas.geyer-schulz@kit.edu

Thomas Hummel
Information Services and Electronic Markets, Institute of Information Systems and Marketing, (IISM), Karlsruhe Institute of Technology (KIT), Germany, e-mail: Thomas.Hummel@kit.edu

Claire Roederer
EM Strasbourg Business School, France, e-mail: claire.roederer@em-strasbourg.eu

Victoria-Anne Schweigert
Information Services and Electronic Markets, Institute of Information Systems and Marketing, (IISM), Karlsruhe Institute of Technology (KIT), Germany, e-mail: Victoria-Anne.Schweigert@kit.edu

References

KAHLE, Lynn R. (1983): *Social Values and Social Change: Adaption to Life in America*. Praeger, New York.

MITCHELL, A. (1983): *The Nine American Life Styles: Who We Are and Where We're Going*. Macmillan, New York.

ROKEACH, M. (1973): *The Nature of Human Values*. Free Press, New York.

Application of Extreme Value Theory in the measurement of household investment risk

Krzysztof Jajuga

Abstract Statistical data analysis methods have been applied in many financial problems, particularly in the financial market analysis, banking and corporate finance. It is observed growing interest in the application of data analysis methods in household financial decisions and financial planning, including household investments. In the paper it is presented the framework for the analysis of risk of household financial investments. The particular attention is paid to extreme risk,resulted from the events which have very low probability of occurrence but lead to very large losses. The elements of Extreme Value Theory are applied to evaluate the risk of household investments. Given several models of investment processes, the extreme risk measures, like Expected Shortfall or Median Shortfall are derived. Theoretical considerations are exemplified for typical households.

Keywords

extreme risk; financial investments

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

Application of ordered multinomial logit model to identification of determinants of LAU2's financial condition

Andrzej Pawel Woloszyn, Romana Glowicka-Woloszyn, and Feliks Wysocki

Abstract As part of the public finance sector, local finances are charged with tasks of local interest, which include addressing essential needs of the community, fostering local development of technical and social infrastructure, ensuring public safety and social order, implementing land-use plans, and attending to ecological issues. Under extant social and economic conditions carrying out these tasks depends on the amount of revenues collected from own and external sources, and on efficiency of their expenditure, which together translates into financial condition of local administrative units (LAUs). Multidimensional analysis of the condition concerns Polish LAU2 units, or communes, and comprises two steps. First, communes' financial condition is evaluated synthetically using TOPSIS method with selected indicators. Next, determinants of the condition are examined with ordered multinomial logit model. The model's dependent variable is discreet with values corresponding to the typological classes of financial condition already determined in step one, while the independent variables represent the communes' local development conditions. Calculations were performed using the STATA program. The study's source data comprise financial indicators published by Polish Ministry of Finance and Local Data Bank of the Central Statistical Office on 2479 communes from 2005-2013.

Keywords

ordered multinomial logit model, TOPSIS, financial condition of communes

Andrzej Pawel Woloszyn
Poznan University of Life Sciences, Poland, e-mail: andrzej.p.woloszyn@gmail.com

Romana Glowicka-Woloszyn
Poznan University of Life Sciences, Poland, e-mail: roma@up.poznan.pl

Feliks Wysocki
Poznan University of Life Sciences, Poland, e-mail: wysocki@up.poznan.pl

IRT-based conjoint analysis in the optimization of banking products

Justyna Brzezińska-Grabowska, Aneta Rybicka, and Adam Sagan

Abstract Conjoint measurement and analysis have common underlying psychometric and statistical assumption concerning axioms of additivity and two-way frame of reference in preference measurement. However, whereas the former concept is widely used in fundamental measurement of subject $\times$ object dominance structures as in IRT and Rasch measurement models, the latter is utilized in a broad family of object $\times$ object dominance structures in both compositional (i.e. Thurestone case III and V) as well as decompositional (classical conjoint experiments and BTL/alpha simulation) preference measurement models. These two traditions are rarely combined in one measurement model and research design that integrates subject $\times$ object $\times$ object measurement (Neubauer 2001). The aim of the paper is to adopt and compare three types of preference measurement models in the area of banking products in Poland:1/ paired-comparisons and rating scale conjoint experiment,2/ IRT-based conjoint (Rasch and Birnbaum politomous models) and 3/ compositional Thurstone III/V models (Bockenholt 2006). Part-worth utilities are used for product optimization and comparison across the estimated models.

Keywords

conjoint measurement; item response theory; banking products

Justyna Brzezińska-Grabowska
University of Economics in Katowice, Poland, e-mail: justyna.brzezinska-grabowska@ue.katowice.pl

Aneta Rybicka
Wroclaw Economic Uniwersity, Poland, e-mail: aneta.rybicka@ue.wroc.pl

Adam Saganat
Cracow Economic University, Poland, e-mail: sagana@uek.krakow.pl

References

BOCKENHOLT, U. (2006): Thurstonian Based Analysis: Past Present and Future Utilities. *Psychometrica*, 71(4), 615–629.

NEUBAUER, G. (2003). *An IRT-Approach for Conjoint Analysis*. In: A. Ferligoj and A. Mrvar (Eds.): Developments in Applied Statistics. Metodoloski zvezki, 19, Ljubljana, 35–47.

Accounting for attribute non-attendance in subset-conjunctive choice models

Michel Meulders

Abstract Standard choice models assume that respondents examine all attributes and all alternatives across all choice tasks in the same fully compensatory manner. However, research has indicated that respondents often resort to screening rules if the choice task is considered too complicated: In a first stage respondents may identify a consideration-set of alternatives that needs further evaluation (i.e. consideration-set screening) or eliminate attributes they find irrelevant (i.e. attribute non-attendance). In a second stage they make a choice using a standard compensatory model on the outcome of the screening-stage. Probabilistic t subset-conjunctive models assume that respondents include an alternative as part of the consideration set if it is acceptable on at least t attribute levels. The involved screening process is probabilistic as respondents are assumed to classify the attribute levels of the alternatives in each choice set as acceptable/unacceptable with a certain (possibly class-specific) probability. As existing subset-conjunctive models only account for consideration-set screening, we will extend these models to also account for attribute non-attendance. This extension is important as it may lead to models with improved fit and predictive accuracy. As an illustration, the models are applied to analyze preferences of students for student rooms.

Keywords

subset-conjunctive model; screening rule; attribute non-attendance

Michel Meulders
KU Leuven, Belgium, e-mail: michel.meulders@kuleuven.be

References

HENSHER, D., ROSE, J. and GREENE, W., (2005): The Implications on Willingness to Pay of Respondents Ignoring Specific Attributes. *Transportation*, 32(3), 203–222.

JEDIDI, K. and KOHLI, R. (2005): Probabilistic Subset Conjunctive Models for Heterogeneous Consumers. *Journal of Marketing Research*, 17, 483–494.

KOHLI, R. and JEDIDI, K. (2005): Probabilistic subset conjunction. *Psychometrika*, 70(4), 737–757.

An adaptive item selection method for curtailment

Niels Smits

Abstract Health questionnaires are often built up from sets of questions which are totaled to obtain a sum score; often, this score is subsequently used to classify respondents. An important consideration in designing questionnaires is to minimize respondent burden. Curtailment is an efficient method of questionnaire administration aimed at classification into categories, such as 'at risk' and 'not at risk'. Curtailment uses a prediction model for forecasting observed class membership; the strategy is to stop testing when not yet administered items are unlikely to change the respondent's classification. The item administration of curtailment is static, i.e., uses the same sequence of items for all respondents, and a dynamic item selection could increase efficiency. The current paper uses a method for adaptive item selection which stems from Data Mining. The item selection method will be studied using several real data sets. Benefits and limitations of this new methodology are discussed.

Niels Smits
Department of Methods Faculty of Psychology and Education VU University Amsterdam, The Netherlands, Netherlands, e-mail: n.smits@vu.nl

A Random Matrix Theory study of Covariance Matrices of Continuous and Categorical Variables

Graziano Vernizzi and Miki Nakai

Abstract A novel approach to the analysis of social survey data sets based on Random Matrix Theory (RMT) is presented. The statistical analysis of social survey data often requires the use of correlation (or covariance) matrices. However, the empirical determination of correlation matrices from sociological survey data is typically perturbed by statistical noise. RMT is capable of modelling analytically the spectrum of a pure random matrix describing a finite time series of strictly uncorrelated variables. Deviations from the random matrix case hint to the presence of true information in the correlations. Therefore, it has been shown in the literature that the spectrum of a random covariance matrix can be used to distinguish between the noisy component and the information component. Here it is suggested using RMT to the analysis of survey data by utilizing a recent unified geometrical description of the measure of association between data of heterogeneous type (i.e. where both continuous and categorical data are concurrently present). A discussion with explicit examples of how RMT can be applied to the variance/covariance analysis of sociological surveys on social stratification and mobility, within such a novel geometrical framework, is provided.

Keywords

Random Matrix Theory; Covariance matrix; Social survey data

Graziano Vernizzi
Department of Physics and Astronomy, Siena College, 515 Loudon Road, Loudonville, NY 12211, USA, e-mail: gvernizzi@siena.edu

Miki Nakai
Department of Social Sciences, College of Social Sciences, Ritsumeikan University, Tojiinkita-machi, Kita-ku Kyoto, Japan, e-mail: mnakai@ss.ritsumei.ac.jp

References

AKEMANN, G., BAIK J., and DI FRANCESCO P. (2011): *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, Oxford, Eds.

VERNIZZI, G. and NAKAI, M (2015): A Geometrical Framework for Covariance Matrices of Continuous and Categorical Variables, *Sociological Methods and Research*, 44, 48–79.

Large covariance matrix estimation by composite minimization

Matteo Farnè

Abstract A method to regularize large-dimensional covariance matrices under the assumption of approximate factor model will be presented. Existing methods perform estimation by recovering principal components and sparsifying the residual covariance matrix. In our setting this task is achieved recovering the low rank plus sparse decomposition by least squares minimization under nuclear norm plus l_1 norm penalization. In the literature, the best known algorithm to solve this problem is soft thresholding plus singular value thresholding and consistency of estimators is derived under specific assumptions on the eigenvalues of the low rank component matrix. In this paper consistency of the proposed estimator will be derived under the pervasive condition, providing the identification of low rank and sparse spaces by introducing the unshrinking of estimated eigenvalues. Algorithm derivation and convergence analysis are provided, and the new procedure is compared with the existing ones under the same assumptions. The performance of our minimizer is described in a wide simulation study, where various low rank plus sparse settings are simulated according to different parameter values.

Keywords

regularization; nuclear norm; unshrinking

References

FAN,J., LIAO, Y. MINCHEVA M. (2013): Large Covariance Estimation by Thresholding Principal Orthogonal Complements (with discussion). Journal of Royal

Matteo Farnè
University of Bologna, Italy, e-mail: matteo.farne2@unibo.it

Statistical Society B, 75, 603–680.
LUO, X. (2013): Recovering Model Structures from Large Low Rank and Sparse Covariance Matrix Estimation. *arXiv.org*

Robust regularized discriminant analysis based on implicit weighting

Jan Kalina

Abstract Standard classification procedures are sensitive to the presence of outlying measurements in the data. We propose a new robust method based on a regularized version of the minimum weighted covariance determinant estimator. The method is suitable for data with the number of variables exceeding the number of observations. The method is based on implicit weights assigned to individual observations. Our approach is a unique attempt to combine regularization and high robustness, allowing to down-weight outlying high-dimensional observations. Classification performance of new methods and some ideas concerning classification analysis of high-dimensional data are illustrated on real raw data as well as on data contaminated by severe outliers.

Keywords

robust classification; breakdown point; regularization

Jan Kalina
Institute of Computer Science CAS Dept. of Medical Informatics and Biostatistics Pod Vodarenskou vezi 2 182 07 Praha 8 Czech Republic, e-mail: kalina@cs.cas.cz

Revisit on Big Data Analysis Framework - How to Make Proper 'Mini data'? -

Hiroyuki Minami and Masahiro Mizuta

Abstract In the era of Big Data, the amount of raw data has been rapidly growing over hundreds of tera bytes. However, due to technical limitation, most popular analytic tools may not handle them straightforwardly.In addition, we empirically realize that raw big data have messy information.Thus, toward practical analysis and proper interpretation, an adequate reduction should be applied to them. We call the reduced intermediate data 'Mini data', compared with big data, and discuss the features focused on the size and the quality. We can formulate how to make them by try and error, including random sampling, data-mining techniques and statistical approaches. To carry them out sufficiently, we need an environment that all raw data are stored, handled in real time, and also has much computing functions regarding information retrieval (represented by SQL language) and statistics. We introduce our implementation in cloud based on 'Mini data' idea and offer some practical examples with the big datasets, air dose radiation rates in Japan including Fukushima Prefecture.

Keywords

Cloud computing; Massive data handling

Hiroyuki Minami
Hokkaido University, Japan,e-mail: min@iic.hokudai.ac.jp

Masahiro Mizuta
Hokkaido University, Japan, e-mail: mizuta@iic.hokudai.ac.jp

References

H. MINAMI and M. MIZUTA (2014): Big Data Oriented Symbolic Data Analysis in Cloud. In Second European Conference on Data Analysis (ECDA 2014), 130.
W. MCKINNEY (2013): *Python for Data Analysis*. O'Reilly.

Determining the number of clusters from decompositions of a Rand index-based measure of partitioning stability

Patrice Bertrand, Lassad El Moubarki, and Ghazi Bel Mufti

Abstract A number of cluster validity indices aim to evaluate a clustering structure by measuring its degree of adequacy with the examined data set. Geometric or density-based assumptions are then generally required to be verified by the data set being considered, which leads to restrict the relevance of such cluster validity indices. In order to avoid this drawback, different cluster validity indices aim to assess a clustering structure simply on the basis of its stability on perturbed data sets. Along this line, we propose different decompositions of a Rand index-based measure of stability. The obtained components of these decompositions provide estimates of the cohesion, isolation and validity of the clusters of the partition being assessed. We then derive different stability-based indexes which aim to determine the optimal number of clusters of a data set partitioning. Based on experimentations achieved both on simulated and real data sets, these indexes are then evaluated by comparing their results with those obtained by well known methods for predicting the number of clusters.

Keywords

Partitioning stability; Adjusted Rand index; Resampling

Patrice Bertrand
Université Paris-Dauphine, France, e-mail: bertrand@ceremade.dauphine.fr

Lassad El Moubarki
Université de Sfax, Tunisia, e-mail: elmoubarki.lassad@yahoo.fr

Ghazi Bel Mufti
EESEC-Tunis, Tunisia, e-mail: belmufti@yahoo.com

References

BEN-DAVID, S., VON LUXBURG, U. (2008): Relating clustering stability to properties of cluster boundaries. In: R. Servedio and T. Zhang (Eds.): *Proceedings of the 21st Annual Conference on Learning Theory*. Springer, Berlin, 379–390.

BEN-HUR, A., ELISSEEFF, A., GUYON, I. (2002): A stability based method for discovering structure in clustered data. In: *Proc. Pacific Symposium on Biocomputing*, 6–17.

HENNIG, C. (2007): Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 51(1), 258–271.

On comparing partitions

Marjan Cugmas and Anuka Ferligoj

Abstract Rand (1971) proposed the Rand Index to measure the stability of two partitions of one set of units. Hubert and Arabie (1985) corrected the Rand Index for chance (Adjusted Rand Index). In this paper, we present some alternative indices. The proposed indices do not assume one set of units for two partitions. Here, one set of units can be a subset of the other set of units. According to the purpose of the comparison of two partitions, the merging and splitting of clusters in two partitions can have different impact on the value of the indices. Therefore, we proposed different modified Rand Indices.

Keywords

Rand Index; Adjustment for chance; comparing partitions

Marjan Cugmas
Researcher at Centre for Methodology and Informatics, Faculty of Social Sciences, University of Ljubljana, e-mail: marjan.cugmas@fdv.uni-lj.si

Anuka Ferligoj
Professor, Faculty of Social Sciences, University of Ljubljana, e-mail: anuska.ferligoj@fdv.uni-lj.si

Set valued prototypes through the consensus analysis

Mario Fordellone and Francesco Palumbo

Abstract Let X be a generic $N \times J$ data matrix, where each row represents a statistical unit described by J features, and let U be a set of descriptions $U = \{u_1, u_2, \ldots, u_K\}$ in the feature space, prototyping consists in defining a rule that associates each row of X to the elements of U. Given the matrix Y of order $N \times K$, under the general constraint that $\sum_{k=1}^{K} y_{ik} = 1$, in the fuzzy logic the general element of Y $y_{ik} = [0,1]$ represents the membership degrees of the row vectors x_i to the descriptions u_k, the crisp membership function assumes that $y_{ik} = \{0,1\}$. Fuzzy c-means and archetypal analysis permits to define fuzzy partitions over the data: the former seeks K homogenous groups with respect to their barycentres, whereas, the latter identifies a set of extreme points, called archetypes, and creates a group around each archetype. When more than one partition can be defined on the same data, consensus analysis has been proposed with a twofold aim: (i) to find an unique partition solution as synthesis of all partitions; (ii) to measure the agreement among the different partitions and between the synthesis and all the partitions. Using the consensus analysis, set valued prototypes are defined as the compromise between the Fuzzy c-means and archetypes partitions.

Keywords

Prototyping; Fuzzy c-Means; Archetypal Analysis; Consensus Analysis

Mario Fordellone
University of Padua, Italy, e-mail: fordellone@stat.unipd.it

Francesco Palumbo
Federico II University of Naples, Italy, e-mail: fpalumbo@unina.it

References

HASTIE, T., et al (2009): *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer;

CUTLER, A., and BREIMAN L. (1994): Archetypal analysis. *Technometrics* 36.4 : 338–347;

HUBERT, L.; ARABIE, P. (1985): Comparing partitions. *Journal of classificatio*n, 2.1: 193–218.

Correction for chance and correction for maximum value

Matthijs J. Warrens

Abstract In data analysis and classification association coefficients are used to quantify the association in contingency tables. Various coefficients are chance-corrected, that is, they have value zero under statistical independence. Examples are the phi coefficient, Cohen's kappa, and the adjusted Rand index. Other coefficients are corrected for maximum value.Correction for chance and correction for maximum value are studied as functions on a space of association coefficients for contingency tables. Both functions are idempotent, and the two functions commute under composition. Furthermore, the composed function maps a coefficient and all its linear transformations given the marginal totals to the same coefficient. The algebraic structure of the two functions, the their composition, and the identity function, turns out to be an idempotent commutative monoid.

Matthijs J. Warrens
Leiden University, Netherlands, e-mail: warrens@fsw.leidenuniv.nl

Estimating Binary Spatial Autoregressive Models for Rare Events

Raffaella Calabrese and Johan A. Elkink

Abstract Policy adoptions, regime transitions, state failures, but also currency crises or credit and bank defaults are all rare events, and are all likely to be interdependent either geographically or through network connections such as credit lines. Furthermore, many of these events are rare. The binary and rare nature of the dependent variable and the spatial autoregressive structure, generate special challenges for the statistical estimation of models explaining these outcomes. The most used spatial regression models for binary dependent variable consider a symmetric link function (logit or probit functions). When the dependent variable represents a rare event, a symmetric link function is not coherent. We propose the quantile function of the Generalized Extreme Value (GEV) distribution as link function in a spatial generalized linear model and we call this model the Spatial GEV regression model. To estimate the parameters of such model, a Gibbs sampler is proposed that mergers existing samplers for rare events and for spatial autoregressive data. We analyse the performance of our model by Monte Carlo simulations and evaluating the prediction quality in empirical data on bank defaults.

Keywords

rare events; spatial autoregression; probit

Raffaella Calabrese
Essex Business School, University of Essex, United Kingdom, e-mail: rcalab@essex.ac.uk

Johan A. Elkink
University College Dublin, Ireland, e-mail: jos.elkink@ucd.ie

Stability Transmission and Risk of Islamic Banking Networks in the MENA Region

Shatha Qamhieh Hashem and Paolo Giudici

Abstract The aim of this paper is to empirically investigate the proposition that Islamic banking services can support financial stability and to examine this relation at the different levels of Islamic banking services used in the MENA region banking sector. The stability that Islamic banking transmits through the financial system will be gauged with a dual backward-forward modelling approach to capture the direct and indirect contagion effects that move along different transmission channels. To achieve this aim, we will first regress stock-market prices on systematic effects such as country and bank type. We then regress the same prices on balance sheet indicators selected in accordance with the CAMEL framework. Next we identify the banks that have the highest contagion impact, with a correlated ability to lead the financial system to breakdown as they fail. This will be performed using a graphical Gaussian model that will enable to distinguish the correlations between banks due to idiosyncratic characteristics inherited in the previous balance sheet indicators, from the correlations between countries that can be attributed to macroeconomic behaviour.

Keywords

Financial risk; Systemic risk; Systematic risk

Shatha Qamhieh Hashem
An-Najah National University, Nablus, Occupied Palestinian Territory, e-mail: shathaqamhieh@yahoo.com

Paolo Giudici
University of Pavia, Italy, e-mail: paolo.giudici@unipv.it

Discovering SIFIs in interbank community.

Alessandro Spelta

Abstract This paper proposes a new methodology to identify Systemically Important Financial Institutions (SIFIs) and to detect the community structure of the interbank network. The proposed technique evaluates the exposures of individual financial institutions in the communities they operate as well as the exposures that they place at the system-wide level. Using interbank transactions data from the e-Mid platform, we show that the systemic importance associated with Italian banks decreased during the 2007-2009 financial crisis while the opposite happened for foreign institutions. We show that, as the transaction volume grew, the number of communities raised as well. Moreover results indicate that, during the financial crisis period, banks strongly operate into non overlapping communities with few institutions playing the role of SIFIs. On the contrary during business as usual years banks act in different and overlapping modules.

Keywords

interbank market; systemically important financial institutions; community detection

References

Psorakis, Ioannis, et al. (2011). Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E* 83.6: 066114.

Cao, Xiaochun, et al. (2013). Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization. *Scientific reports*, 3 (2013).

Alessandro Spelta
Università Cattolica del Sacro Cuore, Italy, e-mail: alessandro.spelta@unicatt.it

Kleinberg, Jon M. (1999) Authoritative sources in a hyperlinked environment. *Journal of the ACM* (JACM) 46.5 (1999): 604–632.

On the choice of the number of groups in the context of model-based clustering

Laura Anderlucci and Geoff McLachlan

Abstract Determining the number of clusters in a data set is a difficult problem that has been widely discussed in the literature. Many methods have been proposed so far but an overall ranking is not possible, mostly because there is not always a clear definition of what a cluster is and because different clustering methods imply different criteria. This work considers some of the widely used methods to assess the number of groups in a model-based perspective, which also allow to statistically test a grouping structure. For a fair comparison we included an extension of the popular Clest method (Dudoit and Fridlyand 2002) to the context of model-based clustering. A simulation study together with results on real data are presented.

Keywords

model-based clustering; clest; likelihood ratio test; sigclust; gap

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

Geoff McLachlan
University of Queensland, Australia, e-mail: g.mclachlan@uq.edu.au

The joint role of trimming and constraints in robust estimation for mixtures of skew normal

Francesca Greselin, Agustin Mayo-Iscar, and Luis-Angel Garcia-Escudero

Abstract In recent years, asymmetric departures from normality observed in some real subpopulations, suggested the introduction of asymmetric components (such as the skew normal and the skew t) in the classical mixture model approach. Mixtures of skew t, in particular, retain reasonable tractability and are more robust against outliers when compared to mixtures of skew normal. Our proposal is to address robust estimation of mixture of skew normal, to resist sparse outliers and even pointwise contamination that could arise in data collection. To this aim, we incorporate a trimming procedure along the iterations of the EM algorithm. The key idea is that a small portion of observations, which are highly unlikely to occur under the current fitted model assumption, is discarded from contributing to the parameter estimates. Furthermore, aiming at reducing spurious solutions and to avoid singularities of the likelihood, we implement a constrained ML estimation for the component covariances. Monte Carlo experiments show that bias and MSE of the estimators in several cases of contaminated data are dramatically inflated, while they return to be comparable to results obtained on skew data without noise, when the joint effect of trimming and constrained estimation is applied.

Keywords

Finite mixture models; Skew-normal; Robust estimation

Francesca Greselin
University of Milano-Bicocca, Italy, e-mail: francesca.greselin@unimib.it

Agustin Mayo-Iscar
University of Valladolid, Spain, e-mail: agustin@med.uva.es

Luis-Angel Garcia-Escudero
University of Valladolid, Spain, e-mail: lagarcia@eio.uva.es

References

AZZALINI, A. and CAPITANIO, A. (2014) *The Skew-Normal and related Families*, Cambridge University Press, IMS Monographs series, UK.

LEE, S. X. and McLACHLAN G. J. (2013) On mixtures of skew normal and skew t-distributions, *Advances in Data Analysis and Classification*, 7, 241–266.

GARCIA-ESCUDERO L., GORDALIZA A., MAYO-ISCAR A. (2014) A constrained robust proposal for mixture modeling avoiding spurious solutions, *Advances in Data Analysis and Classification*, 8 (1), 27–43.

Using mixture models with random effects to test the differences between classes for supervised classification

Shu-Kay (Angus) Ng

Abstract Many real problems in supervised classification involve high dimensional feature data measured for individuals of known origin from two or more classes. When the dimension of the feature vector exceeds the number of individuals, it presents significant challenges to construct a discriminant rule (classifier) for assigning unclassified individuals to one of the classes. One way to handle this high-dimensional problem is to eliminate irrelevant and redundant features for constructing a classifier. Here a new approach is considered, where a mixture model with random effects is used firstly to cluster the features into groups and then the relevance of each feature variable for differentiating the classes is formally tested and ranked using cluster-specific contrasts of mixed effects. The method is illustrated using several publicly available data sets in cancer research for the discovery of biomarkers relevant to the cancer prognosis. The results show that this new method outperforms typical multiple hypothesis testing methods for identifying differentially-expressed biomarkers from a huge pool of genes, with lower proportion of false discoveries or higher power for a specified level of the false discovery rate. The new method provides a list of biomarkers that improves classifications of disease phenotypes and the prediction of disease outcomes.

Keywords

Contrast, Mixed effects, Mixture models

Shu-Kay (Angus) Ng
Griffith University, Australia, e-mail: s.ng@griffith.edu.au

References

NG, S.K., McLACHLAN, G.J., WANG, K., NAGYMANYOKI, Z., LIU, S., and NG, S.W. (2015): Inference on differences between classes using cluster-specific contrasts of mixed effects. *Biostatistics*, 16, 98–112.

A new spatial prediction method for georeferenced data streams

Antonio Balzanella, Antonio Irpino, and Rosanna Verde

Abstract Massive datasets having the form of continuous streams, with no fixed length, are becoming very common due to the availability of sensor networks which can perform, at a very high frequency, repeated measurements of some variable. In many real world applications such data streams depend on the geographic location of each sensing device so that the records collected by near sensors are more likely to be similar than data collected in distant places.This paper proposes a strategy for monitoring the spatial dependence among data streams and for the prediction of data at spatial locations where there is no recording by sensors. The strategy is based on distributed processing. At each sensor it is performed a summarization of the data by means of a micro-clustering strategy for histogram data. At the central processing node, it is performed the measurement of the spatial dependence and the prediction at unsampled location through a kriging based approach. In order to evaluate the effectiveness of the proposed strategy we have performed extensive tests on real data.

Keywords

Data stream mining; histogram data analysis; spatial data mining

Antonio Balzanella
University of Naples Federico II, Italy, e-mail: antonio.balzanella@unina2.it

Antonio Irpino
University of Naples Federico II, Italy, e-mail: antonio.irpino@unina2.it

Rosanna Verde
University of Naples Federico II, Italy, e-mail: rosanna.verde@unina2.it

Functional data analysis for optimizing strategies of cash flow management

Francesca Di Salvo, Marcello Chiodi, Pietro Patricola, Fabrizio Mineo, and Claudio Lo Piccolo

Abstract The cash management deals with problem of automating and managing cash flow processes. Optimization of the management processes greatly reduces overall cash handling costs. The present analysis is an empirical study of cash flows from and to bank branches, to derive an underlying theoretical framework, which can in a reasonable way be connected with the optimal strategy. Functional data analysis is considered an appropriate framework to analyse the dynamics of the time series behavior of cash flows: since the observations are not equally spaced in time and their number is different for each series, they are converted in a collection of random curves in a space spanned by finite dimensional functional bases. A central issue in the analysis is describing specific patterns of the curves, taking into account the temporal dependence, and the dependence between curves. The analysis provides a dynamic cash management model that is applied with alternative strategies for programming a cash transit for the difference between cash inflows and cash outflows in a fixed period t. As we hypothesize that the strategies are affected by changes in the behavior of the cash flows, the dynamic model outperforms more traditional approaches in identifying the optimal strategy.

Francesca Di Salvo
University of Palermo, Italy, e-mail: francesca.disalvo@unipa.it

Marcello Chiodi
University of Palermo, Italy, e-mail: marcello.chiodi@unipa.it

Pietro Patricola
University of Palermo, Italy, e-mail: pietro.patricola@alice.it

Fabrizio Mineo
Sikelia Service S.p.a., Italy, e-mail: fabrizio.mineo@sikeliaservice.it

Claudio Lo Piccolo
Sikelia Service S.p.a., Italy, e-mail: claudio.lopiccolo@sikeliaservice.it

Keywords

Functional data; Time series; Cash flow management

References

BALZANELLA, A. ROMANO, E., VERDE, R. (2011): Summarizing an Mining Streaming data via a Functional Data Approach. In: B. Fichet, D. Piccolo, R. Verde and M. Vichi (Eds.): *Classification and Multivariate Analysis for Complex Data Structures*. Springer Berlin, 409–116.

LAUKAITIS, A. (2008): Functional Data Analysis for cash flow and transactions intensity continous-time prediction using Hilbert-valued autoregressive processes. *European Journal of Operational Research*, 185, 1607–1614.

LEUNG,L. T. and HAIPENG,X. (2007) Nonparametric functionals of spectral distributions and their applications to time series analysis. *Journal of Statistical Planning and Inference*, 137, 3, 1066–1075.

Fuzzy clustering of distribution-valued data usingan adaptive L2 Wasserstein distance

Antonio Irpino, Francisco De Assis Tenhorio De Carvalho, and Rosanna Verde

Abstract A fuzzy c-means algorithm based on an adaptive L2-Wasserstein distance for distribution-valued data is proposed. The adaptive distance induces a set of weights associated with the components of histogram-valued (a particular type of distributional-valued data) and thus of the variables. The minimization of the criterion in the fuzzy c-means algorithm is performed according three steps such that the representation, the allocation and the weights associated to the components of the variables are alternately computed until a the convergence of the solution to a local optimum. The effectiveness of the proposed algorithm is demonstrated through experiments with synthetic and real-world datasets.

Keywords

distributional data; fuzzy cmeans; adaptive distances

References

DE CARVALHO, F. A. T. and LECHEVALLIER Y. (2009): Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern*

Antonio Irpino
Second University of Naples, Dept. of Political Sciences "J. Monnet", Italy, e-mail: antonio.irpino@unina2.it

Francisco De Assis Tenhorio De Carvalho
Universidade Federal de Pernambuco, Brazil, e-mail: fatc@cin.ufpe.br

Rosanna Verde
Second University of Naples, Dept. of Political Sciences "J. Monnet", Italy, e-mail: rosanna.verde@unina2.it

Recognition, vol. 42, no. 7, 1223–1236.

IRPINO, A., VERDE, R. and DE CARVALHO, F. A. T. (2014): Dynamic clustering of histogram data based on adaptive squared wasserstein distances. *Expert Systems with Applications*, vol. 41, no. 7, 3351–3366.

BEZDECK, J. C. (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Missing data imputation by Multitree

Agostino Di Ciaccio

Abstract With big administrative data, often we have a large number of variables with different measurement levels and many missing data. The correct approach to handle these situations depends on the type of data and the purpose of analysis. However, we can not simply delete the incomplete records, because it amounts to a substantial loss of costly collected data. Single imputation or multiple imputation can be applied to obtain different aims, create an 'imputed' data matrix with the same characteristics of the observed data or take account, in the estimation of a model, of the additional variability due to the imputation process. For big administrative data, several approaches have been proposed in literature. In this paper we compare different approaches, considering both single and multiple imputation, and we propose a new method, named Multitree. By some simulations, we show that Multitree is competitive with the best methods considered in literature.

Keywords

imputation; missing data

Agostino Di Ciaccio
Dept. of Statistics. Sapienza, University of Rome, Italy, e-mail: agostino.diciaccio@uniroma1.it

Missing Data Imputation and Its Effect on the Accuracy of Classification

Lynette Hunt

Abstract Multivariate data sets frequently have missing observations scattered throughout the data set. These missing values can have no particular pattern of occurrence. Several methods have been proposed to address missing data values including imputation, likelihood and weighting approaches. Many machine learning algorithms assume that there is no particular significance in the fact that a particular observation has an attribute value missing. A common approach in coping with these missing values is to replace the missing value using some plausible value, and the resulting completed data set is analysed using standard methods. We evaluate the effect that some commonly used imputation methods have on the accuracy of classifiers in supervised leaning. The effect is assessed in simulations performed on several classical datasets where observations have been made missing at random in different proportions. Our analysis finds that missing data imputation using hot deck, iterative robust model based imputation (IRMI) and factorial analysis for mixed data (FAMD) perform in a similar manner regardless of the amount of missing data and have the highest mean percentage of observations correctly classified. Other methods investigated did not perform as well.

Keywords

missing data; imputation; classification

Lynette Hunt
University of Waikato, Department of Statistics, New Zealand, e-mail: lah@waikato.ac.nz

Handling missing data in observational clinical studies concerning cardiovascular risk: an evaluation of alternative approaches

Nadia Solaro, Daniela Lucini, and Massimo Pagani

Abstract In observational clinical studies, subjects' health status is empirically assessed according to research protocols that prescribe aspects to investigate and methods for investigation. Commonly to many fields of research, such studies are frequently affected by incompleteness of information, a problem that, if not duly handled, may seriously invalidate conclusions drawn from investigations. Regarding cardiovascular risk assessment, usual coronary risk factors (e.g. high blood pressure) and proxies of neurovegetative domain (e.g. heart rate variability) are individually evaluated through direct measurements taken in laboratory. Apart from subjects refusing to undergo tests, a major cause of missingness can be ascribed to the fact that overall sets of collected data typically derive from aggregation of a multitude of sub-studies, undertaken at different times and under slightly different protocols that might not involve the same variables. Data on certain variables can thus be missing if such variables were not included in all protocols. Referring to a specific case study, this issue is addressed by first introducing diagnostic tools for assessing the patterns of missingness compared to the complete part of data, and then detecting the most adequate imputation methods by comparing the performance of alternative (both parametric and data-driven) approaches through a MC simulation study.

Keywords

MC simulations; non-parametric imputation; parametric imputation

Nadia Solaro
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy, e-mail: nadia.solaro@unimib.it

Daniela Lucini
BIOMETRA, University of Milano, Italy, e-mail: daniela.lucini@unimi.it

Massimo Pagani
BIOMETRA, University of Milano, Italy e-mail: massimo.pagani@unimi.it

References

LUCINI, D., SOLARO, N., PAGANI, M. (2014): May autonomic indices from cardiovascular variability help identify hypertension? *Journal of Hypertension*, 32, 2, 363–373.

MOLENBERGHS, G., KENWARD, M.G. (2007): *Missing Data in Clinical Studies*. Wiley, Chichester.

SOLARO, N., BARBIERO, A., MANZI, G., FERRARI, P.A. (2015): A sequential distance-based approach for imputing missing data: The Forward Imputation. *Under review*, 1–19.

Correlation analysis for multivariate functional data

Tomasz Górecki and Waldemar Wolynski

Abstract Data in the form of a continuous vector function on a given interval are referred to as multivariate functional data. These data are treated as realizations of multivariate random processes. In the literature, there are various measures to explore the dependence between two sets of variables. This work presents methods of analysis of association between two multivariate random processes. It generalizes some measures of correlation in the case of functional data. The functional version of canonical correlation analysis is also discussed. Correlation methods for multivariate functional data are presented, illustrated and discussed in the context of analyzing real times series.

Keywords

multivariate functional data analysis; correlation analysis; RV coefficient

References

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

JOSSE, J. and HOLMES, S. (2014): Tests of independence and Beyond. arXiv:1307.7383v3.

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

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

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

Covariance based classification in multivariate and functional data analysis

Francesca Ieva, Anna Maria Paganoni, and Nicholas Tarabelloni

Abstract Literature is plenty of classification methods focusing on differences in location of data clouds. Yet in some cases, e.g. in biostatistics, data may be scarcely distinguishable in terms of location, while show differences in their variability. We propose a new algorithm to perform classification of multivariate and functional data when the difference between the two populations lies in their covariances, rather than in their means. The algorithm relies on a proper quantification of distance between the estimated covariance operators of the populations, and identifies as clusters those groups maximising the distance between their induced covariances. The naive implementation of such an algorithm is computationally forbidding, so we propose an heuristic formulation with a much lighter complexity and we study its convergence properties, along with its computational cost. We also propose to use an enhanced estimator for the estimation of finite-dimensional approximation of covariances of functional data, namely a linear shrinkage estimator, in order to improve the precision of the classification. We establish the effectiveness of our algorithm through applications to both synthetic data and a real dataset coming from a biomedical context, showing also how the use of shrinkage estimation may lead to substantially better results.

Keywords

Covariance-based classification; Functional Data Analysis; Covariance operators

Francesca Ieva
Universit degli Studi di Milano, Italy, e-mail: francesca.ieva@unimi.it

Anna Maria Paganoni
Politecnico di Milano, Italy, e-mail: anna.paganoni@polimi.it

Nicholas Tarabelloni
Politecnico di Milano, Italy, e-mail: nicholas.tarabelloni@polimi.it

Generalization, Combination and Extension of Functional Clustering Algorithms

Christina Yassouridis and Friedrich Leisch

Abstract Clustering functional data is mostly based on the projection of the curves onto an adequate basis and building random effects models of the basis coefficients. The parameters can be fitted by an EM-algorithm. Alternatively distance based models are used in the literature. Such as in the multidimensional case, a variety of derivations of these models has been published. Although their calculation procedure is similar, their implementations are very different including distinct hyperparameters and data formats as input. This makes it difficult for the user to apply and particularly to compare them. Furthermore they are mostly limited to specific basis functions. This article aims to show the common elements between existing models in highly cited articles, first on a theoretical basis. Later their implementation is regarded and it is illustrated how common code chunks can be extracted and how the algorithms can be improved and extended to a more general level. A special consideration is given to those models including the possibility of sparse measurements. Finally they are compared on simulated datasets. An R-package is in the process of being designed, including the modified algorithms and integrated into a unique framework.

Keywords

functional mixed effects; functional clustering; generalization; sparse models

Christina Yassouridis
International Biometric Society,e-mail: christina.yassouridis@boku.ac.at

Friedrich Leisch,
University of Natural Resources and Life Sciences Vienna, Austria e-mail: friedrich.leisch@boku.ac.at

References

BOUVEYRON, C. and JACQUES, J. (2011): Model-based clustering of time series in group-specific functional subspaces, *Advances in Data Analysis and Classification*, 281-300

CHIOU, J.M. and LI, P.L. (2007): Functional clustering and identifying substructures of longitudinal data, Journal of the Royal Statistical Society: Series B, *Statistical Methodology*, 69 (4), 679-699

JACQUES, J. and PREDA, C. (2013): Funclust: a curves clustering method using functional random variables density approximation, *Neurocomputing*, 112, 164-171

JAMES, G. M. and SUGAR, C. A. (2003): Clustering for Sparsely Sampled Functional Data, *Journal of the American Statistical Association*, 98 (462), 397-408

PENG, J. and MLLER, H. (2008): Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions, *The Annals of Applied Statistics*, 3, 1056-1077

RICE, J. and WU, CO. (2001): Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics*, 57 (1), 253-259

SERBAN, N. and JIANG H. (2012): Clustering Random Curves Under Spatial Interdependence With Application to Service Accessibility, *Technometrics*, 54 (2), 108-119

Fuzzy bi-clustering, with application to open-ended questionnaires

François Bavaud, Pascale Deneulin, Laurent Gautier, and Yves Le Fur

Abstract Fuzzy block-models postulate the joint row-column frequencies $f_{ik} = n_{ik}/n_{\bullet\bullet}$ associated to a contingency table (n_{ik}) to be of the form $\pi_{ik} := \sum_{uv} c_{uv}\, a_i^u\, b_k^v$, where the latter ingredients are non-negative and normalised to $c_{\bullet\bullet} = a_\bullet^u = b_\bullet^v = 1$. Local minima of the Kullback-Leibler divergence $K(\pi\|f)$ can be obtained by alternative minimisation on complete distributions, whose (possibly original) explicit formulation yields an EM algorithm for bi-clustering, fuzzy in the sense that supports of distinct group distributions a^g and $a^{g'}$ (as well as b^v and $b^{v'}$) generally overlap. Imposing further constraints (such as sharp distributions a^g or b^v; diagonal, symmetric or independent (c_{uv}), etc.) yields nested models (sharp block-modelling; PLSI; block quasi-symmetry; independence, etc.) with fewer parameters, amenable to nested as well as direct hypothesis testing. Quadratic residuals resulting from the second-order expansion of $K(\pi\|f)$ can be factor-analysed and visualised on a symmetric bi-plot (generalised CA). The theory can be applied to e.g. square matrices of flows, or to document-term matrices. The paper illustrates the inferential, clustering and visualisation results on an online survey addressed between 2011 and 2014, containing the open-ended textual responses (in French, containing 4353 distinct terms) of 1898 wine professionals and 1697 wine consumers, with known socio-economic and behavioral characteristics.

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

Pascale Deneulin
University of Applied Sciences and Arts, Western Switzerland, Changins - Viticulture and Oenology, Switzerland, e-mail: pascale.deneulin@changins.ch

Laurent Gautier
University of Bourgogne - Dijon, France, e-mail: laurent.gautier@u-bourgogne.fr

Yves Le Fur
AgroSup Dijon, France, e-mail: yves.le-fur@agrosupdijon.fr

Keywords

latent models; EM-algorithm; soft clustering

A Self-tuning Region-Growing Algorithm for Deriving Upwelling Areas on Sea Surface Temperature Images

Susana Nascimento, Boris Mirkin, and Sérgio Casca

Abstract A version of the Seeded Region Growing approach for the automatic recognition of coastal upwelling from Sea Surface Temperature (SST) images is proposed. Our algorithm, derived from an approximation clustering model derives a homogeneity criterion in the format of a product rather than the conventional difference between a pixel value and the mean of values over the region of interest. It involves a boundary-oriented pixel labelling so that the cluster growing is performed by expanding its boundary iteratively. We introduce a self-tuning version of the algorithm in which the homogeneity threshold is locally derived from the approximation criterion over a window around the pixel under consideration. The window serves as a boundary regularizer. The algorithm has been applied to a set of 28 SST images of the western coast of mainland Portugal, and compared against a supervised version fine-tuned by maximizing the F-measure with respect to manually labelled ground-truth maps. The areas built by the unsupervised version of our algorithm are significantly coincident over the ground-truth regions in the cases at which the upwelling areas consist of a single continuous fragment of the SST map.

Susana Nascimento
Department of Computer Science and NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa 2829-516 Caparica, Portugal, e-mail: snt@fct.unl.pt

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

Sérgio Casca
Department of Computer Science and NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa 2829-516 Caparica, Portugal, e-mail: sergiofcasca@gmail.com

Keywords

Seeded region growing; approximate clustering; homogeneity criterion

Marked Point Processes for MicroArray Data Clustering

Khadidja Henni, Olivier Alata, Abdellatif El Idrissi, Brigitte Vannier, Lynda Zaoui, and Ahmed Moussa

Abstract Microarray technologies become a powerful techniques for simultaneously monitoring the expression patterns of thousands of genes under different conditions. However, it is important to identify gene groups that manifest similar expressions and are activated by similar conditions. "Mode Detection based on a Marked Point Processes - *k*-Nearest Neighbors" (MDMPP-KNN) is a new Microarray data clustering algorithm performed in two steps: the first one (MDMPP) allows to detect modes of clusters representing regions of high density concentration of observations in the raw space. Based on the well known RJMCMC algorithm, where we consider several movements like birth and death, this algorithm allows to identify prototype observations of each cluster. The second step of the algorithm is the KNN assignation that allows to affect the remaining observations to the corresponding clusters. We experiment MDMPP-KNN on several microarray datasets which offer

Khadidja Henni
LSSD Laboratory, Department of Computer Science, University of Sciences and Technologies Oran "Mohamed Boudiaf" USTO-MB, Oran 31000, Algeria, e-mail: khadidja_henni08@yahoo.fr

Olivier Alata
UMR 5516, Jean Monnet University, Saint-Etienne 42000, France, e-mail: olivier.alata@univ-st-etienne.fr

Abdellatif El Idrissi
ENSA-Tangier, Abdelmalek Essaadi University, BP1818 Route Ziaten 90 000, Morocco, e-mail: latf.elid@gmail.com

Brigitte Vannie
Poitiers University, Poitiers 86073, France, e-mail: brigitte.vannier@univ-poitiers.fr

Lynda Zaoui
Department of Computer Science, University of Sciences and Technologies Oran "Mohamed Boudiaf" USTO-MB, Oran, Algeria, e-mail: zaoui_lynda@yahoo.fr

Ahmed Moussa
ENSA, Abdelmalek Essaadi University, Morocco, e-mail: amoussa@uae.ac.ma

the complexity and large scale. The results show the efficiency of the MDMPP-KNN algorithm compared with K-means, spectral clustering and mean-shift.

Keywords

Marked Point Processes for MicroArray Data Clustering

References

ALATA, O., BURG, S. and DUPAS, A. (2011): Grouping/degrouping point process, a point process driven by geometrical and topological properties of a partition in regions. *Computer Vision and Image Understanding*, 115(9), 1324–1339.

MOUSSA, A., SBIHI, A. and POSTAIRE, J. (2008): A markov random field model for mode detection in cluster analysis. *Pattern Recognition Letters*, 29(9), 1197–1207.

GIANCARLO, R., LO BOSCO, G., PINELLO, L. and UTRO, F. (2013): A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. *BMC Bioinformatics*, 14(S- 1):S6.

A pairwise likelihood approach to simultaneous clustering and dimensional reduction of ordinal data

Monia Ranalli and Roberto Rocci

Abstract The literature on clustering for continuous data is rich and wide; differently, that one developed for categorical data is still limited. In some cases, the problem is made more difficult by the presence of noise variables/dimensions that do not contain information about the clustering structure and could mask it. The aim of this paper is to propose a model for simultaneous clustering and dimensionality reduction of ordered categorical data able to detect the discriminative dimensions discarding the noise ones. Following the underlying response variable approach, the observed variables are considered as a discretization of underlying first-order latent continuous variables distributed as a Gaussian mixture. To recognize discriminative and noise dimensions, these variables are considered to be linear combinations of two independent sets of second-order latent variables where only one contains the information about the cluster structure while the other contains noise dimensions. The model specification involves multidimensional integrals that make the maximum likelihood estimation cumbersome and in some cases unfeasible. To overcome this issue the parameter estimation is carried out through an EM-like algorithm maximizing a pairwise log-likelihood. Examples of application of the model on real and simulated data are performed to show the effectiveness of the proposal.

Keywords

Mixture models; Reduction ordinal data; Pairwise Likelihood

Monia Ranalli
Department of Statistics, The Pennsylvania State University, USA e-mail: monia.ranalli@psu.edu

Roberto Rocci
IGF Department, Università Tor Vergata, Italy e-mail: roberto.rocci@uniroma2.it

References

LINDSAY, B. (1988): Composite likelihood methods. *Contemporary Mathematics*, 80, 221–239.

RANALLI, M., and ROCCI, R. (2014): Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, DOI:10.1007/s11222-014-9543-4.

RANALLI, M., and ROCCI, R. (2015): A pairwise likelihood approach to simultaneous clustering and dimensional reduction of ordinal data. *arXiv:1504.02913*.

Machine learning diagnoses on patients presenting abdominal pain

Hong Gur

Abstract Currently a massive amount of data have been collected in the electronic database of the emergency department information system across provinces in Canada. This valuable data source is currently not well utilized. Patient diagnosis is a multi-class classification problem. Our purpose is to build an automated diagnosis tool for physicians. It will provide a list of most possible diagnoses. The physician could refer to this list to check for any possibilities that were overlooked. This diagnostic tool can also assess the most informative additional tests, which will reduce the time and resources used on unnecessary diagnostic tests. The statistical problem is not a standard multi-class classification problem since different test results are available for different patients. We propose a model framework that can combine different models for groups of patients with different variables. The same technique can be used to combine data from different hospitals into one national system, which can take into account local trends in pathology as well as being able to draw on the whole national database to give the best diagnosis possible.

Keywords

Multiclass classification; missing values; multi-model combination

Hong Gu
Department of Math. & Stat. Dalhousie University, Canada, e-mail: hgu@dal.ca

Credibility Classification with Missing Data

Toby Kenney

Abstract Credibility theory combines the information from a small but relevant dataset and a larger but more general dataset. Here the information we wish to extract is a classifier, and different samples represent different sets of predictor variables available. Using all common variables we have a lot of data, but are missing critical predictors. Restricting attention to datapoints with a particular variable, we have less data, so overfit is possible. We apply credibility theory to create a compromise classifier. To achieve this, we develop a new variant of credibility theory combining the features of crossed-classification credibility theory and hierarchical credibility theory.

Keywords

Credibility Theory; Classification

Toby Kenney
Dalhousie University, Canada, e-mail: tkenney@mathstat.dal.ca

Clinical Decision Support System for HCC using Classification Models

Taerim Lee

Abstract The purpose of this paper is to construct and optimize performance of a classification model to aid management of Clinical Decision Support System and to evaluate performance implementation effectiveness and barriers to adoption. We used 8 classification models including RF, SVM, k-nearest neighbor, linear discriminant analysis, logistic regression, boosting and ANN were trained and their performance were compared for predicting patients prognosis. Variable selection was performed and only clinical variables relevant to outcomes were utilized for predicting the outcome to avoid over fitting the model and to detect the surveillance clinical exams significance. After that we plan to compare the performance of RF, ANN and SVM to other classification results using ROC curves. Using several packages of Random forest (package for Random Forest), Support vector machine (package e1071) Shrunken centroid (package pamr) LDA9package (sml), KNN (package class), and logistic regression (package stats), Boosting (package boost), ANN (Neural networks package) we can get the results of the classification for clinical support system. We find that several clinical variables and SNP data were selected as important factor for clinical decision of HCC patient prognosis and manage surveillance clinical exams using 8 classification models. For the comparison of models and methods all models were run evaluation step and validation step using 10 fold cross validation and the accuracy, sensitivity, specificity, positive predictive value and negative predictive value, ROC curves and area under ROC (AUC) with Mann Whitney test. Random Forest seems to be overall the best performing model in terms of accuracy and balance of sensitivity and specificity. Using clinical decision support system Physicians can improve their diagnosis and decision making assisted and it could be efficient and cost effective management of medical resources and surveillance clinical exam.

Taerim Lee
Dept. of Information & Statistics, Republic Of Korea, e-mail: trlee@knou.ac.kr

Keywords

clinical decision support system; ROC; surveillance clinical exam

References

CHU, A., AHN H., HALAWAN, B. and KALMIN, B.(2008): A decision support system to facilitate management of patients with acute gastrointestinal bleeding. *Artificial Intelligence in Medicine*, 42, 247–259

KAVISHWAR BALWANT WAGHOLIKEAR, KATHY L MACLAUGHLIN (2013): Formative evaluation of the accuracy of clinical decision support system for cervical cancer screening, 20, 749–757

DELONG, E. R., DELONG, D. M. and CLARKE-PEARSON, D. L. (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44,837–845.

Machine-learning classification methods with the ability to hesitate

Michal Trzesiok

Abstract Making decision is sometimes very difficult. Even in "yes or no" situations, regardless of the amount of information given, sometimes we (as human beings) hesitate, because we feel it can go both directions. The situation is similar in machine-learning tasks. We can use very sophisticated classification methods to support our decision-making process. The machine is learned, the model is built, but then it seems reasonable to expect the machine to give us at least a warning when the prediction is unstable (which means that it is sensitive to small changes in explanatory variables' values). In such a situation the user can search for additional information to make the decision more reliable (e.g. ask a few new questions in credit scoring problem in order to reject or grant a loan request). The main goal is to present a procedure for providing the machine with the ability to show hesitation, when it is justified. The proposed procedure is based on sensitivity analysis. We illustrate the procedure on a real-world data set using the Support Vector Machines, but the procedure is universal and it can be also used with other classification methods.

Keywords

classification; sensitivity analysis; SVM

Michal Trzesiok
University of Economics in Katowice, Poland, e-mail: michal.trzesiok@ue.katowice.pl

References

DIEDERICH, J. (Ed.). (2008): *Rule extraction from support vector machines* (Vol. 80). Springer Science & Business Media.

SHEN, K. Q., ONG, C. J., LI, X. P.,, WILDER-SMITH, E. P. (2008): Feature selection via sensitivity analysis of SVM probabilistic outputs. *Machine Learning*, 70(1), 1–20.

Improving Predictions for Tree Ensembles using Distributions of Estimated Probabilities with Applications in Record Linkage

Samuel L. Ventura and Rebecca Nugent

Abstract A random forest is an ensemble of classification trees built using randomized subsets of the covariates and bootstrap samples of the observations at each split in each of the T trees in the ensemble. In a two-class problem, random forests aggregate the predictions of each underlying tree using a "majority vote" scheme, assigning the class with the majority vote amongst the underlying trees as the predicted class of the random forest. Predicted/estimated probabilities for a given class are obtained by finding the proportion of trees that voted for that class. However, much information is discarded in this process, including the individual predicted probabilities from each of the T underlying trees, which we show to often have multimodal or heavily skewed distributions. A new prediction approach is introduced that extracts and incorporates information from these distributions of tree probabilities for the two-class problem. A second classifier is trained, modeling the binary outcome given summary statistics from the distribution of tree probabilities. This approach is assessed on a large, labeled record linkage dataset of death records from the Syrian Civil War conflict.

Keywords

random forest; classification trees; record linkage

Samuel L. Ventura
Carnegie Mellon University, USA, e-mail: sventura@stat.cmu.edu

Rebecca Nugent
Carnegie Mellon University, USA, e-mail: rnugent@stat.cmu.edu

References

BREIMAN, L. (2001): Random Forests. *Machine Learning*, 45(1), 5–32.

PRICE, M., GOHDES, A., BALL, P. (2014): Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic. *Human Rights Data Analysis Group*, commissioned by the United Nations Office of the High Commissioner for Human Rights.

FELLEGI, I.P., SUNTER, A.B. (1969): A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328).

T-sharper images and T-level cuts of fuzzy partitions

Slavka Bodjanova

Abstract A fuzzy partition is usually described by a matrix U with elements from the unit interval. For a variety of reasons, U is often approximated by simpler matrices. The most popular are the Maximum membership approximations or approximations based on alpha-level cuts of U. Because they create crisp matrices with elements either zero or one, most of the detailed information from U is lost. Two methods of gradual non-crisp approximations of U based on a set of parameters T derived from U are proposed. The first produces T-sharper images of U with a gradually decreasing amount of fuzziness. The final image is a convex combination of all possible Maximum membership approximations of U. The second method generates T-level cuts of U. They represent a gradual transformation from the lower crisp approximation of U by its core (alpha-cut at level alpha=1) to the upper crisp approximation of U by its support (strong alpha-cut at level alpha=0). Properties of both methods are discussed and illustrated with examples.

Keywords

fuzzy partitions; gradual approximation, partition entropy

Slavka Bodjanova
Texas A&M University-Kingsville, United States, e-mail: kfsb000@tamuk.edu

Minkowski weighted k-means clustering with a median-based consensus rule

Renato Cordeiro de Amorim and Vladimir Makarenkov

Abstract The intelligent Minkowski weighted k-means is a recently developed clustering algorithm capable of computing feature weights. Its cluster-specific weights follow the intuitive idea that a feature with a low dispersion at a specific cluster should have a higher weight than one with a high dispersion. During the clustering process, the intelligent Minkowski k-means algorithm uses feature weights representing the cluster-specific degree of relevance. The final clustering provided by this algorithm obviously depends on the selection of a Minkowski exponent. The median consensus rule we will introduce helps in the selection of the optimal Minkowski exponent. Our rule takes into account the values of the adjusted rand index (ARI) between clustering solutions obtained for different Minkowski exponents and selects the clustering that provides the highest average value of ARI. Our simulations, carried out with real and synthetic data, show that the new median-based consensus procedure usually outperforms the clustering strategies that select the solution corresponding to the highest value of the Silhouette or Calinski-Harabasz clustering validation index over all considered Minkowski exponents.

Keywords

K-Means; Minkowski distance; Consensus rule

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

Vladimir Makarenkov
Departement d'informatique, Universit du Quebec a Montreal, C.P. 8888 succ. Centre-Ville, Montreal (QC) H3C 3P8, Canada, e-mail: makarenkov.vladimir@uqam.ca

References

AMORIM, R.C. and MIRKIN, B. (2012): Minkowski Metric, Feature Weighting and Anomalous Cluster Initialisation in K-Means Clustering. *Pattern Recognition*, 45(3), 1061–1075.

AMORIM, R.C. and KOMISARCZUK P. (2012): On Initializations for the Minkowski Weighted K-Means. Lecture Notes in *Computer Science*, 7619, 45–55.

AMORIM, R.C. (2015): Feature relevance in Ward's hierarchical clustering using the Lp norm, *Journal of Classification*, 32.

A Note on Spherical k-Means++ Clustering

Yasunori Endo

Abstract Recently, information from a large-scaled social dataset has great effect on many aspects of society. We can mention recommendation systems as an example. When a person uses online markets, the recommendation system estimates commodities he prefers from his purchase history and show the commodities on the display. Some data mining tools to retrieve useful information from the social data set play very important role in such systems. One of the most representative tool is spherical k-means clustering algorithm (SKM)[1]. The algorithm classifies the data with the norm one and it is sufficient because many data in the social data set are normalized when we retrieve useful information from the social data set. However, the SKM has a big problem, that is initial value dependence (i.v.d.). Therefore, this presentation shows an algorithm in which the i.v.d. problem is solved, called spherical k-means++ clustering algorithm (SKM++). This work is inspired by k-means++ clustering algorithm (KM++)[2,3]. First, we revise the dissimilarity between data in SKM to satisfy the triangle inequality. Second, clustering results by the SKM with the revised dissimilarity is equivalent to the original SKM. This fact is necessary to construct SKM++. Third, we show a way to select initial values in the clustering process, and prove that the way solves the i.v.d. of SKM. Forth, we show that i.v.d. of SKM++ is theoretically estimated as a half of KM++.

Keywords

clustering; spherical k-means++; initial value dependence; social data set; recommendation system

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

References

[1] HORNIK, K., FEINERER, I., KOBER, M. and BUCHTA, C. (2012): Spherical *k*-Means Clustering, Vol. 50, Issue 10.

[2] ARTHUR, D. and VASSILVITSKII, S. (2007): k-means++: the advantages of careful seeding, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027-1035.

[3] DASGUPTA, S. (2013): Lecture 3 -Algorithms for k-means clustering, `http://cseweb.ucsd.edu/~dasgupta/291-geom/kmeans.pdf`.

Clustering based on adaptive Mahalanobis kernels

Marcelo Ferreira and Francisco de Carvalho

Abstract Over the last years several clustering methods have been modified to incorporate kernels and various kernel clustering methods have been proposed. The Gaussian kernel is the most commonly used kernel function and usually delivers good results. Despite the good characteristics of this kernel, it is based on the Euclidean distance which assumes that patterns are more likely distributed within an hyperspherical region, i.e., each variable has the same variance and there is no covariance. However, patterns in two different groups are more likely distributed within two different hyper-ellipsoidal regions, respectively. The Mahalanobis distance, which takes into account the correlations between variables and is scale-invariant, is a better choice to deal with hyper-ellipsoidal regions. We propose K-means type algorithms based on adaptive Mahalanobis kernels. These kernels are built based on adaptive quadratic distances defined by symmetric positive definite matrices that change at each iteration of the algorithm and either can be the same for all clusters or different from one cluster to another. The adaptive Mahalanobis kernel takes into account the correlations between variables, allowing the discovery of clusters with non-hyperspherical shapes. Experiments with simulated as well as benchmark data illustrate the usefulness of the methods.

Keywords

Kernel clustering; Mahalanobis kernel; Adaptive distance

Marcelo Ferreira
Federal University of Paraiba, Brazil, e-mail: marcelorpf@gmail.com

Francisco de Carvalho
Federal University of Pernambuco, Brazil, e-mail: fatc@cin.ufpe.br

Unit level small area model with covariates perturbed for disclosure limitation.

Serena Arima

Abstract Small area estimation deals with the problem of estimating area level aggregates, when the sampling design is not guaranteed to produce sufficiently large subsamples for all areas of interest. Direct survey estimates may be unreliable and improved estimates can be obtained using mixed effects regression models that link the small areas and borrow strength from similar domains. In this paper we focus on a small area model in which covariates are measured with error. We consider the situation in which errors in covariates are artificially introduced by a mechanism of confidentiality protection. Statistical disclosure control (SDC) is commonly applied to prevent reidentification of respondents. An area model that includes auxiliary variables perturbed by disclosure limitation methods has been proposed in Polettini and Arima (2015) However, in a disclosure limitation context unit-level models arise more naturally since the perturbation is usually performed at the unit-level. In this paper we extend the aforementioned model as a unit-level model: we investigate the performance of the model in estimating the regression parameters and predicting the small area means. We also study the model capability in re-identifying respondents by predicting the true value of the perturbed variables for each unit.

Keywords

Small area estimation; disclosure limitation; MCMC

Serena Arima
Sapienza University of Rome, Italy e-mail: serena.arima@uniroma1.it

References

POLETTINI, S. and ARIMA, S. (2014). Small Area Estimation with Covariates Perturbed for Disclosure Limitation, *Proceedings of the XLVII Scientific Meeting of the Italian Statistical Society*, ISBN: 978-88-8467-874-4, CUEC Cooperativa Universitaria Editrice Cagliaritana, p. 1–6 (10–14 June 2014 Cagliari)

SparseStep: Approximating the Counting Norm for Sparse Regularization

Gertjan Van den Burg and Patrick Groenen

Abstract A common goal in regression problems is feature selection, that is, selecting only those predictors that significantly influence an observed outcome generally from a large group of potential predictors. Existing solutions to this problem either use tedious enumeration of all possible subsets, ad hoc search procedures such as forward selection and backward elimination, or regularization of the loss function such as the Lasso. This latter approach forces the influence of some predictors to zero by shrinking the absolute size of the parameters. However, to study the true effects of the predictors it could be preferable to find a subset of predictors without or with little shrinkage of the non-zero coefficients. Theoretically this is possible by using the stepwise counting norm as penalty term. Previously, a smooth approximation of the counting norm was proposed in the context of a sparse deconvolution problem. We present an Iterative Majorization algorithm for minimizing the regularized loss function for the regression problem using this approximated counting norm, and name the resulting method SparseStep. We furthermore present the results of a simulation study where the proposed method is compared with existing sparsity-inducing regularization approaches, both as a modelling tool and as a prediction method.

Keywords

Regression; Regularization; Sparsity

Gertjan Van den Burg
Erasmus University Rotterdam, Netherlands, e-mail: burg@ese.eur.nl

Patrick Groenen
Erasmus University Rotterdam, Netherlands, e-mail: groenen@ese.eur.nl

References

DE ROOI, J. and EILERS, P. (2011): Deconvolution of pulse trains with the L 0 penalty. *Analytica chimica acta*, 705(1), 218–226.

Sparse Principal Covariates Regression for high-dimensional data

Katrijn Van Deun and Eva Ceulemans

Abstract Prediction in a context of high-dimensional data, this is data with many more covariates than observations, is an ill-posed problem. Popular solutions are the introduction of penalties that perform variable selection (e.g., the lasso or elastic net in regression) and the use of dimension reduction methods to reduce the covariates to a few components (e.g., principal covariates regression that simultaneously optimizes the reduction of the covariates to a few components and the prediction of the outcome by these components). From an interpretational point of view it is attractive to reduce the covariate space to a few meaningful components. However, in a high-dimensional context interpretation of the components is daunting as the components are based on a linear combination of a huge number of variables. To account for this interpretational issue, we propose a sparse principal covariates regression approach that selects a limited number of relevant variables to construct the components and that uses an alternating least squares and coordinate descent estimation procedure. We will compare with sparse partial least squares and illustrate with a systems vaccinology example with the aim to predict the antibody titers for subjects vaccinated against the flu by thousands of genes.

Keywords

high-dimensional data; sparse principal covariates regression

Katrijn Van Deun
Tilburg University, Netherlands, e-mail: k.vandeun@uvt.nl

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

Dual model selection for principal covariates regression

Marlies Vervloet, Katrijn Van Deun, Wim Van den Noortgate, and Eva Ceulemans

Abstract Dimension-reduction based regression methods reduce the predictors to a few components and predict the criterion using these components. When applying such methods in the behavioral sciences, it is often not only important to achieve good prediction of the criterion, but also desirable to gain correct information about the underlying structure of the predictors (i.e., recovery of the underlying components). In contrast to PLS and PCR, PCovR explicitly aims at achieving both goals simultaneously. Moreover, the extent to which both aspects play a role in the construction of the components can be determined flexibly with a weighting parameter. This has as a downside that a dual model selection strategy is needed: selection of the number of components and selection of the weighting parameter value. Therefore, four model selection strategies are proposed, and the optimality of the components they select is studied in comparison to those resulting from PCR and PLS analyses. We start off by discussing the concept of optimality. Next, we present a simulation study. Finally, based on the results of this study, concrete guidelines are offered on how to combine multiple strategies to obtain the optimal model.

Keywords

regression; dimension reduction; model selection

Marlies Vervloet
KU Leuven, Belgium, e-mail: marlies.vervloet@ppw.kuleuven.be

Katrijn Van Deun
Tilburg University, Netherlands, e-mail: k.vandeun@tilburguniversity.edu

Wim Van den Noortgate
KU Leuven, Belgium,e-mail: wim.vandennoortgate@kuleuven-kulak.be

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

References

DE JONG, S. and KIERS, H. A. (1992): Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, 14(1-3), 155–164.

On Associative Confounder Bias

Priyantha Wijayatunga

Abstract Conditioning on sufficient set of confounders that affect both treatment and outcome causally is necessary for eliminating bias in causal effect estimates of the treatment on the outcome. This is done through including them in propensity score calculations in so-called potential outcome framework for causal inference whereas in causal graphical framework usual conditioning is done. However in the former, things get confused when the modeller finds a variable that is non-causally associated with both the treatment and the outcome. Some researchers have argued that those associated factors should also be included in propensity score calculations for removing bias while others have argued that they cause no bias and have shown that conditioning on them introduces spurious dependence between the treatment and the outcome, therefore resulting some bias in the causal effect estimates. They use common cause principle for their argument. Here, we show that there are errors in both the arguments if we assume common cause principle. We show that when such a factor is observed neither of the actions are appropriate. We show that it is needed to consider the strengths of dependencies between the factor and, the treatment and the outcome in order to select the action. We use a simple geometric figure for our argument, that can be used to view Simpson's paradox. We also discuss more complex case as for the current discussion, that is often over looked by causal modelers.

Keywords

causal effects; estimation; conditioning; latent common cause

Priyantha Wijayatunga
Department of Statistics, Umea University, SE-901 87 Umea, Sweden, e-mail: priyantha.wijayatunga@stat.umu.se

References

LANGFORD, E., SCHWERTMAN, N., and OWENS, M. (2001): Is the property of being positively correlated transitive? *The American Statistician*, **55**(4): 322–324.

PEARL, J. (2009): *Causality*, Second Edition, Cambridge University Press, UK.

PEARL, J. (2009): Letter to the Editor. *Statistics in Medicine* 2009; **28**: 1415–1416.

RUBIN, D. (2005): Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, **100**(469), 322–331.

RUBIN, D. (2009): Author's Reply. *Statistics in Medicine* 2009; **28**: 1420–1423.

SJÖLANDER, A. (2009): Letter to the Editor. *Statistics in Medicine* 2009; **28**: 1416–1420.

SHRIER, I. (2009): Letter to the Editor. *Statistics in Medicine* 2009; **28**: 1315–1318.

SIMPSON, E. H. (1951): The interpretation of interaction in contingency tables. *Journal of Royal Statistical Society, Series B*, **13**(2), 238–241.

YULE, G. U. (1903): Note on the theory of association of the attributes in statistics. *Biometrika*, **2**(2), 121–134.

Clustering of links in networks

Jernej Bodlaj and Vladimir Batagelj

Abstract An approach to cluster links of a network, proposed by Evans and Lambiotte, is to construct a corresponding line graph and apply to it a network nodes clustering algorithm. The problem with this approach is that the obtained line graph can be very large – a node of degree d contributes a complete subgraph on d nodes to the line graph. Ferligoj and Batagelj adapted standard clustering algorithms for solving the clustering problem with relational (network) constraints. Because these algorithms are based on the dissimilarity matrix they are too much space (and time) consuming for large networks.To obtain an efficient agglomerative clustering algorithm for large networks the following approach is used: (1) compute the dissimilarities between units (nodes of network) only for end nodes of existing links of the network; (2) define the dissimilarities between clusters based only on the dissimilarities of the corresponding links and derive the update relations. For selected dissimilarities between clusters (minimum, maximum, and average) the Bruynooghe reducibility property holds. This allows to speed-up the hierarchical clustering procedure by using the RNN (reciprocal nearest neighbors) approach. Combining these ideas an efficient agglomerative algorithm for clustering the network links was developed. The obtained link clusterings induce corresponding overlapping node clusterings.The developed algorithm will be illustrated by clustering some real life large networks.

Keywords

clustering links in network; large networks; reciprocal nearest neighbors

Jernej Bodlaj
University of Ljubljana, FRI, Slovenia, e-mail: bodlaytm@gmail.com

Vladimir Batagelj
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia, e-mail: vladimir.batagelj@uni-lj.si

References

EVANS, T.S., LAMBIOTTE, R. (2009): Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80, 016105, 1–8.

FERLIGOJ, A., BATAGELJ, V. (1983): Some Types of Clustering with Relational Constraints. *Psychometrika*, vol. 48, no. 4, 541–552.

MURTAGH, F. (1985): Multidimensional Clustering Algorithms. *Compstat lectures*, 4, Vienna: Physica-Verlag.

Two-stage agglomerative hierarchical clustering using medoids for network clustering

Sadaaki Miyamoto

Abstract Clustering of nodes of networks is studied by researchers interested in web information such as SNS networks. Since objects are not points in the Euclidean space, K-means cannot be applied to such networks. Nevertheless, a center of a cluster is useful. Hence the medoid is used instead of the centroid of a cluster. Let $X = \{x_1, \ldots, x_N\}$ be the set of objects for clustering; $D(x,y)$ is a dissimilarity measure defined on X. A medoid of cluster C is: $\text{medoid}(C) = \arg\min_{y \in C - \{y\}} \sum_{x \in C} D(y,x)$. We propose a two-stage agglomerative hierarchical algorithm in which the first stage uses a one-pass K-medoids or K-medoids++, while the second stage uses Medoid agglomerative clustering or a Medoid-Ward clustering applied to the set of medoids of clusters obtained in the first stage. The both algorithms use medoids instead of centroids in the centroid clustering and the Ward clustering, with the use of $D(x,y)$ instead of the squared Euclidean distance. Repository data and a real Twitter network data are tested using different algorithms combining K-medoids or K-medoids++, and Medoid agglomerative clustering or Medoid-Ward clustering. Evaluation criteria of the effectiveness of the results are considered.

Keywords

agglomerative hierarchical clustering; medoid: two-stage hierarchical algorithm

Sadaaki Miyamoto
Dept. of Risk Eng. University of Tsukuba, Japan, e-mail: miyamoto@risk.tsukuba.ac.jp

Network Tools and Homophily Measures for Brand Image Analysis

Agnieszka Stawinoga, Simona Balbi, and Germana Scepi

Abstract In Text Mining, large corpora are explored, in order to discover and synthesize their content, in an automatic, time-saving way. If the aim is to understand the similarity of the documents in the corpus, it is interesting to represent them as a network measuring their proximity in relation to the words they use. Therefore, statistical tools, developed for the analysis of Social Networks can be applied. In Social Network Analysis, it is often important to introduce information related to the nodes, such as characteristics identifying their belonging to a specific group. In literature different indices have been proposed in order to measure the tendency of actors to have relations with actors similar to themselves. Analogously, in the analysis of textual data, better results can be achieved by introducing information related to each document (e.g. characteristics of the author). Here we focus our attention on advertisements, and their textual component. In competitive markets there is a complex relation between the image of a brand, and the messages of the advertising campaigns over time. We propose a statistical procedure for analyzing the evolution of the brand image through the different campaigns of a famous brand on the basis of several homophily measures.

Keywords

Network Analysis; Textual Data

Name of Second Author
University of Naples Federico II, Italy e-mail: agnieszka.stawinoga@unina.it

Name of Second Author
University of Naples Federico II, Italy, e-mail: sb@unina.it

Germana Scepi
University of Naples Federico II, Italy, e-mail: scepi@unina.it

Predicting the evolution of a constrained network: a beta regression model

Luisa Stracqualursi and Patrizia Agati

Abstract Social network analysis allows to map and measure relationships and flows (links) between people, groups, computers, URLs, or other connected knowledge entities (nodes). In this context, a relevant issue is the treatment of constrained scale-free networks such as, for example, the network of student transfers between degree courses offered by an University, that are strongly influenced by a number of institutional decisions. In the analysis of such a system, special attention has to be paid to identify current or future "critical points", that is nodes characterized by a high number of outcoming or incoming links, on which to act in order to optimize the network. To predict the evolution of a constrained system over time in dependence of constraint modifications, a beta regression model is proposed, that fit links represented by quantities varying between 0 and 1. The algorithm was successfully applied to the network of student transfers within the University of Bologna: the link was defined by the out-transfer rate of the degree course (computed as the ratio of the number of out-transfers to the number of students enrolled) and the critical points of the system were defined by the courses characterized by a high out-transfer rate.

Keywords

Beta regression model; social network analysis; constrained scale-free networks

Luisa Stracqualursi
University of Bologna, Italy, e-mail: luisa.stracqualursi@unibo.it

Patrizia Agati
University of Bologna, Italy, e-mail: patrizia.agati@unibo.it

References

BARABASI, A.L. and ALBERT, R. (2002): Statistical mechanics of complex networks, *Review of Modern Physics*, 74, 47–97.

BOLLOBAS, B. (1985): *Random Graphs*, Academic Press, London

JOHNSON N.L., KOTZ S., BALAKRISHNAN N. (1995): *Continuous univariate distributions*, vol. 2, Wiley, New York

Clustering via Mixture Models with Flexible Components

Geoff McLachlan and Sharon Lee

Abstract This talk considers extensions to the multivariate normal distribution for a parametric approach to the clustering of data with potentially skewed and long-tailed observations.In particular, we introduce a finite mixture of canonical fundamental skew t (CFUST) distributions for use in situations where the clusters are asymmetric and possibly long-tailed. The family of CFUST distributions includes the so-called restricted multivariate t (rMST) and the unrestricted multivariate skew t (uMST) distributions as special cases. In recent years, a few versions of the multivariate skew t (MST) mixture model have been put forward, together with various EM-type algorithms for parameter estimation. These formulations adopted either a restricted or an unrestricted characterization for their MST densities. We present a natural generalization of these developments, employing the CFUST distribution as the parametric family for the component distributions, and point out that the restricted and unrestricted characterizations can be unified under this general formulation. Also, various examples are presented to illustrate the limitations of the rMST and uMST models compared to the CFUST model.

Keywords

Mixture distributions; skew components

Geoff McLachlan
University of Queensland, Australia, e-mail: g.mclachlan@uq.edu.au

Sharon Lee
University of Queensland, Australia, e-mail: sharon.lee1@uqconnect.edu.au

References

LEE, S.X. and McLACHLAN, G.J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, 7, 241–266.

LEE, S. and McLACHLAN, G.J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24, 181–202.

LEE, S.X. and McLACHLAN, G.J. (2015). Finite mixtures of canonical fundamental skew t-distributions: the unification of the restricted and unrestricted skew t-mixture models.*Statistics and Computing*. To appear (Advance Access published 28 February, 2015; doi:10.1007/s11222-015-9545-x).

Density-based clustering multiplex networks

Giovanna Menardi and Domenico De Stefano

Abstract Within large communities, individuals sparsely interact with each others but set a tight releationship with a limited number of subjects only.These aggregations depend on the nature of the relationship, being guided by geographic neighbourhood,task sharing, homophily, and other agglomerative processes. In social network analysis this is translated into the definition of multiple layers, and the actors social behaviour results in the creation of clusters of densely connected actors, loosely connected with actors of other groups. A question of interest is to identify the groups and reveal their formation mechanisms. The correspondence between groups of subjects and the inner connection density, suggests the idea of extending the density-based approach for clustering non-relational data to the network framework. The nonparametric formulation of this approach associates clusters with high-density regions of the sample space. While a probabilistic notion of density is undefined for networks, this lack allows us to consider ad-hoc measures depending on the kind of aggregation mechanism one is interested to uncover.The proposed method allows to deal with very general relational structures such as the so-called multiplex networks - networks spanned on the same actors interacting through different relationships - for which very few methods have been proposed.

Keywords

density function; modal clustering; multiplex networks

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

Domenico De Stefano
Department of Political and Social Sciences, University of Trieste, Italy,e-mail: ddestefano@units.it

Modal Clustering and cluster inference

Surajit Ray

Abstract In this talk I will present new methods developed in the framework of mode based inference. Li, Ray and Lindsay (2007) proposed the method of modal clustering that identifies local mode by starting at any point based on kernel density estimates and further clustering the data that converge to the same mode. Previously Ray and Lindsay (2005) introduced the concept of ridgeline manifold, which has been used by various researchers to develop techniques threshold based approaches for cluster merging. Here we develop a kernel density based asymptotic and bootstrap based inference framework to identify the significance of modes and use them for cluster merging. The inference procedure is applied on both simulated and real datasets.

Keywords

Modal Clustering; Density based clustering; High Dimension; Big Data; Inference

References

LI, J., RAY, S. and LINDSAY, B.G. (2007) A Nonparametric Statistical Approach to Clustering via Mode Identification. *Journal of Machine Learning Research*, 8, 1687–1723.

RAY, S. and LINDSAY, B.G.(2005) The Topography of Multivariate Normal Mixtures. *The Annals of Statistics*, 33, 2042–2065.

CHENG, Y., and RAY, S. (2014) Multivariate modality inference using Gaussian kernel. *Open Journal of Statistics*, 4(5)

Surajit Ray
University of Glasgow, United Kingdom, e-mail: surajit.ray@glasgow.ac.uk

Prediction error in distance-based generalized linear models

Eva Boj del Val, Teresa Costa Cor, and Josep Fortiana Gregori

Abstract We describe how to calculate prediction error in distance-based generalized linear models. In generalized linear models, the mean squared prediction error can be approximated with the sum of two components: variability in the data (process variance) and variability due to estimation (estimation variance). The estimation variance can be calculated by using the corresponding formula or, alternatively, can be approximated by using bootstrap methodology. When we use bootstrap methodology we are able to obtain, in addition, the predictive distribution of the estimations. We apply these concepts to the actuarial problem of claim reserving, where data are collected in a run-off triangle, and it is of interest the use of generalized linear models and the calculus of prediction error. We illustrate the computations with a well-known data set. The related distance-based generalized linear models are fitted using the `dbglm` function of the `dbstats` package for R.

Keywords

distance-based generalized linear model; prediction error; dbstats

Eva Boj del Val
Universitat de Barcelona, Spain, e-mail: `evaboj@ub.edu`

Teresa Costa Cor
Universitat de Barcelona, Spain, e-mail: `tcosta@ub.edu`

Josep Fortiana Gregori
Universitat de Barcelona, Spain, e-mail: `fortiana@ub.edu`

References

BOJ, E., DELICADO, P., FORTIANA, J., ESTEVE, A. and A. CABALLÉ (2012). Local distance-based generalized linear models using the dbstats package for R. *Documentos de Trabajo de la Xarxa de Referència en Economia Aplicada (XREAP)*, XREAP2012–11.

BOJ, E., A. CABALLÉ, P. DELICADO, and J. FORTIANA (2014). dbstats: Distance-Based Statistics (dbstats). *R package version 1.0.4*.

TAYLOR, G. and F.R. ASHE (1983). Second Moments of Estimates of Outstanding Claims. *Journal of Econometrics*, 23, 37–61.

Prediction Accuracy in Logistic Biplots for categorical data.

Jose Luis Vicente-Villardón and Julio Cesar Hernandez-Sanchez

Abstract Classical biplot methods allow for the simultaneous representation of individuals (rows) and variables (columns) of a numerical data matrix. When data are binary, nominal or ordinal, classical linear biplots are not adequate; other techniques such as multiple correspondence analysis (MCA), latent trait analysis (LTA) or item response theory (IRT) for categorical items should be used instead. We have recently extended the biplot to categorical data. The resulting method is termed " logistic biplot"(LB) because the resulting procedure is related to logistic responses in the same way classical biplots are related to linear responses. For the nominal case, variables are represented as convex prediction regions rather than vectors; using the methods from computational geometry, the set of prediction regions is converted to a set of points in such a way that the prediction for each individual is established by its closest "category point". Then interpretation is based on distances rather than on projections. For the binary and ordinal cases, the final representation is more like a traditional biplot with straight lines for predicting probabilities for each variable. The prediction regions are delimited by parallel straight lines and then a line with the adequate marks is enough to visualize the model. We evaluate prediction accuracy of logistic biplots compared to MCA and IRT. The main differences between the LB and MCA are shown with data from demographic and labor market variables of doctorate (PdH) holders in the region of Castilla-Leon in Spain, using the 'Survey on the careers of doctorate holders (CDH)' carried out by Spanish Statistical Institute jointly with Eurostat, the Organization for Economic Co-operation and Development (OECD) and UNESCO's Institute for Statistics (UIS).

Jose Luis Vicente-Villardón
Department of Statistics Universidad de Salamanca, Spain, e-mail: villardon@usal.es

Julio Cesar Hernandez-Sanchez
Instituto Nacional de Estadística, Spain, e-mail: juliocesar.hernandez.sanchez@ine.es

Keywords

logistic biplot; categorial data

References

VICENTE-VILLARDÓN, J. L., GALINDO VILLARDÓN, M. P., BLÁZQUEZ ZABALLOS, A. (2006): Logistic biplots. In: Greenacre, M, Blasius, J. (Eds.): *Multiple correspondence analysis and related methods*. Chapman & Hall, London, 503–521.

DEMEV, J. R., VICENTE-VILLARDÓN, J. L., GALINDO VILLARDÓN, M. P., ZAMBRANO, A. Y. (2008): Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, 24, 24, 2832–2838.

HERNÁNDEZ SÁNCHEZ, J. C., VICENTE-VILLARDÓN, J. L. (2013): Logistic biplot for nominal data. *arXiv preprint arXiv:1309.5486*.

VICENTE-VILLARDÓN, J. L., HERNÁNDEZ SÁNCHEZ, J. C. (2014). Logistic Biplots for Ordinal Data with an Application to Job Satisfaction of Doctorate Degree Holders in Spain. *arXiv preprint arXiv:1405.0294*.

GOWER, J. C., HAND, D. J. (1995). *Biplots* (Vol. 54). CRC Press.

GOWER, J. C., LUBBE, S. G., LE ROUX, N. J. (2011). *Understanding biplots*. John Wiley & Sons.

HERNÁNDEZ SÁNCHEZ, J. C., VICENTE-VILLARDÓN, J. L., (2014): *NominalLogisticBiplot: Logistic Biplot Representations for Nominal Data*. R Package versión 0.4. `http://cran.r-project.org/web/packages/NominalLogisticBiplot/index.html`

HERNÁNDEZ SÁNCHEZ, J. C, VICENTE-VILLARDÓN, J. L., (2015): *OrdinalLogisticBiplot: Logistic Biplot Representations for Ordinal Data*. R Package versión 0.4. `http://cran.r-project.org/web/packages/OrdinalLogisticBiplot/index.html`

Cluster Analysis and Distance Stability in Multidimensional Scaling

José Fernando Vera

Abstract Stability or sensibility analysis is an important topic in data analysis that has received little attention in the application of Multidimensional Scaling, for which the available approaches are formulated in terms of a coordinate-based analytical jackknife methodology. Although in MDS the prime interest is in rating the stability of the points in the configuration, any coordinate-based methodology may be influenced by imprecisions derived from the inherent required Procrustean method. Hence, this localization error may have influence in the estimated values of stability and cross-validation measures for the analysis of the suitability of any adjusted model, and in particular when some parameters are fix by the problem. In this work, a special leave-n-out cross-validation procedure is proposed to study stability in MDS in terms of the estimated distances, which is not influenced by Procrustean errors. Although the proposed procedure is not a coordinate-based method, a stable configuration can be proposed related to the best approach to the dispersion of the leave-n-out procedure in terms of clustered Euclidean distances.

Keywords

Multidimensional Scaling; jackknife; stability; analysis of dispersion; Cross-Validation; Cluster Difference Scaling

References

DE LEEUW, J. and MEULMAN, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification 3*, 97– 112.

José Fernando Vera
Department of Statistics and O.R. University of Granada, Spain, e-mail: jfvera@ugr.es

HEISER, W. J. and GROENEN, P. J. F. (1997). Cluster differences scaling with a within cluster loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika, 62*, 529–550

VERA, J.F., MACÍAS, R. and ANGULO, J.M., (2008). Non-stationary spatial covariance structure estimation in oversampled domains by cluster differences scaling with spatial constraints. *Stochastic Environmental Research and Risk Assessment, 22*, 95–106.

Multi-Dimensional Scaling of Sparse Block Diagonal Similarity Matrix

Tadashi Imaizumi

Abstract When a $n \times n$ similarity matrix ,**S**, of n objects is given,it is valuable to find the relationship of objects hidden in **S**.Then one model and method to be applicable is Multidimensional scaling (MDS).In MDS as the dimensionality reduction techniques,n objects are represented as n points in lower, usually 2 dimensional space. However, MDS models and methods have some difficulty for analyzing a block diagonal similarity matrix of large n. So, the method of MDS for analyzing the block diagonal matrix will be proposed. Similarity values between two objects are represented as distances between two points in MDS. To handle block diagonal similarity matrix in the proposed model, the following are assumed: (i) each object of n objects belongs to one cluster of G clusters,(ii) given similarity values within same cluster are represented as distances between corresponding points in that cluster, (iii) and the similarity values between two objects belonging to different clusters will be represented by distances between two cluster. The configuration and clusters of n objects will be obtained by minimizing a loss function that have two terms, one is for disparity within cluster and the other is for disparity between clusters. Application to real data set are also presented.

Keywords

Sparse; clustering

Tadashi Imaizumi
Tama University, Japan, e-mail: imaizumi@tama.ac.jp

The Classification and Visualization of Twitter Trending Topics Considering Time Series Variation

Atsuho Nakayama

Abstract The purpose of this study was to examine word clustering in detecting Twitter trending topics considering time series variation. Thus, we collected Twitter entries about new products based on specific sentiment or interest expressions. To identify market trends, analysis of consumer tweet data has received much attention. It is significant to consider time series variation of trending topics when we perform word clustering to detect trending topics on Twitter. Personal concerns will be influenced by new product strategies, such as marketing communication strategies, and will change as time passes. In this study, we sought to detect time series variation in topics about new products by classifying words into clusters based on the co-occurrence of words in Twitter entries. We classified the words extracted from the tweet data using non-negative matrix factorization for dimensionality reduction of the vector space model. Then, we proposed a visualization method for text classification to assist interpretation of the results by using multidimensional scaling model. It is said that joint uses of MDS and cluster analysis are often desirable. In this paper, we showed that the joint use of them also worked well in this case.

Keywords

Cluster analysis; multidimensional scaling; Text analysis

Atsuho Nakayama
Department of Business Administration, Tokyo Metropolitan University, Japan, e-mail: atsuho@tmu.ac.jp

References

KRUSKAL, J. B. (1977). The Relationship Between Multidimentional Scaling and Clustering. In J. Van Ryzin (Ed.): *Classification and Clustering*. Academic Press, New York, 17–44.

LEE, D.D. and SEUNG, H.S. (2000). Algorithms for Non-Negative Matrix Factorization. In K. T. Leen, T. G. Dietterich, and V. Tresp (Eds.): *Advances in Neural Information Processing Systems* 13. MIT Press, Cambridge, MA, 556–562.

SAWAKI, M. and HAGITA, N. (1996). Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning. *IEICE Transactions on Information and Systems*, E79-D (5), 491–497

A study on the overlapping cluster analysis for the large data

Satoru Yokoyama

Abstract Several overlapping cluster analysis models have been suggested. ADCLUS/MAPCLUS are for one-mode two-way data, INDCLUS is for two-mode three-way data.Moreover, a model for one-mode three-way data is suggested by the author and co-researchers.These models were applied to various kinds of data by several researchers.However most of these researches were dealing with data which has at most 15 objects, there are almost no research to deal with larger data.The author applies the overlapping cluster analysis to real data which has the large number of objects like marketing data, and examines the algorithmic problems or the representation of the result. Specifically, the author considers the algorithmic problems.MAPCLUS is the most famous and general algorithm for overlapping cluster analysis.This algorithm consists of two stages, one is an alternating least squares approach, the other is a combinatorial optimization.When the data which has large number of objects are analyzed using MAPCLUS, the combinatorial optimization takes much time, however, it is large possibility that the VAF is not improved enough.The author investigate this problem.

Keywords

Overlapping cluster analysis; proximity data; big data

References

ARABIE, P. and CARROLL, J. D. (1980): MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, 45, 211-235.

Satoru Yokoyama
Teikyo University, Tokyo, Japan, e-mail: satoru@main.teikyo-u.ac.jp

YOKOYAMA, S. (2014): Analysis of marketing data using overlapping cluster analysis and its problems [summary]. The 42nd Annual *Meeting of the Behaviormetric Society of Japan*, pp. 66-67. (in Japanese)

Exhaustive biomarker selection techniques

Hans A. Kestler

Abstract We present a methodology for rapid biomarker selection in supervised and unsupervised classification. We utilize a new feature traversal strategy that exploits the tight interaction of cross-validation experiments and nearest neighbor classification and clustering. The new approach approximately improves the runtime by a factor of ten in contrast to a conventional implementation. The same elements that lead to an acceleration of the cross-validation for nearest neighbor classifiers can also be utilized for giving a lower and upper bound on this classification model.

Keywords

exhaustive classification; nearest neighbour; cross-validation bound

Hans A. Kestler
Leibniz Institute for Age Research, Jena, Germany, e-mail: hkestler@fli-leibniz.de

AUC-based splitting criteria for random survival forests

Andreas Ziegler, Marvin Wright, and Matthias Schmid

Abstract Since their introduction in 2001, random forests have become a successful technique for statistical learning and prediction. Ishwaran et al. (2008) extended the original method for classification and regression by proposing a random forest technique for right-censored time-to-event outcomes ("random survival forests"). We present a new AUC-based splitting criterion for random survival forests that is inspired by the concordance index ("Harrell's C") for survival data. Using simulation studies and real-world data, we compare the proposed splitting criterion to traditional methods such as logrank splitting. The performance of the new criterion is evaluated with regard to sample size and censoring rate, and also w.r.t. various tuning parameters such as the forest size and the number of predictor variables selected in each split. AUC-based splitting criteria are implemented in the R package `ranger`, which represents a versatile software environment for random survival forests in both high- and low-dimensional settings.

Andreas Ziegler
Universität zu Lübeck, Germany, e-mail: `ziegler@imbs.uni-luebeck.de`

Marvin Wright
Universität zu Lübeck, Germany, e-mail: `wright@imbs.uni-luebeck.de`

Matthias Schmid
Rheinische Friedrich-Wilhelms-Universität Bonn, Germany, e-mail: `schmid@imbie.meb.uni-bonn.de`

Ensembles of selected classifiers applied to genomic data

Asma Gul, Berthold Lausen, Zardad Khan, and Osama Mahmoud

Abstract We discuss our recent proposals to improve cancer classification based on genomic data. After preprocessing the microarray data we use a recent proposal to improve feature selection of microarray data based on a proportional overlapping score. Using as benchmark of several micro array data sets we compare random forests and new classification methods based on ensembles of selected *k*-nearest neighbours and tree classifiers. Moreover, we illustrate our approach by the classification between invasive and noninvasive colorectal cancer.

Keywords

feature selection; ensemble methods; classifier selection

References

MAHMOUD, O., HARRISON, A.P., PERPEROGLOU, A., GUL, A., KHAN, Z., METODIEV, M., LAUSEN, B. (2014), A feature selection method for classifica-

Asma Gul
University of Essex, United Kingdom, e-mail: agul@essex.ac.uk

Berthold Lausen
University of Essex, United Kingdom, e-mail: blausen@essex.ac.uk

Zardad Khan
University of Essex, United Kingdom, e-mail: zkhan@essex.ac.uk

Osama Mahmoud
University of Essex, United Kingdom, e-mail: ofamah@essex.ac.uk

tion within functional genomics experiments based on the proportional overlapping score, *BMC Bioinformatics*, 15 (1). p. 274.

GUL, A., KHAN, Z., MAHMOUND, O., MIFTAHUDDIN, M., ADLER, W., PERPEROGLOU, A.; LAUSEN, B. (2015, to appear), Ensemble of k-Nearest Neighbour Classifiers for Class Membership Probability Estimation, *Proceedings of ECDA2014*, Springer, Berlin.

KHAN, Z., GUL, A., MAHMOUND, O., MIFTAHUDDIN, M., PERPEROGLOU, A., ADLER, W., LAUSEN, B. (2015, to appear),An Ensemble of Optimal Trees for Class Membership Probability Estimation, *Proceedings of ECDA2014*, Springer, Berlin.

Comparing the performance of non-parametric change point detection methods for capturing response concordance.

Jedelyn Cabrieto, Francis Tuerlinckx, and Eva Ceulemans

Abstract Response concordance is a key concept in the behavioral sciences. It can be defined as the occurrence of changes in response patterning (change in mean) and/or response synchronization (change in covariation) in a multivariate time series. Revealing response concordance can be viewed as a change point detection problem, where the number of change points is unknown a priori. To solve this problem, DeCon was recently developed, which detects change points by combining a moving windows approach and robust PCA. Yet, in the literature, several other methods have been proposed that employ other non-parametric tools: E-divisive, Multirank and KCP. The relative performance of all these methods for capturing response concordance is still unknown, however. Therefore, we compare E-divisive, Multirank, KCP and Decon, through extensive simulations. Specifically, we use the simulation settings of Bulteel et al. implying changes in mean and in correlation structure and those of Matteson et al. implying different numbers of (noise) variables. KCP emerged as the best method in almost all settings. However, in case of two or more noise variables, only DeCon performed adequately.

Keywords

Response Concordance; Synchronization; Change Point Detection

Jedelyn Cabrieto
KU Leuven, Belgium, e-mail: jed.cabrieto@ppw.kuleuven.be

Francis Tuerlinckx
KU Leuven, Belgium, e-mail: francis.tuerlinckx@ppw.kuleuven.be

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

References

BULTEEL, K., CEULEMANS, E., THOMPSON, R., WAUGH, C., GOTLIB, I., TUERLINCKX, F. and KUPPENS, P. (2014): DeCon: A tool to detect emotional concordance in multivariate time series data of emotional responding. *Biological Psychology*, 98, 29–42.

Matteson, D. and James, N. (2014): A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, 109:505, 334–345.

Time Series Rolling-Window Cluster Analysis on geological data

Carlo Drago, Fabio Matano, and Germana Scepi

Abstract Time series analysis is currently applied to geological data, both to historical data derived by natural archives and to cumulative data, adding up over time in the case of on-going monitoring projects. In the last case data are continuously updated and so the statistical analysis need to take into account this peculiar characteristic. In particular it could be necessary to perform a rolling window analysis that is useful to detects structural changes over time. The algorithm we propose works performing a rolling cluster analysis on different moving windows. In that way we are able to detects different clusters over time. In particular we can identify the different clusters related to the different windows observed on times and the changes which can occur. In this way, by considering the different windows, we can measure the level of stability of the clusters. We will work both on simulated time series and toward on an application based on a concrete real case. The contributions of our statistical approach for geological applications are the following: (1) recognition of qualitative changes within a given dataset, (2) characterization of the random process that describes the evolution of the data, (3) recognition of cycles.

Keywords

Time Series; Rolling Analysis

Carlo Drago
Università degli Studi "Niccolo Cusano" Telematica Roma, Italy, e-mail: c.drago@mclink.it

Fabio Matano
Istituto per l'Ambiente Marino Costiero (IAMC), Consiglio Nazionale delle Ricerche (CNR), Italy, e-mail: fabio.matano@cnr.it

Germana Scepi
Università degli Studi di Napoli "Federico II", Italy, e-mail: scepi@unina.it

Analysis of Influence Scores for Detecting a Change Point

Kuniyoshi Hayashi and Koji Kurihara

Abstract Detecting the point at which the population parameters for target data change is closely related to the success of preventing security failure, discovering new knowledge from the target time series, and performing statistical diagnostics at each time point. Therefore, the change point method has received extensive attention from many researchers in various fields. Thus far, to obtain valuable statistical diagnostics at each time point, we have developed a framework for detecting change points using influence scores calculated by the influence function and then tested the framework by applying our method to some numerical examples. In those studies, we only calculated the average influence scores along the timeline and searched for the time point that gave the largest value among these scores. Therefore, we have not focused on prediction analysis of the movement of the average influence scores. In this paper, by applying the analysis of time series to the average influence scores, we analyze their movement from the viewpoint of prediction. We numerically show the predicted results for the movement on the basis of the idea of autoregressive models. Finally, we present the usefulness of them in detecting change points using influence scores.

Keywords

Influence function; Time series analysis

Kuniyoshi Hayashi
Graduate School of Environmental and Life Science, Okayama University, Japan Science and Technology Agency, CREST, Japan,e-mail: k-hayashi@ems.okayama-u.ac.jp

Koji Kurihara
Graduate School of Environmental and Life Science, Okayama University, Japan Science and Technology Agency, CREST, Japan, e-mail: kurihara@ems.okayama-u.ac.jp

References

HAYASHI, K. and KURIHARA, K. (2014): Detecting a Change Point Using Statistical Sensitivity Analysis Based on the Influence Function. In: Proceedings of the *Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems*. SCIS & ISIS 2014, Kitakyushu, Japan, 506–511.

Focused graphical model estimation

Lourens Waldorp, Sara Jahfari, Eugen Pircalabelu, and Gerda Claeskens

Abstract We develop a focused information criterion for graphical models. The proposed method can be tailored to specific research questions in the sense that the selection of the edges in the graph is done in such a way that a user-specified focus is estimated with a small mean squared error. Penalized estimation methods can be included in case of large graphs that contain more nodes than observations. The method is applied to fMRI data which measure brain activity in a number of brain regions, during a certain time period. Graphical models are used to determine brain connectivity. Such models may include autoregressive times series components, or pool data via random effects. Examples illustrate the method.

Lourens Waldorp
University of Amsterdam, The Netherlands, e-mail: L.J.Waldorp@uva.nl

Sara Jahfari
University of Amsterdam, The Netherlands, e-mail: sara.jahfari@gmail.com

Eugen Pircalabelu
Research Centre for Operations Research and Business Statistics (ORSTAT), Leuven, Belgium e-mail: eugen.pircalabelu@kuleuven.be

Gerda Claeskens
KU Leuven, Belgium, e-mail: gerda.claeskens@kuleuven.be

Seriation benchmarking environment for different permutations and measures of goodness

Innar Liiv

Abstract Seriation is an unsupervised data mining technique that reorders objects into a sequence along a one-dimensional continuum to make sense of the whole series (Liiv, 2010). Clustering assigns objects to groups, whereas seriation assigns objects to a position within a sequence. An online web-based environment `http://data.ttu.ee/seriation/` is presented, which enables researchers to test and evaluate different matrix permutations on different seriation problems with different seriation measures of goodness for those permutations. Besides contributing to scrutinizing seriation research results for improved repeatability, it fosters researchers to identify interesting cross-discipline patterns (Liiv et al., 2012). Problems can be queried and results can be submitted using a simple REST API. Initially supported data structures include two-way one/two-mode binary matrices (e.g. supporting symmetric and asymmetric similarity matrices and other two-way/two-mode general matrices).

Keywords

seriation, matrix reordering, benchmarking

References

LIIV, I. (2010): Seriation and matrix reordering methods: an historical overview. *Statistical Analysis and Data Mining*, 3, 70–91.

Innar Liiv
Department of Informatics, Tallinn University of Technology, Estonia, e-mail: `innar.liiv@ttu.ee`

LIIV, I., OPIK, R., UBI, J. STASKO, J. (2012): Visual matrix explorer for collaborative seriation. Wiley Interdisciplinary Reviews: *Computational Statistics*, 4(1), 85–97.

Representable Hierarchical Clustering Methods for Asymmetric Networks

Facundo Memoli

Abstract This work introduces the generative model of representability for hierarchical clustering methods in asymmetric networks, i.e. the possibility to describe a method through its action on a collection of networks called representers. We characterize the necessary and sufficient structural conditions needed on these representers in order to generate a method which is scale preserving and admissible with respect to two known axioms [1,2]. Based on this result, we construct the family of cyclic clustering methods. Moreover, we show that every representable clustering method can be factored into two maps and, building on this fact, we present an algorithm to implement cyclic clustering methods. Finally, we illustrate the value of cyclic methods by clustering a network of United States' economy.

Keywords

Clustering; Asymmetric; Networks

References

[1] CARLSSON, G. and MEMOLI, F. (2012): Classifying Clustering Schemes. *Journal of Foundations of Computational Mathematics*.

[2] CARLSSON, G., MEMOLI, F., RIBEIRO, A. and SEGARRA, S. (2013): Axiomatic construction of clustering on asymmetric networks, 2013. http://arxiv.org/abs/1301.7724

Facundo Memoli
Mathematics Department, The Ohio State University, United States, e-mail: memoli@math.osu.edu

Social Differentiation of Cultural Taste and Practice in Contemporary Japan: Nonhierarchical Asymmetric Cluster Analysis

Miki Nakai

Abstract In sociological debate on cultural consumption, it has been considered that cultural consumption and lifestyles represent a person's taste and preference, which establish boundaries among social class. This is well-known Bourdieu's hypothesis of cultural reproduction. On the other hand, another hypothesis postulates that omnivorous cultural taste pattern has become prevailing among the socially privileged. The aim of this paper is to explore the stratification of cultural taste based on survey data on cultural lifestyle. Nonhierarchical asymmetric cluster analysis is used to gain better understanding how various cultural consumption relate to one another among some subgroups of respondents in terms of social class and gender. Drawing on detailed information on a broad range of cultural consumption practices of the 2,915 respondents from a nationally representative sample in Japan collected in 2005, the paper demonstrates that there are some notable dissimilarity in cultural participation practices between males and females, as well as among social class. The main factor of cultural taste segmentation seems to be social class, especially among women. The results obtained from the present approach are compared to those obtained from the asymmetric multidimensional scaling. Results are basically consistent.

Keywords

cultural taste; social class; Nonhierarchical Asymmetric Cluster Analysis

New Zealand
Ritsumeikan University, Japan, e-mail: mnakai@ss.ritsumei.ac.jp

References

NAKAI, M. (2011): Class and Gender Differences in Cultural Participation: Asymmetric Multidimensional Scaling of Cultural Consumption. In: Proc. of the *Annual Conference of the German Classification Society*, 246.

OKADA, A. and IMAIZUMI, T. (1997): Asymmetric Multidimensional Scaling of Two-Mode Three-Way proximities. *Journal of Classification*, 14, 195–224.

OKADA, A. and YOKOYAMA S. (2013): Nonhierarchical Asymmetric Cluster Analysis. In: T. Minerva, I. Morlini and F. Palumbo (Eds.): Books of Abstract, CLADAG 2013, 353–356.

An Algorithm of Nonhierarchical Asymmetric Cluster Analysis

Akinori Okada and Satoru Yokoyama

Abstract A method of nonhierarchical asymmetric cluster analysis has been introduced earlier by the authors. The method tries to find clusters or groups of objects where each cluster consists of a dominant object and the other less dominant objects for the predetermined number of clusters. The earlier procedure examines all ${}_nC_k$ combinations to find the optimum results, where n is the number of objects and k is the number of clusters, and is not efficient. A new procedure, which is more efficient than the earlier one, is introduced. The method is characterized by analyzing the skew-symmetry of two conjugate similarities weighted by the sum of the two similarities. While the cluster analysis would analyze similarities among objects, the present method analyzes $(s_{ik} - s_{ki}) \times (s_{ik} + s_{ki})$, where s_{ik} is the similarity from non-dominant object i to object k, which is the dominant object of cluster k. Non-dominant objects belongs to cluster k where $(s_{ik} - s_{ki}) \times (s_{ik} + s_{ki})$ is maximized. The Equation tells that the skew-symmetry between a dominant object and a less dominant objects is weighted more as the sum of the similarities becomes larger. The new procedure is compared with the earlier one, and applied to real data.

Keywords

ACLUSKEW, Asymmetry, Cluster analysis, Nonhierarchical, Skew-symmetry

Akinori Okada
Tama University, Graduate School of Management and Information Sciences, Japan, e-mail: okada@rikkyo.ac.jp

Satoru Yokoyama
Teikyo University, Department of Business Administration, e-mail: satoru@yokoyamalab.org

References

KAUFMAN, L. and ROUSSEUW, P. J. (1990): *Finding Groups in Data*. John Wiley & Sons, New York.

MACQUEEN, J. B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. Proceeding of the *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.

OKADA, A. and YOKOYAMA, S. (in press): *Asymmetric CLUster Analysis Based on SKEW-symmetry: ACLUSKEW*. In: I. Morlini, T. Minerva, and M. Vichi (Eds.): Advances in statistical models for data analysis. Springer-Verlag, Heidelberg.

Advances in clustering asymmetric proximity data

Donatella Vicari

Abstract In several applications data are intrinsically asymmetric as it frequently happens in case of preferences, exchanges (e.g. import-export, brand switching), migration data, confusion data. Clustering such asymmetric proximity data is a challenging task because in addition to the amount of the exchanges also the directions of the exchanges or flows deserve to be investigated. A model is presented which relies on the decomposition of the asymmetric proximity matrix into *symmetric* and *skew-symmetric* effects both decomposed in *within* and *between* cluster effects due to the partition of the objects. Specifically, clusters of objects are identified which mainly play a role of origins (destinations) of the exchanges directed towards (from) some other groups of objects and form closed systems of internal exchanges. For each object two sets of weights (*between* and *within* cluster weights) are estimated so that the observed imbalances between objects are fitted as differences between the corresponding weights. An extension to the three-way case is also presented when asymmetric proximity data are available from different occasions or sources. The model is fitted in a least-squares framework and an appropriate algorithm is provided.

Keywords

Clustering; Skew-symmetry; Least-squares

Donatella Vicari
Dipartimento di Scienze Statistiche - Sapienza Università di Roma, Italy, e-mail: donatella.vicari@uniroma1.it

References

SAITO, T. and YADOHISA, H. (2005): *Data analysis of asymmetric structures*. Advanced Approaches in Computational Statistics. Marcel Dekker. New York.

VICARI, D. (2014): Classification of Asymmetric Proximity Data. *Journal of Classification*, 31, 386–420.

VICARI, D. (2015): CLUSKEXT: CLUstering model for SKew-symmetric data including EXTernal information. *Advances in Data Analysis and Classification*, in press.

Correcting Jaccard and other similarity indices for chance agreement in cluster analysis

Ahmed Najeeb Albatineh

Abstract Correcting a similarity index for chance agreement requires computing its expectation under fixed marginal totals of a matching counts matrix. For some indices, such as Jaccard, Rogers and Tanimoto, Sokal and Sneath, and Gower and Legendre the expectations cannot be easily found. We show how such similarity indices can be expressed as functions of other indices and expectations found by approximations such that approximate correction is possible. A second approach is based on Taylor series expansion. A simulation study illustrates the effectiveness of the resulting correction of similarity indices using structured and unstructured data generated from bivariate normal distributions.

Keywords

Similarity indices; Matching counts matrix; Correction for chance agreement; Jaccard index; Cluster analysis; Comparing partitions

References

ALBATINEH A.N., NIEWIADOMSKA-BUGAJ M., MIHALKO D.P. (2006). On similarity indices and correction for chance agreement. *J Classif*, 23:301–313

ALBATINEH A.N., NIEWIADOMSKA-BUGAJ M., (2011). MCS: a method for finding the number of clusters. *J Classif*, 28. doi:10.1007/s00357-010-9069-1

ALBATINEH A.N. (2010). Means and variances for a family of similarity indices used in cluster analysis. *J Stat Plan Inference*, 140:2828–2838

Ahmed Najeeb Albatineh
Kuwait University, Kuwait, e-mail: aalbatineh@hsc.edu.kw

CZEKANOWSKI J. (1932). “Coefficient of racial likeness” und “durchschnittliche Differenz”. *Anthropologischer Anzeiger*, 14:227–249

DICE L.R. (1945). Measures of the amount of ecological association between species.*Ecology*, 26:297–302

FOWLKES E.B., MALLOWS C.L. (1983). A method for comparing two hierarchical clusterings. *J Am Stat Assoc*, 78:553–569

GOWER J.C., LEGENDRE P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J Classif*, 3:5–48

HUBÁLEK Z. (1982). Coefficients of association and similarity based on binary (presence-absence) data: an evaluation. *Biol Rev*, 57:669–689

HUBERT L., ARABIE P. (1985). Comparing partitions. *J Classif*, 2:193–218

JACCARD P. (1908). Nouvelles recherches sur la distribution florale. *Bull Soc Vaudoise Sci Nat*, 44:223–270

JACCARD P. (1912). The distribution of the flora of the alpine zone. *New Phytol*, 11:37–50

JANSON S., VEGELIUS J. (1981). Measures of ecological association. *Oecologia*, 49:371–376

MILLIGAN G., COOPER M. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar Behav Res*, 21:441–458

MILLIGAN G., SOON S., SOKOL L. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure.*Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1), 40–47

MOREY L., AGRESTI A. (1984). The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educ Psychol Meas*, 44:33–37

RAND W. (1971). Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*, 66:846–850

ROGERS D.J., TANIMOTO T.T. (1960). A computer program for classifying plants. *Science*, 132:1115–1118

Application of spatial median of Weber and positional formulation of TOPSIS method to the assessment of financial condition of local government units

Agnieszka Bernadetta Kozera, Aleksandra Łuczak, and Feliks Wysocki

Abstract When assessing financial condition of local government units there may appear features with outliers or strong asymmetry. Such cases do not easily yield themselves to classical methods of construction a synthetic index since a single outlier would in the final classification incorrectly assign an excessively high or low rank to the examined object. Accordingly, when necessary, robust methods should be applied in typological studies of object ranking, including those that utilize the L1-norm (cf. Rousseeuw, Leroy 1987, Vardi, Zhang 2000). Some such methods directly use the spatial median of Weber, characterized by high degree of robustness to outliers. The aim of the paper is to present the application potential of the spatial median of Weber in constructing a synthetic development index based on positional formulation of TOPSIS method. Weber median based transformations of feature values account not only for outliers, but also for possible relationship between examined features. The TOPSIS method first calculates Euclidean distances of each given object from the ideal and anti-ideal of development and then, on these calculations, builds the synthetic index (Hwang, Yoon 1981). This approach is used in the assessment of financial condition of local government units in Poland in 2013.

Keywords

Weber median; TOPSIS method; financial condition

Agnieszka Bernadetta Kozera
University of Life Sciences of Poznan, Poland, e-mail: akozera@up.poznan.pl

Aleksandra Łuczak
University of Life Sciences of Poznan, Poland, e-mail: luczak@up.poznan.pl

Feliks Wysocki
University of Life Sciences of Poznan, Poland, e-mail: wysocki@up.poznan.pl

References

HWANG C.L., YOON K. (1981). *Multiple attribute decision-making: Methods and applications*. Springer, Berlin.

VARDIardi, Y. and ZHANG, C.H. (2000). The multivariate L1-median and associated data depth. *Proc. National Academy of Science*, 97(4), 1423–1426.

ROUSSEEUW P.J., LEROY A. M. (1987). *Robust regression and outlier detection*. John Wiley and Sons, New York.

Distance measures based on the probabilistic information of the data with applications in classification problems

Gabriel Martos Venturini

Abstract There are several distance measures widely used in Data Analysis, such as Euclidean, Cosine, or Manhattan distances just to name a few. These distances are adapted to different data types: Euclidean distance for real data, Jaccard distance for binary data and so on. However, they do not take into account the underlying statistical distribution of data, with the only exception of Mahalanobis distance, that assumes a Gaussian distribution, and it is intended to measure the distance from a point to the center of the distribution.
In this paper we introduce two distances for multivariate data that explicitly integrate information from the underlying distribution: The Cumulative Distribution Function distance and the Minimum Work Statistical distance (that generalizes the Mahalanobis distance). For both of them, the distance between two points does not solely depend on the Euclidean distance between them. We propose too plug-in estimators for these distances for the case where the distribution is unknown (the usual one). The proposed distances are tested using standard classification algorithms like SVM's or k-NN classifiers to solve classification problems with outstanding results.

Keywords

Probability-based distances; classification

Gabriel Martos Venturini
Universidad Carlos III de Madrid, Spain, e-mail: gamartos@est-econ.uc3m.es

Tree-Robinsonian Dissimilarities

Pascal Préa and François Brucker

Abstract We present in this paper a new characterization of dissimilarities whose associated clusters form a totally balanced hypergraph (called tree-Robinsonian dissimilarities). Moreover, any weakly indexed totally balanced hypergraph can be associated with a tree-Robinsonian dissimilarity. Formally speaking, a dissimilarity d on a set X is tree-Robinsonian if there exists a linear order $<$ on X such that for all $i, j, k \in X$ with $i \leq j \leq k$, $d(j,k) \leq max(d(i,j), d(i,k))$. These dissimilarities generalize Robinsonian dissimilarities (defined by $d(j,k) \leq d(i,j)$ and $d(i,j) \leq d(i,j)$. They are a particular case of weak hierarchies. Adapting totally balanced hypergraphs algorithms, one can recognize, approximate and represent these dissimilarities in $O(|X|^3)$.

Keywords

Dissimilarities; hypergraphs; seriation

References

BANDELT, H.J. and DRESS, A.W.M. (1989), Weak hierarchies associated with similarity measures: an additive clustering technique, *Bull. math. Biology* 51, 133–166

Pascal Préa
Ecole Centrale Marseille, LIF Technopole de Chateau Gombert, France, e-mail: pascal.prea@lif.univ-mrs.fr

François Brucker
Ecole Centrale Marseille, LIF Technopole de Chateau Gombert, France, e-mail: francois.brucker@centrale-marseille.fr

LEHEL, J. (1985), A characterization of totally balanced hypergraphs, *Discrete math.* 57, 59–65,

ROBINSON, W.S. (1951), A method for chronological ordering of archeological deposits, *American Antiquity* 16, 293– 301

Novel similarity measures for categorical data based on mutability and entropy

Zdenek Sulc and Hana Rezankova

Abstract In the contribution, we propose two novel similarity measures determined for objects characterized by nominal variables, which can be used in multivariate statistical methods, e.g. in cluster analysis. They are based on variability measures for nominal variables — the nominal variance (mutability) and the entropy. Unlike the previously proposed similarity measures, where the frequencies of certain categories by the currently examined variable were usually used, the newly proposed measures take into account frequencies of all categories. We evaluate their performance in cluster analysis on both real and generated datasets with the aim to cover a wide range of possible situations of their use. Results of cluster analysis are compared with results obtained using some of recently proposed similarity measures (occurrence frequency and inverse occurrence frequency measures, Eskins, Lins and Goodall measures), and moreover with a reference similarity measure, the simple matching coefficient. We use four evaluation criteria based on within-cluster variability (both mutability and entropy based). For generation of datasets and for all necessary computations the R software was used. The preliminary results show that the novel similarity measures perform very well in certain situations.

Keywords

similarity measures; nominal variables; cluster analysis

Zdenek Sulc
University of Economics, Prague, Czech Republic, e-mail: zdenek.sulc@vse.cz

Hana Rezankova
University of Economics, Prague, Czech Republic, e-mail: rezanka@vse.cz

References

BORIAH, S., CHANDOLA V. and KUMAR, V. (2008): Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM *International Conference on Data Mining*, SIAM, p. 243–254. Available at: `http://www-users.cs.umn.edu/~sboriah/PDFs/BoriahBCK2008.pdf`.

BACHE, K. and LICHMAN, M. (2013): UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L. and STOLFO, S., V. (2002): A geometric framework for unsupervised anomaly detection. In: *Applications of Data Mining in Computer Security*, p. 78–100.

GOODALL, V.D. (1966): A new similarity index based on probability. *Biometrics*. Vol. 22, No.4, p. 882.

LIN, D. (1998): An information-theoretic definition of similarity. In: ICML '98: *Proceedings of the 15th International Conference on Machine Learning*. San Francisco. Morgan Kaufmann Publishers Inc., p. 296–304.

MORLINI, I. and ZANI, S. (2012): A new class of weighted similarity indices using polytomous variables. In: *Journal of Classification*, Vol 29, No 2, p. 199–226.

ŘEZANKOVÁ, H., LÖSTER and T., HÚSEK, D. (2011): Evaluation of categorical data clustering. In: Advances in Intelligent Web Mastering 3. Berlin. Springer Verlag, p. 173–182.

Correspondence Analysis in Identification of Structure of Performance Measurement Systems of Polish Firms

Barbara Batóg, Jacek Batóg, Wanda Skoczylas, Andrzej Niemiec, and Piotr Waśniewski

Abstract The very important element of every performance measurement system (PMS) according to its definition is determination of crucial areas of evaluation as well as individual and aggregate indicators. Many current researches indicate four main areas of PMS dedicated to: finance, customers, operational activity ana employees. Every area is usually described by 5-8 indicators measured with different frequency. In the paper the Authors applied multivariate correspondence analysis in order to verify the hypothesis that the structure of PMS is related to the following specific attributes of firms: size, economic sector and the way the measurement system is acquired. This relationship is especially important for the business efficiency and firm's development. The correspondence analysis was used because all examined variables were measured on the nominal scale. The source of the data (contingency tables) was a nationwide survey conducted by the CATI method among Polish companies.

Keywords

multivariate correspondence analysis; performance measurement system

Barbara Batóg
University of Szczecin, Poland, e-mail: barbara.batog@wneiz.pl

Jacek Batóg
University of Szczecin, Poland, e-mail: batog@wneiz.pl

Wanda Skoczylas
University of Szczecin, Poland, e-mail: wanda@wneiz.pl

Andrzej Niemiec
University of Szczecin, Poland, e-mail: andrzej.niemiec@wneiz.pl

Piotr Waśniewski
University of Szczecin, Poland, e-mail: piotr.wasniewski@gmail.com

References

HAFFER R. (2011): Samoocena i pomiar wynikw dzialalnosci w systemach zarzadzania przedsiebiorstw. W poszukiwaniu doskonalosci biznesowej. Wydawnictwo Naukowe Uniwersytetu Mikolaja Kopernika, Torun.

NEELY A. (2004): *Business Performance Measurement*. Theory and Practice. Cambridge University Press, Cambridge.

STANIMIR A. (2005): Analiza korespondencji jako narzedzie do badania zjawisk ekonomicznych. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wroclawiu, Wroclaw.

A principal component method to analyze frequency tables connected by contextual variables

Belchin Kostov, Mónica Bécue-Bertaut, and François Husson

Abstract An original statistical methodology is proposed to analyze open-ended questions answered in different languages, provided that the questionnaire includes the same closed questions for all the samples. For every language, the free answers are encoded in the form of a respondents×words lexical table and the closed answers in the form of a quantitative or categorical respondents×variables, called contextual table in this context. Each lexical table is thus coupled with a contextual table, having both the same respondent-rows. The lexical tables, from one language to another, have neither the row-individuals nor the column-words in common while all the contextual tables share the same column-variables. The contextual variables are used as a link between the lexical tables making a global analysis of the free answers possible. This leads to a method, called Multiple Factor Analysis on Generalized Aggregated Lexical Table MFA-GALT, which extends both Correspondence Analysis on a Generalized Aggregated Lexical Table (CA-GALT) and multiple factor analysis for contingency tables (MFACT). An application on a real survey shows that this method supplies outputs easy to interpret. The method can be applied with profit in other fields provided that the data are coded into a similar structure.

Belchin Kostov
Universitat Polititècnica de Catalunya, Department of Statistics and Operational Research, Spain, e-mail: belchin3541@gmail.com

Mónica Bécue-Bertaut
Universitat Polititècnica de Catalunya, Department of Statistics and Operational Research, Spain, e-mail: monica.becue@upc.edu

François Husson
Agrocampus-Ouest, Applied Mathematics Department, France, e-mail: husson@agrocampus-ouest.fr

Keywords

Multiple Factor Analysis for generalized aggregated lexical tables; Open-ended questions, Mixed data tables

References

BÉCUE-BERTAUT, M. and PAGÈS, J. (2004): A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, 45, 481–503.

BÉCUE-BERTAUT, M., PAGÈS, J. and KOSTOV, B. (2014): Untangling the influence of several contextual variables on the respondents lexical choices. A statistical approach. *SORT*, 38, 285–302.

BCUE-BERTAUT, M. and PAGS, J. (2014): Correspondence analysis of textual data involving contextual information: Ca-galt on principal components. *Advances in Data Analysis and Classification*. DOI 10.1007/s11634-014-0171-9.

A method for recoding ordinal variables

Odysseas Moschidis and Theodore Chadjipadelis

Abstract The similarity of individuals with respect to a number of ordinal variables is the central aim of this research. We consider the case of analyzing by multivariate correspondence analysis k variables measured on ordinal scale for N subjects. In this case we transform each variable to a suitable number of binary variables and we analyze the derived matrix using as similarity measure the X2 metric. As a consequence, we face the loss of information of the initial matrix because ordered variables are treated as categorical. In this paper we propose a method of recoding the initial variables take into account the ordinal scale they are measured. We argue that the proposed transformation gives more accurate results concerning the method used. By the proposed method a variable measured on k categories transformed into a variable of m categories through assigning a probability for each value instead of recoding each value to a new binary variable.

Keywords

Multiple correspondance analysis; coding

Odysseas Moschidis
Dept of Businness Administration, University of Macedonia, Greece, e-mail: fmos@uom.gr

Theodore Chadjipadelis
Professor of Applied Statistics School of Political Sciences, Aristotle University Thessaloniki, Greece, e-mail: chadji@polsci.auth.gr

Big Data Scaling through Metric Mapping: Correspondence Analysis in Very High Dimensional Spaces

Fionn Murtagh

Abstract We present new findings in regard to data analysis in very high dimensional spaces. We use dimensionalities up to around one million. A particular benefit of Correspondence Analysis is its suitability for carrying out an orthonormal mapping, or scaling, of power law distributed data. Power law distributed data are found in many domains. Correspondence factor analysis provides a latent semantic or principal axes mapping. Our experiments use data from digital chemistry and finance, and other statistically generated data. We discuss applications to analysis in high dimensions of social media, and general domains where the objective is to induce, efficiently, the hierarchical clustering of the data.

Keywords

data piling; data analytics; high dimensional data analysis

References

HALL, P., MARRON, J.S. and NEEMAN, A. (2005): Geometric Representation of High Dimension, Low Sample Size Data. *Journal of the Royal Statistical Society Series B*, 67, 427–444.

MURTAGH, F. (2009): The Remarkable Simplicity of Very High Dimensional Data: Application of Model-Based Clustering. *Journal of Classification*, 26, 249–277.

MURTAGH, F. and CONTRERAS, P. (2015): Random Projection towards the Baire Metric for High Dimensional Clustering. In: A. Gammerman et al. (Eds.): *Pro-*

Fionn Murtagh
Department of Computing Goldsmiths, University of London, UK; Department of Computing and Mathematics, University of Derby, UK, e-mail: fmurtagh@acm.org

ceedings SLDS 2015, Symposium on Learning and Data Sciences, Lecture Notes in Artificial Intelligence Volume 9047, 424–431.

Biplot-based visualisations to explore relationships between pneumonia and nasopharyngeal pathogens

Johané Nienkemper-Swanepoel, Sugnet Lubbe, Niël le Roux, Emilee Smith, Heather Zar, and Mark Nicol

Abstract Positive identification of pathogens with pneumonia is vitally important in medicine. Data are considered where a number of nasopharyngeal pathogens were measured as present or absent in patients suffering from pneumonia. For each diseased patient, two healthy controls were also observed. The aim of the analysis is to investigate differences between diseased and non-diseased patients and to identify pathogens specifically related to pneumonia. A selection of graphical multivariate procedures for dichotomous variables are illustrated and compared. These techniques are derived from traditional biplot-based multivariate methods such as Multiple Correspondence Analysis (MCA), Categorical Canonical Variate Analysis (CatCVA), Analysis of Distance (AoD) and Classification Trees. Extensions of the existing methodology are proposed for dealing with the specific complexities of the data at hand. The focus is to provide suitable graphical visualisations to guide

Johané Nienkemper-Swanepoel
Department of Genetics, Stellenbosch University, South Africa, e-mail: jnienkie@gmail.com

Sugnet Lubbe
Department of Statistical Sciences, University of Cape Town, South Africa, e-mail: sugnet.lubbe@uct.ac.za

Niël le Roux
Department of Statistics and Actuarial Science, Stellenbosch University, South Africa, e-mail: njlr@sun.ac.za

Emilee Smith
Centre for Infectious Disease Epidemiology and Research, University of Cape Town, South Africa, e-mail: emilee.smith@uct.ac.za

Heather Zar
Department Paediatrics and Child Health, Red Cross Hospital, Medical Research Council unit on Child and Adolescent Health, University of Cape Town, South Africa, e-mail: heather.zar@uct.ac.za

Mark Nicol
Division of Medical Microbiology, University of Cape Town, South Africa, e-mail: Mark.Nicol@uct.ac.za

and facilitate the statistical classification of patients when multivariate dichotomous pathogen data are available. Both symmetrical (where no distinction is made between the roles played by different variables) and asymmetrical (where a distinction is made between dependent and independent variables) visualisations are considered.

Keywords

binary classification; biplots; multiple correspondence analysis

An efficient implementation of adjacency-constrained hierarchical clustering of a band similarity matrix

Alia Dehman, Guillem Rigaill, Pierre Neuvial, and Christophe Ambroise

Abstract Uncovering a latent blockwise dependency structure from an observed similarity matrix is a common problem in numerous genomic applications, such as the detection of Linkage Disequilibrium (LD) blocks in Genome-Wide Association Studies (GWAS). This problem may be tackled by incorporating an adjacency constraint to a standard hierarchical agglomerative clustering methods. However, such an approach is intrinsically quadratic both in time and space in the number p of items. This is problematic for genomic applications, where p is often of the order of 10^5 to 10^6. It is possible to reduce this complexity to $\mathscr{O}(p(h+log(p)))$ in time and $\mathscr{O}(ph)$ in space, where h is the maximum size of the blocks. We propose an implementation of this algorithm, and illustrate the interest of this novel algorithm in GWAS applications, where h is several orders of magnitude smaller than p.

Keywords

hierarchical clustering; adjacency constraint

Alia Dehman
Université d'Evry Val d'Essonne, France, e-mail: alia.dehman@genopole.cnrs.fr

Guillem Rigaill
Université d'Evry Val d'Essonne, France, e-mail: guillem.rigaill@evry.inra.fr

Pierre Neuvial
Université d'Evry Val d'Essonne, France, e-mail: pierre.neuvial@genopole.cnrs.fr

Christophe Ambroise
Université d'Evry Val d'Essonne, France, e-mail: christophe.ambroise@genopole.cnrs.fr

Tissue-aware age prediction from DNA methylation data

Marcelo Ferreira and Ivan Costa

Abstract Aging is a complex process, which affects every cell in the organism and leads to the deterioration of body functions over time. Similarly, cells undergo a process of replicative senescence during culture expansion. It has been recently demonstrated that organism aging and cell senescence are associated with the loss or gain of methylation of particular DNA regions. DNA methylation arrays allow the measurement of methylation levels of up to 450000 regions at once and enabled genome wide association studies with hundreds of patients at distinct ages. Penalized regression methods have been used for selecting signatures with relevant DNA methylation sites, which were used to predict patient age with high accuracy. However, most of the works either consider only one kind of human tissue or do not take into account that aging occurs differently on different human tissues. We propose tissue-aware modeling using penalized (LASSO and Elastic Net) regression approaches for detection of DNA methylation signatures of aging by including categorized variables representing the tissue types. Experiments with publicly available methylation data sets show that the accuracy of the predictions can be improved.

Keywords

Aging; DNA methylation; Tissue-aware modeling

Marcelo Ferreira
Federal University of Paraiba, Brazil, e-mail: marcelorpf@gmail.com

Ivan Costa
RWTH Aachen University Hospital, Germany, e-mail: ivan.costa@rwth-aachen.de

Bootstrap test of ordered RIG for multiple testing in genomics of Quantitative Trait Loci in yeasts

Evgeny Mirkes, Thomas Walsh, Edward J. Louis, and Alexander N. Gorban

Abstract The problem of identification of pair of loci associated with heat tolerance in yeasts is considered. Interactions of Quantitative Trait Loci (QTL) in heat selected yeast are analysed by comparing them to an unselected pool of random individuals. Data on individual F12 progeny selected for heat tolerance, which have been genotyped at 25 locations identified by sequencing a selected pool, are re-examined. 960 individuals were genotyped at these locations and multi-locus genotype frequencies were compared to 172 sequenced individuals from the original unselected pool. We use Relative Information Gain (RIG) for analysis of associations between loci. Correlation analysis in many pairs of loci requires multi testing methods. Two multi testing approaches are applied for selection of associations: False Discovery Rate (FDR) in the version suggested by J.D. Storey and R. Tibshirani and specially developed Bootstrap Test of ordered RIG (BToRIG). We show that a statistical analysis of entropy and information gain in genotypes of a selected population can reveal further interactions than previously seen. Importantly this is done in comparison to the unselected populations genotypes to account for inherent biases in the original population.

Evgeny Mirkes
University of Leicester, Department of Mathematics, United Kingdom, e-mail: em322@le.ac.uk

Thomas Walsh
Centre for Genetic Architecture of Complex Traits, University of Leicester, Leicester LE1 7RH, United Kingdom, e-mail: tw164@le.ac.uk

Edward J. Louis
Centre for Genetic Architecture of Complex Traits, University of Leicester, Leicester LE1 7RH, United Kingdom, e-mail: ejl21@le.ac.uk

Alexander N. Gorban
Department of Mathematics, University of Leicester, Leicester, LE1 7RH, United Kingdom, e-mail: ag153@le.ac.uk

Keywords

Multitesting; Information gain; Genetic Interaction

References

LITI, G. and LOUIS, E. J. (2012). Advances in quantitative trait analysis in yeast. *PLoS genetics*, 8(8), e1002912.

PARTS, L., CUBILLOS, F. A., WARRINGER, J., JAIN, K., SALINAS, F., BUMPSTEAD, S. J., MOLIN, M., ZIA, A., SIMPSON, J. T., QUAIL, M. A., MOSES, A., LOUIS, E. J., DURBIN, R. and LITI, G. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome research*, 21(7), 1131–1138.

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genome wide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.

Discovering modulators of gene regulation by local energy statistics

Teppei Shimamura and Yusuke Matsui

Abstract Discovering proteins called modulators which control the activity of transcription factors is one of the most important issues for understanding tissue- and content-specific gene regulation. The state-of-the-art methods including MINDy and GEM have been developed for this issue. However, these approaches cannot detect modulators of complicate relationships such as synergistic regulation of gene expression by multiple transcription factors. Here we develop a statistical approach to predict modulators by analyzing a three-way interaction between multiple transcription factors, their target genes, and modulators based on a local version of distance correlation. We also propose a nonparametric approach for testing significance of the local distance correlation by constraint random sampling. We illustrate the performance of the proposed method through both simulation and read data analysis.

Keywords

Biostatistics; Gene expression; Distance correlation

Teppei Shimamura
Nagoya University Graduate School of Medicine, Japan, e-mail: shimamura@med.nagoya-u.ac.jp

Yusuke Matsui
Nagoya University Graduate School of Medicine, Japan, e-mail: ymatsui@med.nagoya-u.ac.jp

References

Wang et al. (2009): Genome-wide identification of post-translational modulators of transcription factor activity in human B cells.*Nat Biotechnol*., 27(9), 829–839.
Babur et al. (2010): Discovering modulators of gene expression. *Nucleic Acids Res.*, 38(17), 5648–5656.

Haplotype estimation, Haplotype Block Identification and statistical analysis for DNA data

Makoto Tomita

Abstract In early research, linkage disequilibrium analysis for SNP data is particularly important. Methods of trait mapping based on theories of linkage disequilibrium analysis have been developing quickly in recent years. In DNA sequences, domain "hotspots" exist at which recombinations have occurred briskly. On the other hand, large domains with infrequent recombinations in which linkage disequilibrium is maintained also exist. Such domain called a "haplotype block" or "LD block". Although the value of represents one of the disequilibrium parameters important for identifying haplotype blocks. In this paper, we introduced haplotype estimation, haplotype block method and so on.

Keywords

DNA marker, spatial clustering, haplotype

References

TOMITA, M., HATSUMICHI, M. and KURIHARA, K. (2008): Identify LD Blocks Based on Hierarchical Spatial Data. *Computational Statistics and Data Analysis*, vol. 52, pp. 1806–1820.

TOMITA, M., KURIHARA, K. and MOON, S. H. (2012): An Application to Select Tag Loci by Using Hierarchical Structures of DNA Markers. *Journal of the Korean Data Analysis Society*, vol. 13, pp. 2749–2762.

Makoto Tomita
Tokyo Medical and Dental University, Japan, e-mail: mtomita@ism.ac.jp

TOMITA, M., HASHIMOTO, N. and TANAKA, Y. (2011): Association Study for the Relationship Between a Haplotype or Haplotype Set and Multiple Quantitative Responses. *Computational Statistics and Data Analysis*, vol. 55, pp. 2104–2113.

Music classification from the user side

Nicolas Greffard and Pascale Kuntz

Abstract Music classification has known a renewed interest since the digital shift. In particular, a large literature is devoted to the automatic classification of music collections into musical genres. The music pieces are described by complex vectors which catch the information at different levels (e.g. spectral flux, pitch, mood) and various (un)-supervised algorithms have been developed to cluster them. However these algorithms are rarely used at the individual level. How do humans classify their own digital music collections which are still growing? Surprisingly, despite the importance of music in everyday life the individual processes of music classification have not undergone extensive research yet and the few existing studies are based on ethnographic surveys. Here, we study real-life music collections stored on laptop hard drives - which remain a support frequently used by young adults -. We analyze the topologies of the collections. We compute the statistical distributions of some basic structural measures and we propose a classification of the various observed topologies based on an adapted hierarchical clustering algorithm. A treemap-based visualization helps to interpret the classes. An additional study of how listeners named their folders contributes to a better understanding of the factors which guide the music collection organization.

Keywords

Music Collection Classification; Music Information Retrieval; digital humanities

Nicolas Greffard
LINA-DUKe, University of Nantes, France, e-mail: nicolas.greffard@univ-nantes.fr

Pascale Kuntz
LINA-DUKe,University of Nantes, France, e-mail: pascale.kuntz@univ-nantes.fr

Transportation clustering

Guillaume Guex, Théophile Emmanouilidis, and François Bavaud

Abstract Optimal transportation performs a soft clustering of origins into destinations. The present study is motivated by the search of optimal children to schools assignment plan for the city of Lausanne, where the pedestrian network, children locations and school capacities are entirely known. The (network) simplex algorithm of linear programming, dealing with metric costs, is demanding in presence of many origins and destinations, and yields one particular solution among the generally multiple optima. We investigate the properties and behavior of a regularized version of the problem, easier to compute, allowing to spatially highlight the boundaries between school recruitment basins (clusters), as well as to determine the embarkation and disembarkation costs of the dual formulation. It consists in minimizing the *free energy*

$$F[P] = \text{trace}(PD') + T\,K(P||P^{\infty})$$

where $P = (p_{ig})$ is the assignment (joint probability of children location i and school g), with fixed margins, $D = (d_{ig})$ is the journey to school pedestrian cost, and $K(P||P^{\infty})$ is the Kullback-Leibler divergence, measuring the dependence between children locations and schools. Minimizers P^T are determined iteratively, and match the *doubly constrained gravity flow* of quantitative geography. In the low-temperature limit $T \to 0^+$, the resulting children-to-school clustering constitutes a soft, constrained K-medoid solution, analogous to a K-means constrained solution, but for D metric rather than squared Euclidean. It enables an arguably original cartography of transportation costs, together with an enhanced visualization of boundaries between school attendance zones, characterized by large conditional school entropies at given children locations.

Guillaume Guex
University of Lausanne, Switzerland, e-mail: gguex@unil.ch

Théophile Emmanouilidis
University of Lausanne, Switzerland, e-mail: temmanouilidis@unil.ch

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

Keywords

optimal transportation; constrained clustering; regularized minimization; pedestrian network

Unsupervised classification of perfusion imaging data using multiple equivalence tests

Fuchen Liu, Yves Rozenholc, and Charles-André Cuénod

Abstract Perfusion imaging using DCE-MRI, -CT or -US allows accessing the functional modifications of the micro-vascularization. Hence it appears as a major marker to follow therapies using anti-angiogenesis treatment in cancer. However, perfusion analysis suffers from low Signal to Noise Ratio (SNR) due to either the need of reducing X-ray dose or the trade-off between the spatial resolution and the temporal resolution. Constructing manually spatial Region Of Interest (ROI) improves the SNR. However analyzes suffer from lack of homogeneity inside these manually drawn ROIs. We propose an automatic construction of ROIs, which preserves functional homogeneity together with local properties of images. Our procedure is based on hierarchical clustering using equivalence test p-values as dissimilarity measure between neighbors. In this framework, neighboring homogeneous voxels are iteratively aggregated into clusters by ensuring the decrease of the dissimilarity. Type I error, which is the probability of aggregating two clusters by mistake, is controlled during the iterations and allows an automatic choice of the number of clusters. At the end of the process, removing neighborhood specifications allows to merge homogenous clusters occurring at long distance, which may come from similar tissue existing in the body at different locations.

Keywords

unsupervised classification; equivalence test; perfusion imaging

Fuchen Liu
LRI, Université Paris Descartes, France, e-mail: fuchen.liu@parisdescartes.fr

Yves Rozenholc
MAP5, Université Paris Descartes, France, e-mail: yves.rozenholc@parisdescartes.fr

Charles-André Cuénod
LRI, Université Paris Descartes, France, e-mail: charles-andre.cuenod@egp.aphp.fr

Using taxonomies and aggregate rankings for measuring research impact.

Boris G. Mirkin and Mikhail A. Orlov

Abstract Research impact of a scientist can be defined using these dimensions: level of scientific results; level of participation in the organization of sciences; level of knowledge transfer; and level of technology innovations. We skip the last item because of difficulties in finding data on technology innovations. We present an approach to measuring the level of research results, for the first time, to the best of our knowledge. The approach involves a taxonomy of the research domain. The level of results is evaluated according to the taxonomy ranks of the subjects that have emerged or have been crucially transformed due to the results by the scientist under consideration. Two conventional dimensions, (a) citation metrics and (b) merit metrics, are taken too. To aggregate individual criteria we develop an in-house criteria-weighting method Linstrat. The method's criterion is that the strata are as tight as possible. We take a sample of thirty scientists in data analysis and machine learning, and the ACM Computing Classification System 2012. Empirical results: (a) Hirsch citation index gets a zero weight; (b) when combining the scales for Citation, Merit, and Taxonomic rank, the latter gets the weight of 80%; (c) the three dimensions are almost uncorrelated.

Keywords

level of research results; taxonomy; weighting criteria

Boris G. Mirkin
National Research University Higher School of Economics Moscow RF and Birkbeck University of London, Russian Federation, e-mail: mirkin@dcs.bbk.ac.uk

Mikhail A. Orlov
National Research University Higher School of Economics Moscow RF, Russian Federation, e-mail: ormian@mail.ru

Cause Related Marketing: a qualitative and quantitative Analysis on Pinkwashing

Gabriella Schoier and Patrizia De Luca

Abstract Cause related marketing is the process of formulating and implementing marketing activities characterized by an offer from the firm to contribute an amount to a specified cause when customers buy their products. In this context many papers deal with problem of greenwash and discuss on the mediation between green consumer confusion and green perceived risk. On the contrary as far as we know there is no literature on the influence on pinkwash on pink trust. The term pinkwash has been created in 2002 by American Breast Cancer Organization for the campaign "think before you pink"; in this campaign it has pointed out that there are companies and organizations that care about breast cancer by promoting a pink ribbon product but at the same time produces manufactures and or sell products linked to the disease. The object of this study is to see the effects of pinkwash on pink consumer confusion and pink perceived risk.We test different hypotheses: the association between pink wash and pink consumer confusion , pinkwash and pink perceived risk, pinkwash and pink trust, pink consumer confusion and pink trust,pink perceived risk and pink trust. In order to do this a cluster analysis based on an on line questionnaire and some bootstrap simulation have been performed.

Keywords

pinkwashing, cluster analysis, bootstrap methods

Gabriella Schoier
University of Trieste, Italy, e-mail: gabriella.schoier@deams.units.it

Patrizia De Luca
University of Trieste, Italy, e-mail: patrizia.deluca@deams.units.it

Hierarchical Disjoint Non-negative Factor Analysis

Maurizio Vichi

Abstract Hierarchical Disjoint Non-negative Factor Analysis (HDFA) is a new latent factor model here proposed for modelling Composite Indicators (CIs). CIs, in general, are multidimensional concepts described by at least a theoretical construct (factor) which is related to a set of measured variables. Frequently CIs have a hierarchical structure, i.e., they are characterized by a set of factors each one corresponding to disjoint, or nested subsets of variables. Hierarchical Confirmatory Factor Analysis can be used to assess the hierarchical structure of the CIs. A Structural Equation Model is in this case used. However, this methodology generally is not easily applicable for the amount of hypotheses to be considered, because the researcher has to know a priori the number of factors and specifically the most relevant relations between variables and factors, and relations between factors of different order in the hierarchy. These assumptions from one side imply the existence of deep theory sustaining the definition of the CI that, in any case needs to be empirically confirmed, and on the other side, represent a limitation in situations -quite realistic and rather recurrent when the researcher has not an exact theory in mind, or the theory demonstrates at least in part of being empirically erroneous or imprecise. Hierarchical Disjoint Factor Analysis is here proposed to model the hierarchical structure of factors that is supposed in part or totally unknown, identifying a reduced set of factors each one related to a disjoint subset of variables. Furthermore, in a definition of a CI each subset of measured variables, used to describe a multidimensional concept, must be internally consistent and reliable, that is, variables related to the factor measure ?consistently? a unique theoretical construct. This implies that variables are concordant with the related factor and loadings must be positive. This last requirement is included as a constraint in the new methodology that for this reason is named Hierarchical Disjoint Non negative Factor Analysis. Properties are discussed for HDNFA. Cross-loadings can also be estimated to increase the fit of the factor model starting from the best HDNFA solution. HDNFA has also the option to constraint a variable to load on a pre-specified factor in order to hypothesize, a

Maurizio Vichi
University La Sapienza, Rome, Italy, e-mail: maurizio.vichi@uniroma1.it

priori, some relations between variables and loadings. An application to optimally identify the dimensions of well-being is used to illustrate the characteristics of the new methodology. A final discussion completes the paper.

Keywords

Exploratory Factor Analysis; Confirmatory Factor Analysis; Disjoint Factor Analysis; Sparse loading matrix

The IFCS Cluster Benchmark Data Repository

Friedrich Leisch and IFCS Task Force on Benchmarking

Abstract Numerous clustering algorithms have been proposed over the last decades. Most papers presenting new cluster algorithms include a demonstration and/or performance comparison on artificial and real world data sets. In the machine learning community the UCI repository provides a collection of data sets which are routinely used for benchmarking supervised learning problems, i.e. regression and classification. Classification data are often also used for benchmarking cluster algorithms. However, a good classification data set need not be good for evaluation of clustering methods. Especially harder classification tasks may be impossible to solve correctly by clustering. A data repository specializing in providing benchmarking data for clustering is currently missing. The new IFCS repository provides tools for collection and distribution of such data sets and will allow for real data as well as data generators for artificial data. A questionnaire accompanies each data upload and provides meta-data such as background information, previous usages, relationships between variables, known group structures, why the data should be clustered, etc. The ultimate goal is to have a collection of data sets which are routinely used for cluster algorithm comparison such that results are reproducible and comparable across scientific publications by different authors.

Keywords

cluster analysis; benchmarking; data repository

Friedrich Leisch
University of Natural Resources and Life Sciences Vienna, Austria, e-mail: Friedrich.Leisch@boku.ac.at

IFCS Task Force on Benchmarking, e-mail: nema.dean@gmail.com

A Statistical Framework for Hypothesis Testing in Real Data Benchmark Experiments

Anne-Laure Boulesteix

Abstract In computational sciences, including computational statistics, machine learning, and bioinformatics, it is often claimed in articles presenting new supervised classification methods that the new method performs better than existing methods on real data, for instance in terms of error rate. However, these claims are often not based on proper statistical tests and, even if such tests are performed, the tested hypothesis is not clearly defined and poor attention is devoted to the type I and type II errors. With the aim to fill this gap, a proper statistical framework for hypothesis tests which compare the performances of supervised learning methods based on several real data sets with unknown underlying distributions is provided in Boulesteix et al. (2015). In the first part of this talk, I will give a statistical interpretation of such tests and outline a simple method of determining the number of data sets to be included in a benchmark experiment to reach an adequate power. These methods are illustrated through three comparison studies from the literature and an exemplary benchmark experiment using gene expression microarray data. In the second part of the talk, I will discuss potential extensions of this statistical framework to unsupervised classification and present a first application to a set of exemplary datasets based on objective performance measures.

Keywords

benchmarking, comparison study, statistical inference

Anne-Laure Boulesteix
Institute for Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-University of Munich, Germany, e-mail: boulesteix@ibe.med.uni-muenchen.de

References

BOULESTEIX, A.L., HABLE, R., LAUER, S. and EUGSTER, M. (2015). A statistical framework for hypothesis testing in real data comparison studies. *The American Statistician*, DOI:10.1080/00031305.2015.1005128.

Benchmarking a New Bayesian Disease Mapping Cluster Detection Method

Nema Dean, Duncan Lee, and Craig Anderson

Abstract A benchmarking study is used to evaluate a new method proposed to identify the geographical extent of spatially contiguous high- or low-risk clusters of areal units in the context of disease mapping. The aim of disease mapping is to estimate the spatial pattern in disease risk for a set of areal units. Existing methods make use of Bayesian hierarchical models with spatially smooth conditional autoregressive priors to estimate disease risk. However, smoothness assumptions mean they are unable to directly identify clusters of areas with similar risks. The new method introduces a two-stage approach: initially identifying proposed cluster structures using a spatially constrained hierarchical agglomerative clustering approach, followed by fitting a Poisson log-linear model where the optimal cluster structure and the spatial pattern in disease risk is estimated via a Markov Chain Monte Carlo algorithm. Both stages are assessed: the first stage in terms of absolute performance in identifying useful cluster structures and then the combined approach is compared to a selection of competing methods. The issue of realistic simulated data settings for the disease mapping context, some real data results and recommendations about the future application of such models are presented for discussion.

Keywords

Cluster Analysis Benchmarking; Disease Mapping; Conditional Autoregressive Priors

Nema Dean
University of Glasgow, United Kingdom, e-mail: Nema.Dean@glasgow.ac.uk

Duncan Lee
University of Glasgow, United Kingdom, e-mail: Duncan.Lee@glasgow.ac.uk

Craig Anderson
University of Technology, Sydney, Australia, e-mail: Craig.Anderson@uts.edu.au

A general method for fuzzy partitioning and component analysis

Maria Brigida Ferraro, Paolo Giordani, and Maurizio Vichi

Abstract A general method for two-mode simultaneous reduction of units and variables of a data matrix is introduced. It consists in a convex linear combination of Reduced K-Means (RKM) and Factorial K-Means (FKM). Both methodologies involve principal component analysis for variables and K-Means for units, even though RKM aims at maximizing the between-clusters deviance without imposing any condition on the within-clusters deviance, while FKM aims at minimizing the within-clusters deviance without imposing any condition on the between one. It follows that RKM and FKM complement each other. In order to take advantage of both methods a convex linear combination of RKM and FKM is proposed. Furthermore, the fuzzy approach to clustering is adopted because of its flexibility in handling the real world complexity and uncertainty. A fast Alternating Least Squares algorithm is introduced and its performance is investigated by simulated and real data.

Keywords

Factorial K-Means; Reduced K-Means; Linear convex combination; Fuzzy approach

Maria Brigida Ferraro
Sapienza University of Rome, Italy e-mail: mariabrigida.ferraro@uniroma1.it

PAolo Giordani
Sapienza University of Rome, Italy e-mail: paolo.giordani@uniroma1.it

Maurizio Vichi
Sapienza University of Rome, Italy, e-mail: maurizio.vichi@uniroma1.it

The multivariate power method: a fast iterative algorithm for repeated dimension reduction

Alfonso Iodice D'Enza, Michel Van de Velden, and Patrick J.F. Groenen

Abstract The visual exploration of big data often requires interactivity as well as the possibility to update an existing solution as new data becomes available as in data flows. For exploratory data visualization can be very useful for such data and is often provided by dimension reduction methods. The core of popular dimension reduction techniques, such as principal component analysis (PCA) and multiple correspondence analysis (MCA) lies in the computation of an eigenvalue or a singular value decompositions. Here we propose an efficient implementation of MCA, called powerCA, that exploits the sparseness of the data for matrix transformations needed, uses fast iterative methods for the computation of eigenvalues, and makes use of smart initializations in case of a repeated analysis. The aim of this paper is to extend the applicability of MCA to computationally demanding applications such as the visualization of streaming text and web-log data, and computer intensive methods such as the bootstrap-based sensitivity analysis.

Keywords

Power method; eigendecomposition algorithms; Multiple Correspondence Analysis; Data visualizations

Alfonso Iodice D'Enza
Università di Cassino e del Lazio Meridionale, Italy, e-mail: iodicede@gmail.com

Michel van de Velden
Erasmus University Rotterdam, Netherlands, e-mail: vandevelden@ese.eur.nl

Patrick J.F. Groenen
Erasmus University Rotterdam, Netherlands, e-mail: groenen@ese.eur.nls

On acceleration methods for Alternating Least Squares algorithm

Michio Sakakihara, Msahiro Kuroda, Yuichi Mori, and Msaya IIzuka

Abstract The alternating least squares (ALS) algorithm is an iterative solution method for the nonnegative matrix factorization such as $min_W H|V - WH|_F^2$ with $W_{ij} > 0$ and $H_{ij} > 0$, which is utilized in e.g., PARAFAC, the document clustering, the molecular pattern discovery and image analysis. We have proposed some acceleration algorithms based on the vector epsilon method for the EM iterations and the ALS algorithm in the nonlinear principal components analysis. In this paper, we discuss those acceleration methods for the factorizations including Lee and Seung's multiplicative update approach. Moreover, we consider the acceleration method by a block wise formulation. In some numerical experiments we evaluate the efficiency of the proposed acceleration methods.

Keywords

ALS; acceleration; vector epsilon method

Michio Sakakihara
Okayama University of Science, Japan, e-mail: sakaki@mis.ous.ac.jp

Msahiro Kuroda
Okayama University of Science, Japan, e-mail: kuroda@soci.ous.ac.jp

Yuichi Mori
Okayama University of Science, Japan, e-mail: mori@soci.ous.ac.jp

Msaya IIzuka
Okayama University of Science, Japan, e-mail: okayama-u.ac.jp

Canonical correlation analysis for three-mode three-way data

Jun Tsuchida and Hiroshi Yadohisa

Abstract Canonical correlation analysis (CCA) is very popular method for investigation of relationships between two sets of variables in many areas. In marketing research and psychometric, we obtain not only two-mode two- way data (object vs. variable) but also three-mode three-way (object vs. variable vs. source). three-mode three way data which have the same sources, we often combine the data into multivariate data or separate each sources from the data before applying CCA, in order to examine the relationship between two sets of variables. However, If there is different relationship between two sets of variable depending on the source, this method is not suitable for interpreting relationship between sets of variables,because this method ignores the difference of sources. In order to overcome this problem , we propose CCA for three-mode three-way data,which imposes the constraints the relationship between sources. In proposed method, we describe the difference in variables between sources, by using the same parameters for the same sources.

Keywords

alternating least squares; dimension reduction; tensor analysis

Jun Tsuchida
Doshisha University, Japan, e-mail: jt.tabakosangyo@gmail.com

Hiroshi Yadohisa
Doshisha University, Japan, e-mail: hyadohis@mail.doshisha.ac.jp

From mixtures of SEMs to mixtures of Double-structure SEMs

Francesca Martella, Marco Alfò, and Paolo Giordani

Abstract Mixtures of structural equation models (SEMs) treat heterogeneity by assuming that, within each mixture component, the observed data are described by a SEM; this in turn, by making use of component-specific individual effects, explains the component-specific relationship between the manifest variables. However, when three-way data structures are considered, that is when the subjects are measured on several manifest variables under different conditions, the model may be extended. For instance, if conditions refer to time occasions, one may also be interested to take into account their impact in the model specification. In this case, the same subject may exhibit different values of the same manifest variable over time occasions and this would not be reflected by the standard mixture of SEMs. After a close look at various types of clustering recently proposed in the three-way data literature, we propose a refinement of the Double-structure SEMs, referred to as mixture of Double-structure SEMs. Thanks to the simultaneous presence of component-specific individual and time occasion effects, this specification allows to consider a general structure of dependence within and between subjects and to take into account individual-specific unobserved heterogeneity.

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy, e-mail: francesca.martella@uniroma1.it

Marco Alfò
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy, e-mail: marco.alfo@uniroma1.it

Paolo Giordani
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy, e-mail: paolo.giordani@uniroma1.it

Keywords

Double-structure mixture SEM; three-way data

References

ARMINGER, G. and STEIN, P. (1997): Finite Mixture of Covariance Structure Models with Regressors: Loglikehood Function, Distance Estimation, Fit Indices, and a Complex Example. *Sociological Methods and Research*, 26, 148–182.

DOLAN, C.V. and VAN DER MAAS, H.L.J. (1997): Fitting Multivariate Normal Finite Mixtures Subject to Structural Equation Modeling. *Psychometrika*, 63, 227–253.

GONZALEZ, J. DE BOECK, P. and TUERLINCKX, F. (2008): A Double-Structure Structural Equation Model for Three-Mode Data. *Psychological Methods*, 13, 337–353.

Analysis of Multivariate Longitudinal Data Subject to Nonrandom Dropout

Mai Sherif Hafez, Irini Moustaki, and Jouni Kuha

Abstract Longitudinal data are collected for studying changes across time. We consider multivariate longitudinal data where multiple observed variables, measured at each time point, are used as indicators for theoretical constructs (latent variables) of interest. A common problem in longitudinal studies is dropout, where subjects exit the study prematurely. Ignoring the dropout mechanism can lead to biased estimates, especially when the dropout is nonrandom. Our proposed approach uses latent variable models to capture the evolution of the latent phenomenon over time while also accounting for possibly nonrandom dropout. The dropout mechanism is modeled with a hazard function that depends on the latent variables and observed covariates. Different relationships among these variables and the dropout mechanism are studied via 2 model specifications. The proposed models are used to study people's perceptions of women's work using 3 questions from 5 waves of the British Household Panel Survey.

Keywords

nonignorable dropout; ordinal variables; structural equation modeling

Mai Sherif Hafez
The London School of Economics and Political Science, United Kingdom, e-mail: m.m.hafez@lse.ac.uk

Irini Moustaki
The London School of Economics and Political Science, United Kingdom, e-mail: i.moustaki@lse.ac.uk

Jouni Kuha
The London School of Economics and Political Science, United Kingdom, e-mail: j.kuha@lse.ac.uk

Two-mode K-Spectral Centroid analysis for studying multivariate dynamical processes

Joke Heylen, Iven Van Mechelen, Eiko Fried, and Eva Ceulemans

Abstract Researchers that study dynamic processes, often collect multivariate time profiles, mapping the evolution of a set of variables over time, for multiple subjects. For instance, many clinical studies focus on the differential effect of an intervention on different symptoms, by repeatedly measuring symptom severity. To parsimoniously describe the huge information in such data and to pursue an insightful overview on how time profiles vary as a function of both subjects and variables, we propose two-mode K-Spectral Centroid (2M-KSC) analysis. This method, that combines the key ideas of multi-mode partitioning and one-mode K-Spectral Centroid analysis, simultaneously reduces subjects to subject clusters and variables to variable clusters. This clustering is based on the shape of the time profiles under study, implying that time profiles that correspond to a specific combination of a person cluster and variable cluster are modeled with one specific reference profile, reflecting the typical evolution over time. Furthermore, each time profile receives an amplitude score, indicating its overall intensity relative to its corresponding reference profile. We apply the new 2M-KSC method to time profiles reflecting the intensity of depression symptoms during citalopram treatment.

Joke Heylen
KU Leuven, Belgium, e-mail: Joke.Heylen@ppw.kuleuven.be

Iven Van Mechelen
KU Leuven, Belgium, e-mail: Iven.VanMechelen@ppw.kuleuven.be

Eiko Fried
KU Leuven, Belgium, e-mail: Eiko.Fried@ppw.kuleuven.be

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

Keywords

clustering; time profiles; two-mode partitioning

Rotation involving subsets of variables to achieve an insight into structures of within-person variability

Marieke Timmerman, Eva Ceulemans, and Henk Kiers

Abstract When one or more subjects are assessed at a large number of consecutive measurement occasions on a large number of variables, it is often of interest to identify the major sources of variance across time and, in case of multiple subjects, to express structural differences and similarities between the subjects. Insight into those structure(s) of within-person variability can be obtained using a suitable component analysis variant, like principal component analysis (PCA), switching PCA or clusterwise simultaneous component analysis. The estimated loadings typically have rotational freedom, which may be exploited to facilitate the interpretation. We propose a class of rotation procedures, involving both orthogonal and oblique rotations, that is useful to identify components expressing the source(s) of variance underlying 1. all variables and/or 2. one or more pre-specified subsets of variables and/or 3. single variables. The class is based on the earlier proposed rotation to a partially specified target and Simplimax rotation. It covers, as a special case, a bifactor target rotation. The use and usefulness of the rotation procedures will be illustrated using empirical analyses of multivariate repeated observations from a single individual, as well as of multiple individuals simultaneously.

Keywords

target rotation; orthogonal rotation; oblique rotation

Marieke Timmerman
University of Groningen, Heymans Institute for Psychological Research Psychometrics and Statistics, Netherlands, e-mail: m.e.timmerman@rug.nl

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

Henk Kiers
Heymans Institute for Psychological Research, University of Groningen, Netherlands, e-mail: h.a.l.kiers@rug.nl

Regime change analysis of interval-valued time series with an application to PM10

Carmela Cappelli, Pierpaolo D'Urso, and Francesca Di Iorio

Abstract In several real life and research situations data are collected in the form of intervals. To analyze interval-valued data, usually researchers summarize the original data into single values, such as the centers or the medians of the intervals, but by doing so some important information in the original data is lost such as the range of the interval. In the last years efforts have been done either to extend classical methods or to develop new approaches to deal with interval valued data. This work addresses the problem of detecting change points in interval-valued times series (IVTS). Various methods proposed in the literature consider the case single-valued time series. In order to deal with IVTS we propose to employ a regression tree based approach called A theoretical Regression Tree using a proper distance measure that accounts for the interval structure of the time ordered units. We present the results of simulation studies pertaining to different scenarios and an empirical application to a real interval valued time series that shows the usefulness of the proposed procedure. In particular, the application considers a time series of an air pollutant, the particulate matter (PM10) that is responsible for harmful effects on health.

Keywords

Interval data; change points; regression trees

Carmela Cappelli
University of Naples Federico II, Italy, e-mail: carcappe@unina.it

Pierpaolo D'Urso
University of Rome La Sapienza, Italy, e-mail: pierpaolo.durso@uniroma1.it

Francesca Di Iorio
University of Naples Federico II, Italy, e-mail: fdiiorio@unina.it

Two-sample test with distributional data and detection of differential DNA methylation

Yusuke Matsui and Teppei Shimamura

Abstract We propose an approach to two-sample test with distributional data under the framework of symbolic data analysis and apply to detecting differential DNA methylation sites. The two-sample test is the key issue in genomics and underlying distributions are often with complicated shapes (*e.g.,* multi-modality). Thus traditional two-sample tests based on summary statistics such as mean and variance are insufficient in many cases. We perform statistical test under the null hypothesis that two distributional data is equal and the alternative is different. We construct the test statistic based on Wasserstein metric and make null distributions with the resampling approach. We show that the proposed method outperforms previous methods of two- sample test in the simulation study. We also demonstrate the method to detect the differential methylation sites in glioblastoma multiforme and lower grade glioma using the TCGA datasets.

Keywords

Two-sample test; Symbolic Data Analysis; DNA methylation

Yusuke Matsui
Nagoya University Graduate School of Medicine, Japan, e-mail: ymatsui@med.nagoya-u.ac.jp

Teppei Shimamura
Nagoya University Graduate School of Medicine, Japan, e-mail: shimamura@med.nagoya-u.ac.jp

References

A. IRPINO, R. VERDE (2014). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification*, pp.1–33, Springer Berlin Heidelberg.

E. DIDAY, M. NOIRHOMME-FRAITURE (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.

H. SUZUKI., *et al.* (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet.*.

Detection of singularities in distribution valued data

Masahiro Mizuta and Hiroyuki Minami

Abstract We focus on distribution valued data or distributional data, in which observations (objects) are described by distributions. Simple structures of distribution valued data are normal distributions or unimodal distributions. It is important to develop a method to find out singular objects or outlier whose distributions are not simple. We propose a method to detect singular objects using principal points.We show an application of the proposed method with an important actual dataset: air dose rates measured every 3 seconds by 32 route buses in Fukushima Prefecture. The records of the data include time, latitude, longitude, and air does rate. Total number of records is 13,508,200. We assign 100m-mesh to object, and each object is described by distribution of the air dose. There are 6791 objects. The proposed method can detect singular objects. We investigate them deeply.

Keywords

Principal points; SDA; environmental monitoring data

References

JAPAN ATOMIC ENERGY AGENCY(2014): *Display radiation dose rates in Fukushima Prefecture at real time*. TOPICS Fukushima 12 Dec 2014 No.56 http://fukushima.jaea.go.jp/english/topics/pdf/topics-fukushima056e.pdf

Masahiro Mizuta
Hokkaido University, Japan, e-mail: mizuta@iic.hokudai.ac.jp

Hiroyuki Minami
Hokkaido University, Japan, e-mail: min@iic.hokudai.ac.jp

Interval-valued logistic regression ensemble vs noisy variables and outliers

Marcin Pelka, Aneta Rybicka, and Justyna Brzezińska-Grabowska

Abstract Ensemble approach based on aggregating information provided by different models has proved to be a very useful tool in context of supervised and unsupervised learning for classical data. However this kind of approach can be useful when dealing symbolic data analysis. In symbolic data analysis objects can be described by different types of variables: nominal, ordinal, interval, ratio, interval-valued, multivalued, multivalued with weights, histogram. Interval-valued data arise in many practical situations such as recording interval temperatures at meteorological stations, daily interval stock prices, etc. This paper presents an evaluation study on ensemble logistic regression for symbolic data when dealing noisy variables and outliers.

Keywords

interval-valued data, ensemble learning, logistic regression

References

BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag,

Marcin Pelka
Wroclaw University of Economics, Poland, e-mail: marcin.pelka@ue.wroc.pl

Aneta Rybicka
Wroclaw Univeristy of Economics, Poland, e-mail: aneta.rybicka@ue.wroc.pl

Justyna Brzezinska-Grabowska
University of Economics in Katowice, Poland, e-mail: justyna.brzezinska-grabowska@ue.katowice.pl

Berlin-Heidelberg.
DE SOUZA, R.M.C.R, QUEIROZ, D.C.F, and CYSNEIROS, F.J.A. (2011): Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, vol. 14, issue 3, p. 273–282.
KUNCHEVA, L.I. (2014): *Combining Pattern Classifiers. Methods and Algorithms*. John Wiley & Sons, Chichester.

Mixtures of Hidden Truncation Hyperbolic Distributions

Ryan Browne

Abstract Non-Gaussian distributions are of great utility in model-based clustering applications for which the underlying distribution of the clusters is unknown. A novel formulation of the hyperbolic distribution, referred to as the hidden truncation hyperbolic (HTH) distribution, is presented for use in model-based clustering. The HTH distribution encapsulates several distributions as special or limiting cases including the skew-t and multivariate-t distributions. The HTH distribution is incorporated into a mixture model and an expectation-maximization algorithm for the estimation of model parameters will be presented. Clustering results will be presented for both real and simulated data and the ability of the HTH distribution to model asymmetric data as compared to other skewed distributions which have appeared in the literature will be discussed.

Keywords

finite mixture model; skewness;

Ryan Browne
McMaster University, Department of Mathematics & Statistics, Canada, e-mail: rbrowne@math.mcmaster.ca

Parsimonious multiple scaled mixtures

Brian C. Franczak, Ryan P. Browne, and Paul D. McNicholas

Abstract Mixtures of multiple scaled distributions have garnered increased attention in the last few years. Generalizations of the multivariate-t and shifted asymmetric Laplace distributions have been introduced, and both are shown to exhibit untraditional physical features that can be beneficial when performing cluster analysis. One issue with the aforementioned mixtures is that their covariance structures become highly parameterized as the dimension of the data, p, increases. Since the multiple scaled distributions are formulated using an eigen-decomposed covariance matrix, a natural way to introduce parsimony is by constraining the constituent elements of this decomposition. This leads to two families of mixture models, that we call parsimonious multiple scaled mixtures (PMSMs). Interestingly, the PMSMs can be derived using two different stochastic relationships. As such, in addition to introducing the PMSMs, we compare these two stochastic relationships and discuss their advantages when deriving the parameter estimates for our families of mixture models. We demonstrate the PMSMs' abilities in both clustering and classification applications using simulated and real data sets, and compare their results to those obtained using the current state-of-the-art.

Keywords

multiple scaled distributions; model-based clustering; model-based classification; cluster analysis

Brian C. Franczak
McMaster University, Canada, e-mail: bfrancza@math.mcmaster.ca

Ryan P. Browne
McMaster University, Canada, e-mail: rbrowne@math.mcmaster.ca

Paul D. McNicholas
McMaster University, Canada, e-mail: mcnicholas@math.mcmaster.ca

Mixtures of Coalesced Generalized Hyperbolic Distributions

Cristina Tortora

Abstract A coalesced distribution is a mixture of two distributions with common parameters. In particular, the coalesced generalized hyperbolic distribution (CGHD) is a mixture of a generalized hyperbolic distribution (GHD) and a multiple scaled generalized hyperbolic distribution (MSGHD). A random variable $\mathbf{R}$ follows a CGHD, i.e., $\mathbf{R} \sim \mathscr{CGHD}(\mu, \Sigma, \alpha, \omega_0, \lambda_0, \omega, \lambda)$, if $\mathbf{R} = U\mathbf{Y} + (1-U)\mathbf{S}$, where $\mathbf{Y} \sim \mathscr{GHD}(\mu, \Sigma, \alpha, \omega_0, \lambda_0)$, $\mathbf{S} \sim \mathscr{MSGHD}(\mu, \Sigma, \alpha, \omega, \lambda)$ and $U \in [0,1]$. The GHD is a flexible distribution, capable of handling skewness and heavy tails, and has many well known distributions as special or limiting cases. A random variable $\mathbf{X}$ following a GHD, i.e., $\mathbf{X} \sim \mathscr{GHD}(\mu, \Sigma, \alpha, \omega_0, \lambda_0)$, can be generated via $\mathbf{X} = \mu + W\alpha + \sqrt{W}\mathbf{V}$, where $W \sim \text{GIG}(\omega_0, 1, \lambda_0)$ and $\mathbf{V} \sim \mathscr{N}(\mathbf{0}, \Sigma)$, with GIG denoting a generalized inverse Gaussian distribution. The MSGHD is obtained by decomposing the generalized hyperbolic distribution component scale matrix using an eigen-decomposition of the matrix Σ, and adding a multidimensional weight function $S \sim \mathscr{MSGHD}(\alpha, \Sigma, \mu, \omega, \lambda)$. A finite mixture of coalesced generalized hyperbolic distributions (MCGHDs) assumes that the population is a convex combination of a finite number of CGHDs. The effectiveness of the MCGHDs is illustrated on real and simulated data sets.

Keywords

Generalized hyperbolic distribution; coalesced distribution; multiple scaled distribution

Cristina Tortora
McMaster University, Canada, e-mail: ctortora@mcmaster.ca

References

BROWNE, R. P. and MCNICHOLAS, (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, to appear.

FORBES, F. and WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering. *Statistics and Computing*, 24(6), 971–984.

TORTORA, C., FRANCZAK, B.C., BROWNE, R.P. and MCNICHOLAS, P. D. (2014). A Mixture of Coalesced Generalized Hyperbolic Distributions.*preprint arXiv:1403.2332*.

A simple approach to bi-clustering discrete data

Marco Alfò, Maria Francesca Marino, and Francesca Martella

Abstract Finite Mixtures of Factor Analyzers have been used as model based clustering approach for high-dimensional data. In the last few years, they have been extended to biclustering purposes or to deal with Gaussian and non-Gaussian, continuous, responses, for example by using t and generalized hyperbolic distributions. We propose an extension of Finite Mixtures of Factor Analyzers to allow for simultaneous clustering of subjects and variables when multivariate discrete outcomes are available. We detail the EM algorithm for ML parameter estimation and discuss the performance of the proposed model when applied to binary synthetic and to real count data, coming from Next-generation sequencing assays.

References

LEE, S., HUANG, J.Z. (2014). A biclustering algorithm for binary matrices based on penalized Bernoulli likelihood. *Statistics and Computing*, 24, 429–441.

MARTELLA, F., ALFÒ, M., VICHI, M. (2008). Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics*, 4(1), 3.

MARTELLA, F., ALF'Ò, M., VICHI, M. (2011). Hierarchical mixture models for biclustering in microarray data. *Statistical Modelling*, 11(6): 489–505.

Marco Alfò
Sapienza Università di Roma, Italy, e-mail: marco.alfo@uniroma1.it

Maria Francesca Marino
Sapienza Università di Roma, Italy, e-mail: mariafrancesca.marino@uniroma1.it

Francesca Martella
Sapienza Università di Roma, Italy, e-mail: francesca.martella@uniroma1.it

Modeling mixtures in genomics

Jeanine Houwing-Duistermaat

Abstract Nowadays epidemiological studies comprise multiple omics datasets, that provide insight in biological mechanisms underlying complex diseases. Clustering methods can be applied to find subclasses of patients with similar profiles. For example osteoarthritis (OA) might be characterized by progressive heterogeneous changes in articular cartilage and subchondral bone resulting in either atrophic or hypertrophic OA-subtypes. For efficient treatment it is essential to know the subtype. A second example is the presence of a mutation which causes imbalance of the transcription of the two copies of a gene in affected cartilage of OA patients. In healthy tissue both copies are equally transcribed while in affected tissue imbalance might be present. The amount of imbalance at a locus tends to decrease with the genetic distance to the unknown mutation.
From 21 OA patients, we have RNAseq data from several candidate genes measured in affected cartilage. For heterozygous SNPs the number of reads per allele are observed. To model these count data the negative binomial distribution is typically used to allow for over dispersion. We will use mixtures of negative binomial distributions to identify patients who show imbalance and to model the deviation from allelic balance over the genetic region.

Keywords

Count data; Negative binomial distribution; RNAseq

Jeanine Houwing-Duistermaat
Leiden University Medical Center, Netherlands, e-mail: j.j.houwing@lumc.nl

Mixture model with multiple allocations for clustering spatially correlated gene expression data

Saverio Ranciati, Cinzia Viroli, and Ernst Wit

Abstract Model-based clustering is a technique widely used to group a collection of units into mutually exclusive groups. There are, however, situations in which an observation could in principle belong to more than one cluster. In the context of *next-generation sequencing* (NGS) experiments, for example, the signal observed in the data might be produced by two (or more) different biological processes operating together and a gene could participate in both (or all) of them. We propose a novel approach to cluster NGS discrete data with a mixture model, allowing each unit to belong to potentially more than one group: these multiple allocation clusters can be flexibly defined via a function combining the features of the original groups without introducing new parameters. The formulation naturally gives rise to a 'background group' in which values close to zero can be allocated, acting as a correction for the abundance of zeros that manifest in this type of data. We take into account the spatial dependency between observations, which is described through a latent *conditional auto-regressive* (CAR) model that can reflect different dependency patterns. We assess the performance of our model within a simulation environment and then we apply it to a RNA-seq real dataset.

Keywords

clustering; spatial dependency; multiple allocations

Saverio Ranciati
Name, University of Bologna, Italy, e-mail: saverio.ranciati2@unibo.it

Cinzia Viroli
Name, University of Bologna, Italy, e-mail: cinzia.viroli@unibo.it

Ernst Wit
Name, University of Groningen, Netherlands, e-mail: cinzia.viroli@unibo.it

A finite mixture IRT model for ordinal responses with nonignorable missingness

Silvia Bacci, Francesco Bartolucci, Leonardo Grilli, and Carla Rampichini

Abstract We propose a multidimensional latent class IRT model for the analysis of ordinal responses subject to nonignorable missingness. The missingness mechanism is driven by 2 sets of latent classes summarizing, respectively, the propensity to respond and the abilities measured by the test items. The model allows for both item covariates and examinee covariates and it is fitted by the EM algorithm. The model is illustrated through an application to university student careers, focusing on the grades of first-year exams, where the missingness of a grade is likely to be nonignorable.

Keywords

keyword1; keyword2; keyword3

References

BACCI, S., BARTOLUCCI, F. (2015): A Multidimensional Finite Mixture Structural Equation Model for Nonignorable Missing Responses to Test Items. *Struc-*

Silvia Bacci
Unversity of Perugia, Italy, e-mail: silvia.bacci@unipg.i

Francesco Bartolucci
Unversity of Perugia, Italy, e-mail: francesco.bartolucci@unipg.it

Leonardo Grilli
University of Florence, Italy, e-mail: grilli@disia.unifi.it

Carla Rampichini
University of Florence, Italy, e-mail: rampichini@disia.unifi.it

tural Equation Modeling, DOI: 10.1080/10705511.2014.937376

Dealing with Large Heterogeneity in Sample Surveys

Stefania Capecchi and Domenico Piccolo

Abstract In psychological and sociological surveys, when ordinal options have to be selected, some items receive a large proportion of uncertainty responses due to several causes: sensible questions, inaccurate wording, inexact knowledge of the problem at hand, etc. This situation is even more frequent when the "Don't know" response option is absent in the questionnaire so that people adopt a lazy strategy. Thus, it is highly probable that the frequency distribution of the response manifests a single category as the preferred one; in addition, indecision may be expressed by means of a large heterogeneity among the categories. The main goal of this paper is to investigate situations where uncertainty is a dominant feature of the responses; thus, the structure we discuss is nested in the class of GeCUB models. In addition, with respect to standard analysis of ordinal data, we emphasize the usefulness of the graphical tools in the framework of CUB models and the easiness of interpretation of the role of the components. Then, we examine a real case study where an extreme indecision of the respondents can be analyzed within the above specified approach. A comparison with more consolidated models and some concluding remarks end the paper.

Keywords

Ordinal data; GeCUB models; CUSH models

Stefania Capecchi
University of Naples Federico II, Italy, e-mail: stefania.capecchi@unina.it

Domenico Piccolo
University of Naples Federico II, Italy, e-mail: domenico.piccolo@unina.it

Partial Possibilistic Regression Path Modeling

Rosaria Romano

Abstract Structural equation models (SEMs) aim to estimate a network of causal relationships among latent variables (LVs) defined by blocks of manifest variables (MVs). Under this framework, Partial Possibilistic Regression Path Modeling (PPR-PM) is a method to analyze complex phenomena where there is an additional source of complexity arising from the involvement of human beings.For instance, in the context of subjective measurement, where ordinal data are collected from rating surveys to measure latent concepts. This is achieved by combining the principles of Possibilistic Regression and Quantile Regression. The main idea is to use Quantile Regression to model the relations between each LV and its respective block of indicators andthe Possibilistic Regression to model the relations among the LVs.This choice allows us on the one hand to have a robust measure of the LVs, through the use of the Quantile Regression. Here, in fact, the items used to measure attitudes and preferences often have a skewed distribution. Moreover, the presence of outliers is very common in this context. On the other, PR allow us to take into account the imprecision inherent in systems where human estimation is influential and the observations cannot be described accurately.PR defines the relation between variables through possibilistic linear functions and considers the error due to the vagueness in the relations among the variables as reflected in the model via interval-valued parameters.

Keywords

Path modeling; possibilistic regression; uncertainty

Rosaria Romano
University of Calabria, Italy, e-mail: rosaria.romano@unical.it

References

BOLLEN, K.A. (1989). *Structural equations with latent variables*, Wiley, New York.

ROMANO R., PALUMBO F. (2013). Partial Possibilistic Regression Path Modeling for subjective measurement, *Journal of Methodological and Applied Statistics*, 15, 177–190.

TANAKA, H., GUO, P. (1999). *Possibilistic Data Analysis for Operations Research*, Physica-Verlag, Wurzburg.

WOLD, H. (1975). Modelling in complex situations with soft information, in: *Third World Congress of Econometric Society*, Toronto, Canada.

Cluster Correspondence Analysis

Alfonso Iodice D'Enza, Michel Van de Velden, and Francesco Palumbo

Abstract A new method is proposed that combines dimension reduction and cluster analysis for categorical data. A least-squares objective function is formulated that approximates the cluster by variables cross-tabulation. Individual observations are assigned to clusters in such a way that the distributions over the categorical variables for the different clusters are optimally separated. In a unified framework, a brief review of alternative methods is provided and performance of the methods is appraised by means of a simulation study. The results of the joint dimension reduction and clustering methods are compared with cluster analysis based on the full dimensional data. Our results show that the joint dimension reduction and clustering methods outperform, both with respect to the retrieval of the true underlying cluster structure and with respect to internal cluster validity measures, full dimensional clustering.The differences increase when more variables are involved and in the presence of noise variables.

Keywords

clustering; dimension reduction; correspondence analysis

Alfonso Iodice D'Enza
Università di Cassino e del Lazio Meridionale, Italy, e-mail: iodicede@unicas.it

Michel van de Velden
Erasmus University of Rotterdam, Netherlands, e-mail: vandevelden@ese.eur.nl

Francesco Palumbo
Federico II University of Naples, Italy, e-mail: fpalumbo@unina.it

Clustering and Dimensional Reduction for mixed variables

Henk Kiers, Donatella Vicari, and Maurizio Vichi

Abstract Various methods have been proposed for clustering and simultaneous dimensional reduction of objects and variables of a two-way two-mode data matrix, but usually only for quantitative variables. The present paper describes a proposal for a new general model including Factorial K-means and Reduced K-means as special cases and that also handles qualitative variables, and mixtures of qualitative, quantitative and ordinal variables. An efficient algorithm has been designed for this, and some results will be presented, as well as potential problems. Furthermore, the method will be related to other methods, like GROUPALS which also caters for qualitative variables.

Henk Kiers
University of Groningen, Netherlands, e-mail: h.a.l.kies@rug.nl

Donatella Vicari
Sapienza University of Rome, Italy, e-mail: donatella.vicari@uniroma1.it

Maurizio Vichi
Sapienza University of Rome, Italy, e-mail: Maurizio.vichi@uniroma1.it

A simultaneous analysis of dimension reduction and clustering with correlated error variables

Michio Yamamoto

Abstract Recently, a lot of procedures to find a cluster structure of individuals and a subspace for the clustering simultaneously have been developed. However, those procedures have a drawback that independent error variables are assumed in the models. Thus, if there are correlated variables that do not contribute to a cluster structure, existing methods fail to find the cluster structure. To overcome the drawback, a joint procedure is introduced, which intends to eliminate the effect of the correlation by estimating a subspace for the correlation simultaneously with the subspace for the cluster structure. The introduced model has a general formulation that subsumes existing methods. Thus, it is needed to select an optimal model by deciding the values of several tuning parameters. For the model selection, we propose a strategy that focuses on a good separation of clusters and a degree of confidence for the estimated cluster structure.

Keywords

clustering; dimension reduction

References

DE SOETE, G. and CARROLL, J.D. (1994): *K-means clustering in a low-dimensional Euclidean space*. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy (Eds.): New Approaches in Classification and Data Analysis. Springer, Heidelberg, 212–219.

Michio Yamamoto
Kyoto University, Japan, e-mail: michyama@kuhp.kyoto-u.ac.jp

VICHI, M. and KIERS, H.A.L. (2001): Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37, 49–64.

YAMAMOTO, M. and Hwang, H. (2014): A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, 41, 115–129.

Stability analyses in human exposure to background air pollution in urban environments

Álvaro Gómez-Losada and José F. Vera-Vera

Abstract Estimation of the stability in measurements obtained from monitoring networks is essential for solving spatial interpolation problems (e.g. air pollution exposure models). In general, station location within air quality monitoring networks (AQMNs) represents a source of error in pollution exposure estimates. Background air pollution (lowest concentration profiles) in cities is of great importance since adverse health effects have been associated with pollution levels falling below air quality standards considered safe for health. A combined methodology is proposed for the spatio-temporal analysis of air pollution data. First, the background air pollution at monitoring sites from one city for key pollutants is estimated; then the analysis of stability of the non-parametric estimation of the spatial dispersion is made on the basis of the estimated background pollution. Extrapolation (to unobserved sites) in terms of a background pollution analyses can be made from the spatial dispersion. The proposed methodology is applied to the Seville (Spain) AQMN for five criteria air pollutants. The obtained results from the stability analyses indicate that in general, this AQMN is optimally configured to provide enough quality background pollution information. However, background pollution exposure estimations from some monitoring locations are more stable with respect to ambient perturbations than others.

Keywords

background pollution; analysis of stability; spatial dispersion; time series

Álvaro Gómez-Losada
University of Seville Spain, e-mail: alvgomlos@alum.us.es

José F. Vera-Vera
University of Granada, Spain, e-mail: jfvera@ugr.es

References

De LEEUW, J., MEULMAN, J. (1986): A special jackknife for multidimensional scaling. *Journal of Classification*, 3, 97–112.

GÓMEZ-LOSADA, Á., LOZANO-GARCÍA, A., PINO-MEJÍAS, R., CONTRERAS-GONZÁLEZ, J. (2014): Finite mixture models to characterize and refine air quality monitoring networks. *Science of the Total Environment*, 485-486, 292–299.

SAMPSON, P.D. and GUTTORP, P. (1992): Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87, 108–119.

An extension of the Adjusted Rand Index for fuzzy partitions

Antonio D'Ambrosio, Sonia Amodio, Carmela Iorio, and Roberta Siciliano

Abstract To compare clustering partitions, Rand index (RI) and Adjusted Rand index (ARI) are commonly used for measuring the agreement between partitions. Both these external validation indexes aim to analyze how close is a cluster to a reference (or to prior knowledge about the data) by counting corrected classified pairs of elements.When the aim is to evaluate the solution of a fuzzy clustering algorithm, the computation of these measures require converting the soft partitions into hard ones. It is known that different fuzzy partitions describing very different structures in the data can lead to the same crisp partition and consequently to the same values of these measures. We compare the existing approaches to evaluate the external validation criteria in fuzzy clustering and we propose an extension of the ARI for fuzzy partitions based on the *normalized degree of concordance*. Through use of real and simulated data, we analyze and evaluate the performance of our proposal.

Keywords

Clustering evaluation; fuzzy partitions; external validity measures

Antonio D'Ambrosio
University of Naples Federico II, Italy, e-mail: antdambr@unina.it

Sonia Amodio
University of Naples Federico II, Italy, e-mail: sonia.amodio@unina.it

Carmela Iorio
University of Naples Federico II, Italy, e-mail: carmela.iorio@unina.it

Roberta Siciliano
University of Naples Federico II, Italy, e-mail: roberta@unina.it

References

HUBERT, L., and ARABIE, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.

HULLERMEIER, E., RIFQI, M., HENZGEN, S., and SENGE, R. (2012). Comparing fuzzy partitions: A generalization of the Rand index and related measures. *Fuzzy Systems*, IEEE Transactions on, 20(3), 546–556.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.

An integrated formulation for merging mixture components based on posterior probabilities

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

Abstract In parametric clustering, groups are usually formed from the components of a finite mixture distribution. The method comprises two steps: first, a finite mixture distribution with the following probability density function

$$f(\,\cdot\,;\pi_1,\ldots,\pi_k,\theta_1\ldots\theta_k) = \pi_1 f(\,\cdot\,;\theta_1) + \cdots + \pi_k f(\,\cdot\,;\theta_k)$$

is fitted to a data set; second, each observation x is assigned to the component j, $1 \leq j \leq k$, with $\hat{\pi}_j f(x;\hat{\theta}_j)$ maximum. After the fitting process, some mixture components might not be separated enough. To deal with this situation, several authors have proposed merging methods that, based on the posterior probabilities incrementally, combine those components that are more similar. Using a generic definition, an integrated formulation to unify such merging methods is presented and discussed. This new formulation opens the way to define new methods based on the posterior probabilities.

Keywords

Mixtures models; Merging components; hierarchical clustering

Marc Comas-Cufí
Universitat de Girona, Spain, e-mail: mcomas@imae.udg.edu

Josep Antoni Martín Fernández
Universitat de Girona, Spain, e-mail: jamf@imae.udg.edu

Glòria Mateu Figueras
Universitat de Girona, Spain, e-mail: gloria@imae.udg.edu

References

BAUDRY, J.P., RAFTERY A.E., CELEUX, G., LO, K., and GOTTARDO, R. (2010). Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, 9(2):332–353.

HENNIG, C. (2010): Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1):3–34.

LONGFORD, N.T. and BARTOSOVA, J. (2014). A confusion index for measuring separation and clustering. *Statistical Modelling*, 14(3):229–255.

DESPOTA: a permutation test algoritm to detect a partition from a dendrogram

Dario Bruzzese and Domenico Vistocco

Abstract The topic of selecting the "best" partition starting from a hierarchical cluster has been extensively treated in the literature. Main proposals exploit cluster validity indexes; at this regard see the seminal works by Milligan (1981) and Milligan and Cooper (1985). A different strategy could exploit the more general GAP method, proposed by Tibshirani et al. (2001), to estimate the number of clusters starting from the output of any clustering algorithm, and then usable also in case of hierarchical clustering. The application of the GAP method to the hierarchical clustering case is effective with the only flaw to require as input the different partitions among which the optimal one has to be selected. We propose a method, DESPOTA (**DE**ndrogram **S**licing through a **P**ermutati**O**n **T**est **A**pproach) that exploits permutation tests (Pesarin and Salmaso, 2010) in order to automatically detect a partition among these embedded in a dendrogram (Bruzzese and Vistocco, 2015). It locally explores the branches of the dendrogram picking up those clusters that will compose the final partition. As main feature, DESPOTA includes in the search space also partitions not corresponding to horizontal cuts of the tree. Applications on both real and synthetic datasets will show the effectiveness of our proposal.

Keywords

Hierarchical clustering; Cluster detection; Permutation tests.

Dario Bruzzese
University of Naples "Federico II", Department of Public Health, Italy, e-mail: dbruzzes@unina.it

Domenico Vistocco
University of Cassino, Department of Economics and Law, Italy, e-mail: vistocco@unicas.it

References

BRUZZESE, D., and VISTOCCO, D. (2015): DESPOTA: DEndrogram Slicing through a PemutatiOn Test Approach. *Journal of Classification*, in press.

MILLIGAN, G.W. (1981): A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika*, 46(2), 187–199.

MILLIGAN, G.W., and COOPER, M.C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Dataset, *Psychometrika*, 52(2), 159–179.

PESARIN, F., and SALMASO, L. (2010): *Permutation Tests for Complex Data*. Theory, Applications and Software, Chichester: John Wiley and Sons.

TIBSHIRANI, R., WALTHER, G., and HASTIE, T. (2001): Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of Royal Statistical Society B*, 83(2), 411–423

On a Comprehensive Metadata Framework for Artificial Cluster Data Generation

Rainer Dangl and Friedrich Leisch

Abstract Solid results in unsupervised model validation require thoroughly tested validation methods and algorithms. We intend to optimize their development by proposing a framework that streamlines the way artificial test data is constructed. This improves comparability between existing and new algorithms and offers a more transparent way of assessing performance. In essence, the framework, developed with R, consists of metadata object definitions for various kinds of data types (e.g. metric, functional or ordinal data). These objects impose a certain structure on the metadata information that enables R to assemble the actual data sets in a way that the user can generate all desired data sets (custom functions for random number generation, location of group means, etc.) while at the same time providing a reliable structure for the metadata that is the same for all data sets. The necessary functions for (meta)data generation have been implemented in R package `bdlp`, which is under development and thus at the moment hosted on R-forge.

Keywords

Clustering; Benchmarking; Artificial Data

Rainer Dangl
Institute for Applied Statistics and Computing, University of Natural Resources and Life Sciences Vienna, Austria, e-mail: `rainer.dangl@boku.ac.a`

Friedrich Leisch
Institute for Applied Statistics and Computing, University of Natural Resources and Life Sciences Vienna, Austria, e-mail: `friedrich.leisch@boku.ac.at`

References

HOTHORN, T., LEISCH, F., ZEILEIS, A., & HORNIK, K. (2005): The Design and Analysis of Benchmark Experiments. *Journal of Computational and Graphical Statistics*, 14 (3), 675–699.

Benchmarking cluster algorithms for ordinal survey data

Dominik Ernst and Friedrich Leisch

Abstract In marketing research survey data of consumer preferences are often clustered in order to identify market segments. In many cases these data sets are on ordinal scale, e.g. ratings on scales from "very important" to "not important at all". Most cluster algorithms on the other hand are designed for metric data, from the standard k-means algorithm to mixtures of Gaussians. We present model-based cluster algorithms for ordinal data and compare their performance on both artificial and real world data sets. Special emphasis is given to resampling methods for cluster validation like subsampling or bootstrapping. Resampling helps to assess the stability of single clusters and complete partitions as well as to identify the best number of clusters. As these simulation-based approaches are parallel in nature they can be efficiently used on standard laptops and desktops as well as HPC systems for everyday cluster analysis applications. All methods shown have been implemented in R and are freely available.

Keywords

cluster analysis; ordinal data; R

Dominik Ernst
BOKU - University of Natural Resources and Life Sciences, Vienna, Austria, e-mail: dominik.ernst@boku.ac.at

Friedrich Leisch
BOKU - University of Natural Resources and Life Sciences, Vienna, Austria, e-mail: friedrich.leisch@boku.ac.at

Modern multivariate data analysis through monitoring

Marco Riani

Abstract Robust methods are little applied although much studied by statisticians. In this paper we sketch what we see as some of the reasons for this failure and suggest a system of interrogating robust analyses, which we call ?monitoring?, whereby we consider fits from very robust to highly efficient and follow what happens to aspects of the fitted model. The resulting procedure provides insight into the structure of the data including outliers and the presence of more than one population. Monitoring overcomes the hindrances to the routine adoption of robust methods, being informative both about the choice between the various robust procedures and the choice of different tuning constants. We also propose some computational improvements of the robust routines and provide a recursive implementation of the so called concentration steps. The output is a set of efficient routines for fast updating of the model parameter estimates, which do not require either data sorting or inverse matrix computations. These efficient routines make the use of robust methods appealing also in presence of big datasets. Finally, we describe the new routines inside the FSDA (Flexible Statistics Data Analysis) toolbox for MATLAB, which go from the possibility of simulating regression or multivariate mixtures with a prespecified degree of overlap among groups (possibly contaminated with outliers), to the implementation of robust clustering routines based on trimming and eigenvalue constraints, from the possibility of brushing and linking different objects which come out from the application of robust methods, to the implementation of new routines for robust heteroskedastic regression. Through all these developments we can build an integrated approach for modern multivariate data analysis which puts together robustness, efficiency and clustering.

Marco Riani
Department of Economics, Division of Statistics and Computing University of Parma, Italy, e-mail: mriani@unipr.it

Generalized Additive Models (GAMs) via Bayesian P-splines using INLA

Cajo ter Braak, María Xosé Rodríguez-Álvarez, Martin Boer, Paul Eilers, and Havard Rue

Abstract IFCS 2015 celebrates Generalized Additive Models (GAMs) with its back-fitting algorithm. With P-splines a direct fitting algorithm is feasible. With a prior on the smoothing parameter, Bayesian GAMs are obtained, which allow credible intervals that account for the uncertainty in the smoothing parameter.Integrated nested Laplace approximation (INLA) makes the Bayesian approach to GAMs practical and the INLA R package (Martins et al., 2013) already includes smoothing functions using piece-wise linear basis functions with peaks at the data points, supported by nice theory and links to geo-statistical modelling.Fitting Bayesian P-splines using INLA combines the advantages of the Bayesian approach using INLA with the flexibility and power of P splines. Our approach maintains sparseness whereas the usual mixed model approach (Wakefield, 2013) does not. We are able to go beyond GAMs without increasing the number of penalty parameters beyond the number of covariates. We wrote a small R library (BayesianPspline) that wraps up the approach. We illustrate our approach using an ecological example (Fraaije et al., 2015).

Cajo Ter Braakr
Wageningen University and Research Center, Netherlands, e-mail: cajo.terbraak@wur.nl

María Xosé Rodríguez-Álvarez
University of Vigo, Spain, e-mail: mxrodriguez@uvigo.es

Martin Boer
Wageningen University and Research Center, Netherlands, e-mail: Martin.Boer@wur.nl

Paul Eilers
Erasmus Medical Center, Netherlands, e-mail: p.eilers@erasmusmc.nl

Havard Rue
Norwegian University of Science and Technology, Norway e-mail: hrue@math.ntnu.no

Keywords

generalized additive models; spline smoothing; Bayesian model

References

FRAAIJE, R. G. A., TER BRAAK, C. J. F., VERDUYN, B., BREEMAN, L. B. S., VERHOEVEN, J. T. A. & SOONS, M. B. (2015). Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. *Functional Ecology*, http://dx.doi.org/10.1111/1365-2435.12441.

MARTINS, T. G., SIMPSON, D., LINDGREN, F. & RUE, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67, 68–83.

WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods*, edn. New York, NY, Springer New York.

Multinomial correspondence analysis

Patrick J.F. Groenen and Julie Josse

Abstract Multiple correspondence analysis (MCA) is a well know visualization technique for studying the relations between the categories of two or more nominal variables and the similarites between individuals. Often, the results are represented in a biplot showing individuals and the categories simultaneously. In this paper, we present multinomial correspondence analysis, a maximum likelihood version of MCA based on the multinomial logit model. Multinomial correspondence analysis fits main effects for the categories along with the coordinates needed for the biplot representation for individuals and categories as is usual in MCA. The likelihood of such a model can not be straightforwardly maximized and we tackle this issue by suggesting a majorization algorithm. The trick is to bound the Hessian with a very simple matrix. It is possible to include in the procedure a ridge and/or nuclear norm penalty to avoid problems due to the large dimensionality of the parameters set. Another advantage is that missing values can be easily handle and we can provide probabilities for each category of the missing variable for an individual.

Keywords

correspondence analysis; majorization; ridge penalty; nuclear norm penalty

Patrick J.F. Groenen
Econometric Institute, Erasmus University Rotterdam, Netherlands, e-mail: groenen@ese.eur.nl

Julie Josse
Department of Statistics, Applied Mathematics Unit, Agrocampus Ouest, Rennes, France, e-mail: josse@agrocampus-ouest.fr

Three-way data analysis with clustered bilinear models

Pieter Schoonees

Abstract A least-squares bilinear clustering framework for modelling three-way data, where each observation consists of an ordinary two-way matrix, is introduced. The method combines bilinear decompositions of the two-way matrices into overall means, row margins, column margins and row-column interactions with clustering along the third way. Different clusterings are defined for each part of the decomposition, so that up to four different classifications are defined jointly. The computational burden is greatly reduced by the orthogonality of the bilinear model, such that the joint clustering problem reduces to separate ones which can be handled independently. Three of these sub-problems are specific cases of k-means clustering.

Keywords

three-way data; bilinear decomposition; cluster analysis

Pieter Schoonees
Rotterdam School of Management, Erasmus University, Netherlands, e-mail: schoonees@rsm.nl

Regularized Generalized Canonical Correlation analysis for Multiway data.

Arthur Tenenhaus and Laurent Le Brusquet

Abstract Several examples of either three-way data or multiblock data can be found in a variety of domains including chemometrics, psychometry, bioinformatics to name but a few. Nowadays, it frequently occurs to encounter data combining multiway and multiblock data. Regularized Generalized Canonical Correlation Analysis (RGCCA) [Tenenhaus & Tenenhaus 2011] is currently geared for the analysis two-way data matrices. In this work, RGCCA is extended to the multiway data configuration (Multiway RGCCA - MGCCA) by adding appropriate kronecker constraints to the RGCCA outer weight vector. The main aim of MGCCA is to study the relationships between a collection of multi-way data table.

Keywords

Multiblock data; Multiway data

References

TENENHAUS, A. and TENENHAUS, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2), 257–284.

Arthur Tenenhaus
L2S-CentraleSupelec, France, e-mail: arthur.tenenhaus@centralesupelec.fr

Laurent Le Brusquet
L2S-CentraleSupelec, France, e-mail: laurent.lebrusquet@centralesupelec.fr

Biplot-based visualizations in latent class modelling

Zsuzsa Bakk, Niel Le Roux, and Jeroen Vermunt

Abstract The use of categorical analysis of distance (AoDcat) biplot methodology to visualize the posterior classifications resulting from a latent class (LC) model is proposed. While most visualization tools for LC models can only show the class response probabilities on all indicators, the AoDcat graphs allow a more in depth visual description of the data: they provide an easy to quantify and visualize measure (alpha bags) of class separation; they aid interpretation by visualizing simultaneously the cases, the group centroids and the category level points associated with the response variables; they show the dispersion of cases belonging to the different classes around the class centroids thus providing visual insight of the relative homogeneity of the respective classes. Ternary plots can also be used for the three class model but the proposed biplot-based visualizations are not restricted to be constructed only for three class models. The proposed AoDcat biplots for LC models are illustrated for a three class as well as a five class LC model. Finally, measures of the goodness of the approximations of the biplot-based visualizations of multidimensional LC models are briefly discussed.

Keywords

biplots; categorical analysis of distance; latent class models; visualization

Zsuzsa Bakk
Tilburg University, Netherlands, e-mail: Z.bakk@tilburguniversity.edu

Niel Le Roux
Stellenbosch University, South Africa, e-mail: njlr@sun.ac.za

Jeroen Vermunt
Tilburg University, Netherlands, e-mail: j.k.vermunt@tilburguniversity.edu

Small sample multi-label discriminant analysis

Nelmarie Louw

Abstract Multi-label learning has many applications, such as text classification, speech categorization and gene function prediction. In a multi-label data set each entity may be associated with a subset of the available variables, as opposed to a single-label data set where each entity is associated with only one label. Many multi-label classification methods have been proposed in the literature and it is an active field of research. In one approach, named multi-label linear discriminant analysis (MLDA), classical linear discriminant analysis is generalized by extending the definition of the between-group and within-group covariance matrices to the multi-label scenario. Label correlations are also taken into account. In many multi-label data sets the number of feature variables exceeds the number of observations, resulting in singularity of the covariance matrices. Small sample multi-label discriminant analysis considers various ways of dealing with this, other than using the pseudo-inverse proposed in the MLDA approach. A further extension to multi-label kernel discriminant analysis for small sample situations is also considered. The performance of the proposed methods is evaluated on different multi-label data sets.

Keywords

Multi-label classification; Small sample size;

References

TSOUMAKAS, G., KATAKIS, I. and VLAHAVAS, I. (2010): *Mining multi-label data*. In Datamining and Knowledge Discovery Handbook, Springer, pp. 667-

Nelmarie Louw
Stellenbosch University, South Africa, e-mail: nlouw@sun.ac.za

685. WANG, H., DING, C. and HUANG, H. (2010): *Multi-label linear discriminant analysis*. In Computer Vision ECCV 2010, Springer, pp.126-139.

Feature selection and kernel specification for support vector machines using multi-objective genetic algorithms

Martin Philip Kidd, Martin Kidd, and Surette Bierman

Abstract Support Vector Machines (SVMs) have shown to be popular for classification problems. There are tuning parameters that need to be specified before fitting SVMs. Genetic algorithms (GA) have been used as optimization algorithm for selecting parameters, but most applications excluded the selection of a kernel function. GA has a further extension called multi-objective GA where multiple criteria are specified and the fitness of possible solutions are determined by their level of dominance. The use of multi-objective GA applied to SVMs is demonstrated where the optimization criteria are prediction error, number of variables and number of support vectors. The kernel function, kernel parameters and cost parameter(C) form part of the member definition of the GA. Benchmark and simulated data sets are used to show how this approach provides a range of solutions that are trade-offs of the various optimization criteria. For the standard GA where prediction error is used as fitness criterion, the fitness has to be determined from a validation set or using cross validation to guard against overfitting. In the multi-objective approach, the number of variables and number of support vectors are part of the optimization, and the possibility of using the full training dataset without cross validation will be discussed.

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

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

Surette Bierman
Dept of Statistics, Stellenbosch University, Stellenbosch, South Africa, e-mail: surette@sun.ac.za

Keywords

Genetic Algorithms; Multi-objective, SVM

References

BIERMAN, S. and STEEL, S.J. (2009): Variable selection for support vector machines. *Communications in Statistics Simulation and Computation*, 38(8), 1640–1658

DEB, K., PRATAB, A., AGARWAL, S. and MEYARIVAN, T. (2002): A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197

Measures of fit for nonlinear biplots

Karen Vines

Abstract Biplots provide a representation of observations and variables on the same diagram. In order to ensure that this diagram is low-dimensional (typically two-dimensional) the representation is usually an approximation - the distances between points approximate the dissimilarities between corresponding observations. Thus for any given biplot it is important to know how good this approximation is. As with linear (PCA) biplots, an overall measure of quality of a nonlinear biplot is given by considering $\text{trace}(\mathbf{Y}_r'\mathbf{Y}_r)\text{trace}(\mathbf{Y}'\mathbf{Y})$ where $\mathbf{Y}$ is an exact representation of the n observations in $n-1$ dimensional space and $\mathbf{Y}_r$ is the representation in r-dimensional space. However if the overall quality is not good, it is of interest of why that might be. For example, does the form of the dissimilarity function mean that differences due an important variable is not represented well by Euclidean distances in small number of dimensions, or are simply too many variables are important to be summarised well on any r dimensional plot? So, taking measures of fit for linear biplots as a starting point, ways of exploring the fit of nonlinear biplots will be explored.

Keywords

Nonlinear biplots; measures of fit

References

GOWER, J.C. and HARDING, S.A. (1988). Nonlinear Biplots, *Biometrika*, 75, 445–455.

Karen Vines
The Open University, United Kingdom, e-mail: karen.vines@open.ac.uk

GOWER, J.C. and HAND, D.J. (1996). *Biplots*. Chapman & Hall, London.
GOWER, J.C., LUBBE, S. and LE ROUX, N.J.R (2011). *Understanding Biplots*. Wiley, Chichester.

Mixture simultaneous factor analysis for modeling structural differences in multivariate multilevel data

Kim De Roover, Jeroen K. Vermunt, Marieke E. Timmerman, and Eva Ceulemans

Abstract Multivariate multilevel data consist of multiple data blocks that all involve the same set of variables. For instance, one may think of personality measures of inhabitants from different countries. The associated research questions often pertain to the underlying covariance structure (e.g., which dimensions underlie the individual scores), and whether this structure holds for each data block (e.g., do inhabitants of different countries vary on the same personality dimensions). To answer such research questions, we present mixture simultaneous factor analysis (MSFA) which performs a mixture model clustering of the data blocks according to their factor structure. In other words, MSFA assumes that the data blocks are sampled from a mixture of multivariate normal distributions with different covariance matrices, which can be perfectly modeled by a low rank common factor model. Note that existing multilevel mixture models were mainly capturing differences in means, often assuming the covariances to be identical, whereas MSFA is purely focused on differences in the covariance structure. MSFA can be applied by means of Latent GOLD.

Keywords

mixture model clustering; factor analysis; multilevel mixture model

Kim De Roover
KU Leuven, Belgium, e-mail: kim.deroover@ppw.kuleuven.be

Jeroen K. Vermunt
Tilburg University, Netherlands, e-mail: J.K.Vermunt@uvt.nl

Marieke E. Timmerman
University of Groningen, Netherlands, e-mail: m.e.timmerman@rug.nl

Eva Ceulemans
KU Leuven, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

A multiple clusterings model based on Gaussian mixture

Andrea Pastore and Stefano F. Tonellato

Abstract We consider, under a Gaussian model based perspective, the problem of identifying different clusterings of a given set of units, where each clustering is described by a subset of the observed variables. In particular, we assume that it is possible to identify, among the set of observed variables, T subsets, each of them giving rise to a different clusterings, and a complementary subset of (unrelevant) variables that provides no information about clustering. We assume that the distributions of the T clustering subsets are independent Gaussian mixtures, while the conditional distribution of the irrelevant variable, given those included in the T subsets, is Gaussian. For the identification of the T subsets of variables and for estimation of the model parameters, we propose a generalization of an algorithm by Raftery and Dean, which is tailored for the selection of variables in Gaussian mixture models, and where the model comparison problem is addressed using approximate Bayes factors. The proposed algorithm provide a forward-stepwise identification of the T clustering subsets of variables. Some results from Monte Carlo experiments and from application to real dataset are presented.

Keywords

Cluster analysis; Gussian mixture model; BIC

Andrea Pastore
Department of Economics - Ca' Foscari University of Venice, Italy, e-mail: pastore@unive.it

Stefano F. Tonellato
Department of Economics - Ca' Foscari University of Venice, Italy, e-mail: stone@unive.it

References

GALIMBERTI, G. and SOFFRITTI, G. (2007): Model-based methods to identify multiple cluster structures in a data set. *Computational Statistics & Data Analysis*, 52, 520–536

RAFTERY, A.E. and DEAN, N. (2006): Variable Selection for Model-Based Clustering. *Journal of the American Statistical Assocation*, 101, 168–178.

Criteria for model selection in model-based clustering

Fumitake Sakaori

Abstract Model-based clustering, or finite mixture models, are important classification methods, and wide range of theoretical and applied researches have been made. One of the most useful advantages of these methods are the applicability of model selection criteria for choosing the number of clusters, selecting variables, selecting structure of model parameters and so on. One might use some basic information criteria, e.g., AIC or BIC. However, regularity conditions for deriving these criteria do not hold in finite mixture models - these models are called "singular models" - and these criteria do not have theoretical justification at all. In this study, we compare the performance of WAIC and WBIC, information criteria applicable for singular models, with the basic criteria numerically in some problem of the model-based clustering.

Keywords

model-based clustering; AIC; WAIC

References

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000): Assessing a mixture model for clustering withe the integrated completed likelihood. *IEEE transaction on pattern analysis and machine intelligence*, 22, 7, 719–725.

Fumitake Sakaori
Faculty of Science and Engineering, Chuo University, Japan, e-mail: sakaori@math.chuo-u.ac.jp

WATANABE, S. (2010): Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. —em Journal of machine learning research, 11, 3571–3594.

Partially Supervised Biclustering of Gene Expression Data with Applications in Nutrigenomics Biomarker Discovery

Monica Hiu Tung Wong

Abstract A family of parsimonious Gaussian mixture models for the biclustering of high-dimensional gene expression data is introduced. Previously, a family of parsimonious Gaussian mixture models was introduced which is based on the mixtures of factor analyzers model. Our family extends these models into the biclustering framework by including a binary and row-stochastic factor loadings matrix. This particular form of factor loadings matrix results in a block-diagonal covariance matrix, which is a useful property in specific gene expression applications in nutrigenomics biomarker discovery. Knowledge of the factor loadings matrix is useful in this application and is reflected in the partially supervised nature of the algorithm. Parameter estimates are obtained through a variant of the expectation-maximization algorithm and the best fitting model is selected using the Bayesian information criterion. We demonstrate our family of models using both simulated and real data.

Monica Hiu Tung Wong
McMaster University, Canada, e-mail: wongm44@mcmaster.ca

Bayes Clustering Operators for Random Labeled Point Processes

Lori Dalton

Abstract Although clustering algorithms aim to group objects based on some similarity criterion with the hope of gaining knowledge about the underlying groups in a problem, they do not optimize error with respect to a probabilistic model to make clustering rigorously predictive. To address this, recent work develops a general risk-based formulation for clustering that parallels classical Bayes decision theory for classification, transforming clustering from a subjective activity to an objective operation. Given an underlying random labeled point process, we formally define notions of risk and error for any clustering operator. We then develop a general analytic procedure to find an optimal clustering operator, called a Bayes clusterer, which corresponds to the Bayes classifier in classification theory. We present solutions for optimal operators and their errors for several classes of Gaussian models, which provide the first fundamental limits of performance in clustering under known models. Owing to computational complexity, we further develop methods to reduce complexity without sacrificing optimality, and develop approximations of the Bayes clusterer to significantly reduce computational complexity under very large point sets.

Keywords

clustering; risk; optimization

Lori Dalton
The Ohio State University, USA,e-mail: dalton@ece.osu.edu

References

DALTON L.A., BENALCAZAR M. E., BRUN M., DOUGHERTY E. R. (2015). Analytic Representation of Bayes Labeling and Bayes Clustering Operators for Random Labeled Point Processes. Accepted in *IEEE Transactions in Signal Processing*.

DALTON L.A. (Nov. 2013) On the optimality of K-means clustering. In *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics*.

Bayesian clustering: a novel nonparametric framework for borrowing strength across populations

Antonio Lijoi, Bernardo Nipoti, and Igor Pruenster

Abstract Non-exchangeable data arise in a number of relevant applied problems and have led to the development of several statistical methodologies tailored to the type of dependence assumed among the data. Here attention is focused on the case where data originate from different studies or refer to related experiments that are performed under different conditions. In such a context, the properties of a novel flexible class of nonparametric priors are analyzed. Specifically, these priors are used to define dependent hierarchical mixture models whose features are explored, especially in terms of the clustering behavior and the borrowing of strength across studies. An extensive simulation study investigates the effect of dependence: the novel model allows for a more appropriate use of the available information and, in turn, leads to a better understanding of the clustering structure underlying the data. The degree of dependence between priors does not need to be set by the experimenter as the Bayesian approach naturally allows the data to determine it.

Keywords

Bayesian clustering; Dependent priors; Hierarchical mixture models

Antonio Lijoi
University of Pavia & Collegio Carlo Alberto, Italy, e-mail: lijoi@unipv.it

Bernando Nipoti
Name, University of Torino & Collegio Carlo Alberto, Italy, e-mail: bernardo.nipoti@carloalberto.org

Igor Pruenster
Name, University of Torino & Collegio Carlo Alberto, Italy, e-mail: igor@carloalberto.org

References

LIJOI, A., NIPOTI, B. and PRUENSTER, I. (2014): Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20, 1260–1291.

LIJOI, A., NIPOTI, B. and PRUENSTER, I. (2014): Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis*, 71, 417–433.

Dirichlet process Bayesian clustering: an application to survival data

Silvia Liverani

Abstract PReMiuM is a recently developed R package for Bayesian clustering using a Dirichlet process mixture models. It is an alternative to regression models, non-parametrically linking a response vector to covariate data through cluster membership (Liverani et al, 2015). Posterior inference is carried out by using Markov chain Monte Carlo simulation and to allow for fast computations, all essential methods in the package are based on efficient C++ code. The model allows binary, categorical, count, survival and continuous response, as well as continuous and discrete covariates. Additionally, predictions may be made for the response, and missing values for the covariates are handled. Several samplers and label switching moves are implemented along with diagnostic tools to assess convergence. A number of R functions for post-processing of the output are also provided. In addition to fitting mixtures, it may additionally be of interest to determine which covariates actively drive the mixture components. This is implemented in the package as variable selection.This talk will include an overview of the features of the package and some of its applications to date (Pirani et al, 2015; Hastie et al, 2013). The latter part of the talk will focus on the clustering of survival data with censoring. We will present a simulation study and a real application to sleep data.

Keywords

Bayesian; variable selection; censored data

Silvia Liverani
Department of Mathematics, Brunel University London, United Kingdom, e-mail: silvia.liverani@brunel.ac.uk

References

LIVERANI, S., HASTIE, D. I., AZIZI, L., PAPATHOMAS, M. and RICHARDSON, S. (2015): PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. *Journal of Statistical Software*, 64(7), 1–30.

PIRANI, M., BEST, N., BLANGIARDO, M., LIVERANI, S., ATKINSON, R. W., and FULLER, G. W. (2015): Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. To appear in *Environment International*.

HASTIE, D. I., LIVERANI, S., AZIZI, L., RICHARDSON, S. and STUCKER I. (2013): A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC Medical Research Methodology*, 13 (1), 129.

Optimizing Mailing Decisions Based on a Mixture of Dirichlet Processes

Harald Hruschka and Nadine Schröder

Abstract Determining the optimal amount of mailings being sent to individual customers is a crucial decision in direct marketing. In this respect it is important to specify relevant mailing variables. By distinguishing different types of mailings and considering their sizes, we set our study apart from the majority of existing studies. To deal with latent heterogeneity we estimate a mixture of Dirichlet processes (MDP) whose components are Tobit-2 models. A policy function approach is used to take endogeneity into account. We investigate whether and how consideration of endogeneity leads to different managerial implications. To this end, we determine mailings by dynamic optimization for customers which are prototypical for the segments discovered by the MDP model. We find evidence that mailings which are too large are sent to about 25 % of the customers in our data base if decisions are based on a model which ignores endogeneity.

Keywords

Marketing; Latent Heterogeneity; Endogeneity

Harald Hruschka
University of Regensburg, Germany, e-mail: harald.hruschka@wiwi.uni-r.de

Nadine Schröder
University of Regensburg, Germany, e-mail: nadine.schroeder@wiwi.uni-r.de

Latent class modeling of markers of day-specific fertility

Francesca Bassi and Bruno Scarpa

Abstract There is a considerable interest in predicting the fertile days in a woman's menstrual cycles in couples desiring a pregnancy and among those wishing to avoid conception by periodic abstinence. Cervical mucus detection is potentially an accurate marker of fertile days. It is therefore of great interest to assess the magnitude of heterogeneity among women and among cycles and among cycles of a given woman, in the evolution in time of the mucus secretions detected during an interval of potential fertility and defined relative to ovulation. In this paper, we study the problem of heterogeneity in cervical mucus hydration at various times relative to the mucus peak, both among cycles and among women, specifying and estimating appropriate multilevel latent class models for longitudinal data. Results showed that heterogeneity in mucus evolution between cycles and women is non-negligible. Model estimates identified different mucus patterns for groups of cycles and women, and the characteristics of the cycles and the women which influence mucus symptom evolution over time.

Keywords

Menstrual cycles; cervical mucus; peak day; multilevel latent class models; multilevel latent growth mixture models

Francesca Bassi
Department of Statistics, University of Padova, Italy, e-mail: francesca.bassi@unipd.it

Bruno Scarpa
Department of Statistics, University of Padova, Italy, e-mail: bruno.scarpa@unipd.it

Attitudes to maternity in Poland - a longitudinal analysis based on latent Markov models with covariates

Ewa Genge and Joanna Trzesiok

Abstract The total fertility rate in Poland is 1.3 and globally only 6% of countries rank lower. Such a low level of fertility has a profound impact on both the state and the society. The difference between the actual fertility rate (1.3) and the desirable level (2.1), attributable to the issues related to well-being and self-fulfillment, can be interpreted as an unmet need of women who give up or postpone maternity for social, economic or personal reasons.The latent Markov (LM) model is a variation of the latent class model that is applied to estimate not only the prevalence of latent class membership, but the incidence of transitions over time. We used the model-based clustering approach for grouping and detecting inhomogeneities of public attitudes to maternity in Poland. We focused especially on the latent Markov models with covariates (having influence on initial and transition probabilities), which additionally allowed us to investigate the dynamic pattern of the Polish attitudes to maternity for different demographic features.We analyzed data collected as part of the Polish Social Diagnosis using the R software.

Keywords

latent Markov model; longitudinal data; model-based clustering

Ewa Genge
University of Economics in Katowcie, Poland, e-mail: ewa.genge@ue.katowice.pl

Joanna Trzesiok
University of Economics in Katowcie, Poland, e-mail: joanna.trzesiok@ue.katowice.pl

References

BARTOLUCCI F., MONTANARI G., PANDOLFI S. (2001): Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis*, 83, 287–301.

GENGE E. (2014): A latent class analysis of the public attitude towards the euro adoption in Poland. *Advances in Data Analysis and Classification*, 8(4), 427–442.

Space-time clustering for radiation monitoring post data based on hierarchical structure

Fumio Ishioka and Koji Kurihara

Abstract The Tokyo Electric Power Fukushima I Nuclear Power Plant accident released large amounts of radioactive materials to the environment. Measurement of radiation dose rate by monitoring post is carrying out by related ministries and agencies, local governments, nuclear operator and related companies in real time. The government aggregates them and opens to the public. The monitoring data have a fixed lat/long location as spatial information, and there is considerable validity in applying various spatial clustering approaches. The purpose of this study is to ascertain where and when the high contaminant cluster lies at the "difficult-to-return zone" in Fukushima prefecture from these monitoring results. The spatial scan statistic (Kulldorff et al., 2009) is a method of detecting clusters based on the likelihood ratio associated with the quantity inside and outside a circular scanning window. However, it is noted that a non-circular shaped cluster, such as the shape made by a river or a road cannot be detected. To solve this problem, an echelon approach was proposed. (Myers et al., 1997; Ishioka et al., 2012). The echelons enable the spatial clustering consisting of the various shapes which have high-likelihood, because the areas are scanned based on the inherent hierarchical structure of data. We apply the echelons to the monitoring post data to know a trend of the movements of radioactive materials released in the environment.

Keywords

spatial scan statistic; echelon analysis; radiation dose rate

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

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

References

ISHIOKA, F. and KURIHARA, K. (2012) :Detection of spatial clusters using echelon scanning method. Proceedings of COMPSTAT2012 (20th International Conference on Computational Statistics), 341–352.

KULLDORFF, M, HUANG, L. and KONTRY, K. (2009): A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, 8, 58.

MYERS, W.L., PATIL, G.P. and JOLY, K. (1997): Echelon approach to areas of concern in synoptic regional monitoring, *Environmental and Ecological Statistics*, 4, 131–152.

Regional Spatial Moving Average and new Spatial Correlation Coefficient

Andrzej Sokolowski, Malgorzata Markowska, Marek Sobolewski, Danuta Strahl, and Sabina Denkowska

Abstract Spatial moving average is used in geography mainly for interpolation of measurements performed in point special locations to find values for other points. Regional Spatial Moving Average which is introduced in the paper deals with measurements referring to areas (regions). The general aim is to smooth statistical data and eliminate random disturbances especially for data obtained from small geographical areas. The basic weighting system is based on border lengths but additional weights can be introduced using other side variables. The problem of relative importance of boarders (between regions and between states) is discussed in the paper. The smoothing process is controlled by the smoothing constant which, taking values from [0,1] interval decides about the proportion of current region in the moving average value. Taking smoothing constant as 0 we can define a new spatial correlation coefficient. It is being compared with Moran's I coefficient. Using the idea of partial correlation we can measure the correlation of two variables taking away the effect of spatial neighbourhood. Some simulation studies results are presented together with real data application for NUTS 2 European Union regions.

Andrzej Sokolowski
Department of Statistics, Cracow University of Economics, Poland, e-mail: sokolows@uek.krakow.pl

Malgorzata Markowska
Chair of Regional Economics, Wroclaw University of Economics, Poland, e-mail: mmarkowska@ae.jgora.pl

Marek Sobolewski
Department of Quantitative Methods, Rzeszow University of Technology, Poland, e-mail: mareksobol@poczta.onet.pl

Danuta Strahl
Chair of Regional Economics, Wroclaw University of Economics, Poland, e-mail: dstrahl@ae.jgora.pl

Sabina Denkowska
Department of Statistics, Cracow University of Economics, Poland, e-mail: sabina.denkowska@uek.krakow.pl

Keywords

spatial correlation; Moran's I; regional analysis

Removing Unwanted Variation for classification and clustering

Laurent Jacob, Johann Gagnon-Bartsch, and Terry Speed

Abstract Large omics studies are often carried out over months or years, and involve multiple labs. Unwanted variation (*UV*) can arise from technical elements such as batches, different platforms or laboratories, or from biological signals such as heterogeneity in age or ethnicity which are unrelated to the factor of interest in the study. Similar issues arise when the goal is to combine several smaller studies. A very important task is to remove these *UV* factors without losing the factors of interest. When neither the factors of interest nor the *UV* are observed, the problem is quite difficult. Recently, we proposed a general framework (called *RUV*) for removing *UV* in microarray data using *control* genes. It showed very good behavior for differential expression analysis (i.e., with a known factor of interest) when applied to several datasets. Our objective in this talk is to describe our recent results doing similar things when carrying out classification and clustering.

Laurent Jacob
Department of Statistics, UC Berkeley, United States, e-mail: laurent@stat.berkeley.edu

Johann Gagnon-Bartsch
Department of Statistics, UC Berkeley, United States, e-mail: johann@stat.berkeley.edu

Terry Speed
Bioinformatics Division, WEHI, Australia, e-mail: terry@wehi.EDU.AU

Unimodal Logistic Discrimination

Joaquim Costa and A. Rita Gaio

Abstract Many supervised classification problems involve classifying instances into classes which have a natural ordering. Our method assumes that in a supervised classification problem with ordered classes, the random variable class $\mathscr{C}_x$ associated with a given query point x should follow a unimodal probability distribution. We apply this paradigm to the usual multinomial logistic discrimination model and derive the maximum likelihood parameters.

Keywords

ordinal; logistic regression; maximum likelihood

References

PINTO DA COSTA, J., CARDOSO, J.S. (2005), Classification of ordinal data using neural networks, *Lecture Notes in Artificial Intelligence*, LNAI 3720:690–697.

PINTO DA COSTA, J., ALONSO, H., CARDOSO, J.S. (2008), The unimodal model for the classification of ordinal data, *Neural Networks*, 21(1):78–91.

PINTO DA COSTA, J., SOUSA, R., CARDOSO, J.S. (2010), An all-at-once Unimodal SVM Approach for Ordinal Classification. In *Proceedings of the Ninth International Conference on Machine Learning and Applications*. (ICMLA 2010), Washington DC, USA.

Joaquim Costa
Univ. Porto, Portugal, e-mail: jpcosta@fc.up.pt

A. Rita Gaio
Univ. Porto, Portugal, e-mail: argaio@fc.up.pt

High-dimensional regression mixture models to perform clustering - application to electricity dataset

Emilie Devijver, Jean-Michel Poggi, and Yannig Goude

Abstract Model-based clustering is useful to understand how data is grouped. We propose to introduce the Lasso-MLE procedure, which uses finite mixture regression models to perform model-based clustering, on electricity dataset. To improve electricity consumption prediction, we cluster together consumers who have the same reliance between two successive days. To deal with functional datasets (response and regressors) we use the wavelet coefficients rather than the discretization. Then, we construct a model collection with more or less components and more or less relevant variables, and select some with the slope heuristic. Relevant variables are detected with the Lasso estimator, whereas we refit estimators with the maximum likelihood estimator. We run this procedure on Thursday 5 and Wednesday 6 January 2010. We analyze clusters done by the procedure with usual electricity consumption criterion, as the temperature, and the tariffs among others.

Keywords

Model-based clustering; functional dataset; data analysis

Emilie Devijver
Université Paris-Sud - INRIA Select, France, e-mail: emilie.devijver@math.u-psud.fr

Jean-Michel Poggi
Université Paris-Sud - INRIA Select, France, e-mail: Jean-Michel.Poggi@math.u-psud.fr

Yannig Goude
EDF R&D - Université Paris-Sud, France, e-mail: yannig.goude@edf.fr

One-class classification based on transvariation probability

Francesca Fortunato

Abstract One class classification is justified when objects from one class only, the *target* class, are available. This is different from the traditional classification problem, which tries to *distinguish between* two or more classes with the training set containing objects from all of them. We propose a method for one-class classification based on Gini's transvariation probability between a group and a constant. We explore one-class classification application in recognizing adulterated Olive Oil whereas only the spectrum of Pure Olive Oil is well-known. The method is employed on dimensionally reduced data through Sparse Principal Component analysis.

Keywords

One-class classification; Transvariation probability; Sparse Principal Component analysis

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

Quantified SWOT method and its use in assessing the financial situation of local administrative units

Romana Glowicka-Woloszyn, Aleksandra Łuczak, and Andrzej Woloszyn

Abstract Quantified SWOT method allows to determine the financial position of an administrative unit (for example LAU2) among other same level units that together form a higher level territorial unit. The method constructs two hierarchic schemes pertaining to external and internal SWOT factors. For all administrative units of a given level simple features corresponding to simple factors are normalized. In the proposed process the financial situation of administrative units is assessed and coordinates of their relative position with respect to external and internal financial conditions are calculated, which in turn ranks them among other units. The aim of the paper was to use the quantified SWOT method to assess financial situation of administrative territorial units, perform synthetic evaluations of their external and internal conditions and thereby identify the types of financial position. This process was applied to the assessment of financial situation of LAU2 units of the Wielkopolska province.

Keywords

Quantified SWOT method; financial situation of LAUs

Romana Glowicka-Woloszyn
Poznan University of Life Sciences, Poland, e-mail: roma@up.poznan.pl

Aleksandra Łuczak
Poznan University of Life Sciences, Poland, e-mail: luczak@up.poznan.pl

Andrzej Woloszyn
Poznan University of Life Sciences, Poland, e-mail: woloszyn@awf.poznan.pl

Football and the dark side of cluster analysis

Christian Hennig and Serhat Akhanli

Abstract In cluster analysis, decisions on data preprocessing such as how to select, transform, and standardise variables and how to aggregate information from continuous, count and categorical variables cannot be made in a supervised manner, i.e., based on prediction of a response variable. Statisticians often attempt to make such decisions in an automated way by optimising certain objective functions of the data anyway, but this usually ignores the fact that in cluster analysis these decisions determine the meaning of the resulting clustering. We argue that the decisions should be made based on the aim and intended interpretation of the clustering and the meaning of the variables. The rationale is that preprocessing should be done in such a way that the resulting distances, as used by the clustering method, match as well as possible the "interpretative distances" between objects as determined by the meaning of the variables and objects. Such "interpretative distances" are usually not precisely specified and involve a certain amount of subjectivity. We will use ongoing work on clustering football players based on performance data to illustrate how such decisions can be made, how much of an impact they can have, how the data can still help with them and to highlight some issues with the approach.

Keywords

transformation; dimension reduction; distance design

Christian Hennig
Department of Statistical Science, UCL, United Kingdom, e-mail: c.hennig@ucl.ac.uk

Serhat Akhanli
Department of Statistical Science, UCL, United Kingdom, e-mail: serhat.akhanli.14@ucl.ac.uk

References

COX, T. F. and M. A. A. Cox (1994). *Multidimensional Scaling*. Boca Raton: Chapman and Hall.

GOWER, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27, 857–874.

HENNIG, C. and B. HAUSDORF (2006). Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna (Eds.), *Data Science and Classication*, pp. 29–38. Springer, Berlin.

Segmentation of Online Consumer Behaviour of Polish Youth: a Means-End Approach

Anna Mirosława Myrda

Abstract The objective of this paper is to present results of research on segmentation of online consumer behaviour of polish youth. The research was based on the Means-End Theory and the basic of segmentation were the Means-End Chains. The Means-End Chains are built with qualitative and sequential data describing the cognitive-motivational structures of consumers. One consumer can generate more than one Means-End Chain, chains can be nested. An exploratory cluster analysis was used to built the segmentation models. 1 000 student of the last classes of high school from Malopolska (Poland) participated in the research. The sample was randomly selected. Two measures of sequence dissimilarity were used in analysis. With each of them alternative models of segmentation were built models differed in the number of groups. A set of quantitative cluster validity indexes were calculated for each of the segmentation model. Models with the same number of groups formed on the basis of the different sequence dissimilarity measures were compared with the adjusted Rand index and the clusters characteristics.One final model was chosen, the influence of the sequence dissimilarity measures on the grouping results was discussed.

Keywords

market segmentation; Means-End Chains; cluster analysis

Anna Mirosława Myrda
Cracow University of Economics, Poland, e-mail: anna.m.myrda@gmail.com

References

REYNOLDS, T.J. and OLSON, J.C. (Eds.): *Understanding Consumer Decision Making*. The Means-End Approach to Marketing Decision Making and Advertising Strategy. Lawrence Earlbaum Associates, Mahwah, NJ

GABADIHNO, A.and RITSCHARD, G. and MULLER, N.S. and STUDER, M. (2011): Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–3

KACIAK, E. and CULLEN, C. (2006): Analysis of Means-End Chain data in Marketing Research. *Journal of Targeting, Measurement and Analysis for Marketing*, 15(1), 12–20

Classification Methods in the Research on the Financial Standing of Construction Enterprises after Bankruptcy in Poland

Barbara Pawełek, Jadwiga Kostrzewska, Artur Lipieta, Maciej Kostrzewski, and Krzysztof Gałuszka

Abstract The problem of the enterprise's bankruptcy is an important issue of economic sciences. There are two types of bankruptcy in Poland: bankruptcy with the possibility to make an arrangement and bankruptcy by liquidation. The court obeys the precept according to which the bankruptcy proceedings should be conducted so that the creditorsf claims could be satisfied as much as possible, and the enterprise could be saved. The applications of multivariate statistical analysis are focused more on the prediction of bankruptcy than the evaluation of the financial standing of enterprises after bankruptcy. Studying the path of recovering from insolvency can be a source of valuable information useful in assessing the probability of a success of the restructuring. The aim of the paper is to present an application of the selected classification methods in the research on the financial standing of construction enterprises after bankruptcy. The classification as "healthy" enterprises and the ones "after bankruptcy" is based on, inter alia, logistic regression, classification tree. To evaluate the classification accuracy we use e.g. sensitivity, AUC. We apply the one-dimensional methods (e.g. Tuckeyfs criterion) and the multi-dimensional methods

Barbara Pawełek
Cracow University of Economics, Department of Statistics, Poland, e-mail: barbara.pawelek@uek.krakow.pl

Jadwiga Kostrzewska
Cracow University of Economics, Department of Statistics, Poland e-mail: jadwiga.kostrzewska@uek.krakow.pl

Artur Lipieta
Cracow University of Economics, Department of Statistics, Poland, e-mail: artur.lipieta@uek.krakow.pl

Maciej Kostrzewski
Cracow University of Economics, Department of Econometrics and Operational Research, Poland, e-mail: maciej.kostrzewski@uek.krakow.pl

Krzysztof Gałuszka
University of Economics in Katowice, Department of Finance, Poland, e-mail: krzysztof.galuszka@ue.katowice.pl

(e.g. depth function) to detect outliers. The research covers construction enterprises in Poland from 2005 to 2009.

Keywords

Classification Methods; Financial Standing; After Bankruptcy

References

ALTMAN, E.I. and BRANCH, B. (2015): The Bankruptcy System?s Chapter 22 Recidivism Problem: How Serious is It?. The Financial Review, 50, 1–26.

HOTCHKISS, E.S., JOHN, K., MOORADIAN, R.M. and THORBURN, K.S. (2008): Bankruptcy and the resolution of financial distress. In: B. Eckbo (Ed.): *Handbook of Corporate Finance: Empirical Corporate Finance*, Vol. 2. Elsevier, North-Holland.

PAWEŁEK, B., KOSTRZEWSKA, J. and LIPIETA, A. (2015): The Problem of Outliers in the Research on the Financial Standing of Construction Enterprises in Poland. In: M. Papież and S. Śmiech (Eds.):*Proceedings of the 9th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-economic Phenomena*. Foundation of the Cracow University of Economics, Cracow.

Clustering and Classification methods for an experimental study on prion diseases

Giorgia Rocco

Abstract We report on a hybrid approach to analyze a dataset derived from an experimental study on prion diseas conducted at the Istituto Superiore di Sanità. The data comes from inoculating different strains (inocula) of the disease to bank voles. The aim of the research is to understand at what extent some phenotipic outcomes such as survival times and profiles of brain lesions are able to detect the underlying heterogeneous multi-level origin of the data. We use first an ensemble of hierarchical clustering through the Gower index, a general coefficient that includes similarity for different metrics in the dataset (quantitative and ordinal data). We have verified the ability of the proposed approach to match some preliminary knowledge on the underlying group structure with some possible hint at detecting a slightly finer structure. We then consider alternative classifiers with the aim of validating alternative clustering structures and predict whether a new observation belongs to a group or another.

Keywords

Unsupervised classification; Supervised classification; Ensemble clustering; Gower coefficient; Cross Validation; Classifiers combination

References

GORDON AD, VICHI M (2001), Fuzzy Partition Models for Fitting a Set of Partitions. *Psychometrika*, 66(2), 229–248.

Giorgia Rocco
University La Sapienza, Rome, Italy, e-mail: g.rocco@uniroma1.it

GOWER, J.C.(1971), A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. (2009), *The elements of statistical learning (Vol. 2, No. 1)*, New York: Springer.

HORNIK, K.(2005), A clue for cluster ensembles. *Journal of Statistical Software*,14(12).

HORNIK, K.(2005), *Cluster ensembles. In Classification the Ubiquitous Challenge*, (pp.65–72). Springer Berlin Heidelberg.

KAKOUROU, A., VACH, W., MERTENS, B. (2014), Combination Approaches Improve Predictive Performance of Diagnostic Rules for Mass-Spectrometry Proteomic Data. *Journal of Computational Biology*, 21(12), 898–914.

MERTENS, B.J., NOO, M.D.,TOLLENAAR, R.A., DEELDER, A.M.(2006), Mass spectrometry proteomic diagnosis: enacting the double cross-validatory paradigm. *Journal of Computational Biology*, 13(9), 1591–1605.

Repeated measures analysis for functional data using two-cumulant approximation - with applications

Łukasz Smaga

Abstract The repeated measures analysis for functional data is investigated. A new testing procedure for the two-sample problem when the data are from the same subject is proposed. This test is based on the two-cumulant approximation for the distribution of the test statistic considered in the literature. It is shown that the estimated critical value tends to its theoretical critical value, as the number of observations tends to infinity. The known permutation and bootstrap approximations for the distribution of the test statistic may be time-consuming in contrast to the new one. Via intensive simulation studies, it is found that in terms of size control and power, the new test is comparable with the known tests. An illustrative example of the use of the tests in practice is also given.

Keywords

functional data; repeated measures analysis; two-cumulant approximation

References

MARTINEZ-CAMBLOR, P. and CORRAL, N. (2011): Repeated measures analysis for functional data. *Computational Statistics & Data Analysis*, 55, 3244-3256.

ZHANG, J.-T. (2005): Approximate and asymptotic distributions of chi-squared-type mixtures with applications. *Journal of the American Statistical Association*, 100, 273-285.

Łukasz Smaga
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland, e-mail: ls@amu.edu.pl

ZHANG, J.-T. (2013): *Analysis of Variance for Functional Data*. Chapman & Hall, London.

Coping with poor information for classifiers to support business decision-making. A case study.

Annalisa Stacchini

Abstract Classifiers are (also) valuable for optimizing business decisions. Still in the 'era of big data', for some business problems information on effective predictor variables is inaccessible, or the quality and-or quantity of corporate data is insufficient for getting not just statistically acceptable, but also economically profitable results. This depends on the capability of predicting the outcomes generating the highest losses or profits, which are generally the most rarely observed. Bayesian inference can be resolutive if subjective or firm-level prior information is retrievable. Otherwise it is necessary to recur to 'non-traditional' strategies. The Maximum Likelihood method can be replaced by a classification rule that directly optimizes the economic result. If data allow not even this, it is advisable to set an uncertainty interval of probability values, defining the cases for which outcomes can be considered equiprobable. As last resort, it is possible to guess some parameters, based on the literature, the qualitative understanding of the context and the characteristics of predicted probabilities. Then simulate 'guess-based data' for estimation and validation. This work presents these strategies to cope with poor information in the case of deciding which fare, a company aiming at minimizing the cost of flights for business travel, should choose.

Keywords

business decision-making; poor information; business travel

Annalisa Stacchini
University of Bologna, Italy, e-mail: annalisa.stacchini2@unibo.it

Analysis of Quality of Life among Hemophiliac Patients Using Scores of Medical Outcomes Study

Shinobu Tatsunami

Abstract Hemophilia is a genetic disease and its definitive treatment hasn't been established. Recently, the regular replacement therapy is adopted as the standard therapy in advanced countries. This therapy is to prevent bleeding in joints, and this resulted in the improvement of patients' activities. In order to evaluate present status of QOL in Japanese patients, scores of Medical Outcomes Study 36-Item Short-Form Health Survey (SF36) were analyzed. Data collected by the Research Committee on QOL Study Regarding Coagulation Disorders in Japan were used. Therein, the number of patients older than 16 years old was 724. Patients with other clotting disorders were also included. Descriptive statistics were performed in various groups classified by background factors such as disease type and severity, age, frequency of bleeding, presence of inconvenient joints, and infection with HIV/HCV. Principal component analysis (PCA) was performed using eight subscale scores of SF36. Regarding almost all the background factors, worse condition was related to lower scores. Although there was no clear subgroup structure on the biplot of patients, the advantage of younger age could be clearly illustrated on it. Young patients are usually under the control of the regular replacement therapy. In addition, young patients are free from the infection with HIV or HCV. Those might be the causes of the advantage of young age.

Keywords

principal component analysis; qol; hemophilia

Shinobu Tatsunami
Unit of Medical Informatics, St. Marianna University School of Medicine, Japan e-mail: s2tatsu@marianna-u.ac.jp>

Progress and Open Problems in Clustering: Beyond Algorithm Development

Margareta Ackerman

Abstract Clustering is a central unsupervised learning task with a wide variety of applications. Due to its wide applicability, the field generates much research activity, the great majority of which focuses on the development of new algorithms. Yet, most users continue to rely on a small number of well-established methods. Furthermore, inherent ambiguity in clustering prevents us from identifying which algorithm is best. In recent years, a new approach for studying clustering algorithms has been developed, which focuses on uncovering essential differences between popular clustering methods instead of providing new ones. This formal approach allows users to utilize prior knowledge about their domain to identify which algorithm is suitable for their application, without having to execute many different techniques. The talk will discuss major developments in this field as well as directions for future work. The talk will address joint work with Shai Ben-David, Simina Branzei, Sanjoy Dasgupta, David Loker, and Sivan Sabato.

Keywords

Clustering; theory

References

ACKERMAN, M. and DASGUPTA S. (2014): Incremental Clustering: The Case for Extra Clusters. *Neural Information Processing Systems Conference* (NIPS).

Margareta Ackerman
Florida State University, United States, e-mail: mackerman@fsu.edu

ACKERMAN, M., BEN-DAVID, S., SABATO, S. and LOKER, D. (2013): Clustering Oligarchies. Proceedings of the *Twelfth International Conference on Artificial Intelligence and Statistics* (AISTATS).

ACKERMAN, M., BEN-DAVID, S., BRANZEI, S. and LOKER, D. (2012): Weighted Clustering. Proc. *26th AAAI Conference on Artificial Intelligence*.

ACKERMAN, M., BEN-DAVID, S. and LOKER, D. (2010): Towards Property-Based Classification of Clustering Paradigms. *Neural Information Processing Systems Conference*, NIPS.

ACKERMAN, M., BEN-DAVID, S. and LOKER, D. (2010): Characterization of Linkage-Based Clustering. 2*3rd International Conference on Learning Theory*, COLT.

ACKERMAN, M. and BEN-DAVID, S. (2009): Clusterability: A Theoretical Study. Proceedings of the *Twelfth International Conference on Artificial Intelligence and Statistics*, AISTATS, JMLR: 5, pp. 1-8.

ACKERMAN, M. and BEN-DAVID, S. (2008): Measures of Clustering Quality: A Working Set of Axioms for Clustering. *Neural Information Processing Systems Conference*, NIPS.

Information-Theoretic Validation of Clustering Algorithms

Joachim M. Buhmann

Abstract Algorithms are usually analyzed according to their time and space requirements, i.e., speed and memory. Machine learning algorithms are additionally required to show robustness to input fluctuations or randomization during execution. This paper elaborates a new framework to measure the *informativeness* of algorithmic procedures and their *stability* against noise. An algorithm is considered to be a noisy channel which is characterized by a generalization capacity ($\mathscr{GC}$). Informative algorithms yield a high generalization capacity, whereas fragile algorithms suffer from a low capacity. $\mathscr{GC}$ objectively ranks different algorithms for the same data processing task based on the bit rate of their respective capacities. The problem of grouping data is used to demonstrate this validation principle for clustering algorithms, e.g. Pairwise Clustering, Correlation Clustering, Normalized Cut and Dominant Set clustering. Our new validation approach selects the most informative clustering algorithm, which filters out the maximal number of stable, task-related bits relative to the underlying hypothesis class of clusterings.(joint work with Morteza Haghir Chehreghani and Ludwig Busse)

Keywords

graph theoretic clustering; information theory; approximation set

Joachim M. Buhmann
ETH Zurich, Switzerland, e-mail: jbuhmann@inf.ethz.ch

Averaging and Asymmetry in Cluster Analysis

Paul McNicholas

Abstract Two issues around the selection of the best clustering model are discussed: averaging and accounting for asymmetry. Both of these issues will be considered within the mixture model-based clustering context. Fitting a family of mixture models and averaging some of the resulting models is a departure from the common single best model paradigm. Whether this departure is likely to improve the situation is a crucial question, and is discussed along with some examples. Consideration is also given to approaches when there is reason to believe that clusters are, or might be, asymmetric. Possible approaches for tackling such situations are discussed as well as strategies for model selection when multiple models are fitted. The talk concludes with a discussion about merging asymmetric components.

Keywords

Asymmetric clusters; mixture models; model averaging.

References

BROWNE, R.P. and MCNICHOLAS, P.D. (To appear): A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, DOI: 10.1002/cjs.11246

WEI, Y. and MCNICHOLAS, P.D. (To appear): Mixture model averaging for clustering, *Advances in Data Analysis and Classification*. DOI: 10.1007/s11634-014-0182-6

Paul McNicholas
McMaster University, Hamilton, Ontario, Canada, e-mail: mcnicholas@math.mcmaster.ca

Disentangling Continuous and Discrete Structure Within Data

Doug Steinley

Abstract When fitting models to data, general assumptions are frequently made automatically without much consideration for their implication on subsequent interpretations. For instance, fitting a standard factor model often presupposes an underlying set of continuous, latent factors. Likewise, when searching for group structure, mixture models (e.g., latent profile analysis, latent class analysis) or cluster analysis are implemented and assume a a set of discrete latent "classes". Usually, the type of model that is fit to the data is governed by the theoretical notions underpinning the substantive question of interest. In this talk, it is shown that both types of structures can be present and correspond to different subsets of the data. A general strategy is discussed for extracting both class structure and factor structure. Demonstrations are given on a data set of internet habits of collegiate students.

Keywords

cluster analysis; factor analysis

Doug Steinley
University of Missouri, USA e-mail: steinleyd@missouri.edu

Financial technical analysis using hidden Markov models

José G. Dias

Abstract Technical analysis, also known as "charting", has been a key instrument among traders and financial analysts for many decades. It uses charts in an attempt to find shapes and patterns that can anticipate trends in financial time series. One of the main difficulties of the technical analysis has been its highly subjective nature as many times the presence of shapes in the historical data is very dependent on the analysts' insights and subjective believes. There are many approaches in technical analysis (e.g., candlestick charting, Dow Theory, and Elliott wave theory), but most analysts combine different techniques between these two extremes: some analysts use subjective judgment to decide patterns and trends; whereas others tend to be more objective using system techniques in the identification of patterns. We introduce a hidden Markov model (HMM) in the context of technical analysis, namely candlestick data. An application uses daily prices from the Standard & Poor's 500 (SP500) from 2 January 1962 to 31 December 2014 drawn from Yahoo!Finance. Results show that HMM reveals the underlying structure in candlestick time series data. In particular, the posterior probability of being in the bear regime can be confirmed as a downturn in the stock market.

Keywords

Time series; Hidden Markov models; Technical analysis; Stock markets

José G. Dias
Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Portugal, e-mail: name@email.address

References

DIAS, J. G., VERMUNT, J. K.,, RAMOS, S. B. (2015): Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research*, 243(3), 852–864.

FISCHER, R., FISCHER, J. (2003): *Candlesticks, Fibonacci, and Chart Pattern Trading Tools: A Synergistic Strategy to Enhance Profits and Reduce Risk*. Wiley, Hoboken, NJ.

NISON, S. (1994): *Beyond Candlesticks: New Japanese Charting Techniques Revealed*. Wiley, New York.

PERSON, J. L. (2004): *A Complete Guide to Technical Trading Tactics: How to Profit using Pivot Points, Candlesticks & other Indicators*. Wiley, Hoboken, NJ.

Power Analysis for the Likelihood Ratio Test in Latent Markov Models: Short-cutting the bootstrap p-value based method

Dereje W. Gudicha, Verena D. Schmittmann, Fetene B. Tekle, and Jeroen K. Vermunt

Abstract In recent years, the latent Markov (LM) model has proven useful to identify distinct unobserved states and transitions between these states over time in longitudinally observed responses. The bootstrap likelihood ratio (BLR) test is becoming a gold standard for testing the number of states, yet little is known about power analysis methods for this test. This paper presents a short-cut to a p-value based power computation for the BLR test. The p-value based power computation involves computing the power as the proportion of the bootstrap p-value (PBP) for which the null hypothesis is rejected. This requires to perform the full bootstrap for multiple samples of the model under the alternative hypothesis. Power computation using the short-cut method involves the following simple steps: obtain the parameter estimates of the model under the null hypothesis, construct the empirical distributions of the likelihood ratio under the null and alternative hypotheses via Monte Carlo simulations, and use these empirical distributions to compute the power. The advantage of this short-cut method is that it is computationally cheaper and is simple to apply for sample size determination.

Dereje W. Gudicha
Tilburg University, Netherlands, e-mail: D.W.Gudicha@uvt.nl

Verena D. Schmittmann
Tilburg University, Netherlands, e-mail: v.d.schmittmann@tilburguniversity.edu

Fetene B. Tekle
Tilburg University, Netherlands, e-mail: f.b.tekle@uvt.nl

Jeroen K. Vermunt
Tilburg University, Netherlands, e-mail: j.k.vermunt@tilburguniversity.edu

Keywords

Latent Markov; Number of States; Likelihood Ratio; Bootstrap; Power Analysis; sample size.

Clusterwise three-way component models to account for heterogeneity in three-way data

Tom F. Wilderjans and Eva Ceulemans

Abstract Nowadays, in many fields of science, challenging research questions often call for the analysis of three-way data sets (e.g., conventional sensory profiling, multi-subject EEG, fluorescence spectroscopy and nuclear magnetic resonance data). To reveal the structure underlying such data, often three-way component methods (e.g., Parafac, Tucker3) have been used. These methods postulate the underlying components being the same for all elements of the three modes, an assumption which may be violated in many situations (e.g., groups of raters adopting different sensory dimensions). To account for this type of heterogeneity, a class of clusterwise three-way component models will be proposed. Members of this class have in common that the elements of one of the modes (e.g., subjects) are clustered and, simultaneously, a three-way component model is fitted to the data within each cluster. As such, qualitative differences (i.e., heterogeneity) in underlying component structure across clusters can be disclosed. The goal of this presentation is to introduce this model class by highlighting some of its representative members (e.g., Clusterwise Parafac) and illustrating them with empirical data. Further, also models in which the number of components is allowed to vary across clusters, resulting in a challenging model selection problem, will be discussed.

Keywords

clustering; three-way data; heterogeneity

Tom F. Wilderjans
Leiden University, Netherlands, e-mail: t.f.wilderjans@fsw.leidenuniv.nl

Eva Ceulemans
Leuven University, Belgium, e-mail: eva.ceulemans@ppw.kuleuven.be

Statistical data depth for clustering macroseismic fields

Claudio Agostinelli, Renata Rotondi, and Elisa Varini

Abstract The size of historical earthquakes is given by the macroseismic intensity, an ordinal variable expressed by different scales and closely related to the effects produced by an earthquake on humans, buildings and natural environment. The collection of intensity values recorded at sites in the area surrounding the seismic source constitutes the macroseismic field of an earthquake.Our aim is to identify clusters of macroseismic fields according to size and shape of their isoseismals (lines of equal shaking). To this end we consider the modified version of the local half-region depth, a nonparametric method especially suitable for ordering irregular curves (related to isoseismal lines in our application) with many crossing points. Then the most central curve represents the global pattern of intensity decay. To deal with the case of possible multiple centres, a hierarchical algorithm is applied to the dissimilarity matrix based on the local modified half-region depth similarity of a pair of curves. The method is first tested on sets of simulated fields divided into groups whose isoseismal lines differ in shape (circle or ellipse), size, eccentricity, and rotation angle. Then we analyse 31 fields associated with earthquakes of intensity IX, drawn from the Italian Macroseismic Database DBMI11.

Keywords

Clustering; isoseismal lines; similarity

Claudio Agostinelli
Department of Environmental Sciences, Informatics and Statistics, Cà Foscari University, Venice, Italy, e-mail: claudio@unive.it

Renata Rotondi
CNR-IMATI Milano, Italy, e-mail: reni@mi.imati.cnr.it

Elisa Varini
CNR-IMATI Milano, Italy, e-mail: elisa@mi.imati.cnr.it

References

AGOSTINELLI, C. (2013): *Local half region for functional data*. Manuscript.

AGOSTINELLI, C. and ROTONDI, R. (2015): Analysis of macroseismic fields using statistical data depth functions. *Bulletin of Earthquake Engineering*, accepted.

LÓPEZ-PINTADO, S. and ROMO, J. (2011): A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55, 1679–1695.

Robust model-based functional clustering of satellite data

Carlo Gaetan, Paolo Girardi, and Roberto Pastres

Abstract Due to human activities or natural causes, eutrophic phenomenon causes many adverse effects for the marine ecosystem and consequently for humans. The European Community indicates the diffuse light attenuation coefficient at 490 nm (Kd490) and the Chlorophyll type-a (Chl-a) concentration as trophic status indicators of the sea-water. High Chl-a and KD490 levels may lead to harmful events. We focus our attention to the Gulf of Gabes in Tunisia, an area knows an increasing eutrophication due to urban interferences. Monthly time series from 2003 to 2011 of Chl-a concentrations and Kd490 levels are available from satellite data sensors. Satellite data are affected by missing values and measurement errors. We will follow a model approach to the functional data clustering in which time series are supposed to be generated by a mixture of models for functional curves. To take into account robustness issues and possible spatial dependences we propose a two-stage approach. In the first stage, conditional to the cluster membership, we model the time series by an asymmetric Laplace distribution and in second stage (cluster membership model) we consider a Markov random field. Our proposal shines some light on the complex task of spatial marine zoning for fisheries and conservation.

Keywords

functional data clustering; laplace distribution; spatial dependence

Carlo Gaetan
Ca' Foscari University of Venice, Italy, e-mail: gaetan@unive.it

Paolo Girardi
Ca' Foscari University of Venice, Italy, e-mail: paolo.girardi@unive.it

Roberto Pastres
Ca' Foscari University of Venice, Italy, e-mail: pastres@unive.it

References

GUYON, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media.

JAMES, G. M., & SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462), 397–408.

JIANG, H., & SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, 54(2), 108–119.

Functional Data Analysis for the investigation of longitudinal response patterns in health care

Sugnet Lubbe, Felix Dube, Mark Nicol, and Heather Zar

Abstract A total of 137 subjects are followed over a year with two weekly intervals. At each measurement the presence or absence of a specific pathogen is noted, and if the pathogen is present, the specific serotype is determined. The aim of the study is to investigate patterns of colonisation and relationships with disease. Functional data analysis is used to model the occurrence of serotypes. Furthermore the subjects are vaccinated at three occasions during the study. Only a few of the serotypes are covered by the vaccine. A complicating factor is that the vaccinations did not all take place at the same point in time for all the subjects. The use of functional data analysis and specifically curve registration to 'align' the vaccination points is illustrated in the comparison of colonisation patterns in response to the vaccinations.

Keywords

Functional Data Analysis; Longitudinal data

Sugnet Lubbe
Department of Statistical Sciences University of Cape Town Country , South Africa, e-mail: Sugnet.Lubbe@uct.ac.za

Felix Duber
Division of Medical Microbiology, University of Cape Town, South Africa e-mail: Felix.Dube@myuct.ac.za

Mark Nicol
Division of Medical Microbiology, University of Cape Town, South Africa e-mail: Mark.Nicol@uct.ac.za

Heather Zar
Department Paediatrics and Child Health, Red Cross Hospital, Medical Research Council unit on Child and Adolescent Health, University of Cape Town, South Africa, e-mail: Heather.Zar@uct.ac.za

Clustering word life-cycles in chronological corpora: what data transformation for differing clustering goals

Matilde Trevisani and Arjuna Tuzzi

Abstract Chronological corpora are collections of texts ordered in time. Texts are often grouped into equal time intervals and, in *bag-of-words* approaches, the processed data are typically the frequencies of individual words in the set of texts referred to the same time-point.The temporal course of a word occurrence is viewed as a proxy of a word diffusion and vitality, *i.e.* of a word life-cycle. Recognition of temporal shapes and clustering of words having similar life-cycles are the basic objective. However, the strong asymmetry of the frequency spectrum typical of textual data (*Large Number of Rare Events*) has to be taken into account when defining the specific purpose of clustering and, hence, the type of any further processing of the data.The crucial decision is whether similarity essentially depends on the degree of synchronization or also on the level of word popularity (in a functional data analysis approach, whether to compare curves horizontally or also in terms of their amplitude variation).Several column normalizations coupled with row normalizations are applied to the word×time contingency table of corpus data. By applying constrained spline smoothing and distance-based curve clustering, the effect of selected data transformations on the generation of word groups is examined.

Keywords

chronological corpora; data normalization; curve clustering

Matilde Trevisani
University of Trieste, Italy, e-mail: matilde.trevisani@deams.units.it

Arjuna Tuzzi
University of Padova, Italy, e-mail: arjuna.tuzzi@unipd.it

References

GIACOFCI, M., LAMBERT-LACROIX, S., MAROT, G., PICARD, F. (2013): Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1), 31–40.

RAMSAY, J. O., HOOKER, G., GRAVES, S. (2009): *Functional Data Analysis with R and Matlab, Use R!*. Springer.

TREVISANI, M. and TUZZI, A. (2014): A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal. *Quality and Quantity*, DOI 10.1007/s11135-014-0050-7.

Estimating the average treatment effect through Balancing Recursive Partitioning

Claudio Conversano, Massimo Cannas, and Francesco Mola

Abstract A tree-based approach for identification of a balanced group of observations in causal inference studies is presented. The method uses an algorithm based on a multidimensional balance measure criterion applied to values of the covariates to recursively split the data. Starting from an ad-hoc resampling scheme, observations are finally partitioned in subsets characterized by different degrees of homogeneity, and causal inference is carried out on the most homogeneous subgroups.The proposed methodology is applied to the NSW data set analyzed originally by Lalonde with the aim to compare the estimators it provides with those currently available in the literature.

Keywords

Regression trees; Resampling; Average Treatment Effect; Balancing Recursive Partitioning

References

ABADIE, A. and IMBENS, G. (2006): Large Sample Properties of Matching Estimators for Average Treatment Effects.*Econometrica*, Vol. 74, No. 1, 235–267

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

Massimo Cannas
University of Cagliari, Italy, e-mail: massimo.cannas@unica.it

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

LALONDE, R. (1986): Evaluating the Econometric Evaluations of Training Programs, *American Economic Review*, 76: 604–620

PORRO, G. and IACUS, S.M. (2009): Random Recursive Partitioning: A matching method for the estimation of Average Treatment Effects. *J. of Appl. Econom*. 24, 163–185

,,

A framework for measuring the stability of recursive partitioning results

Michel Philipp, Thomas Rusch, Kurt Hornik, and Carolin Strobl

Abstract Recursive partitioning approaches, such as classification and regression trees and model-based recursive partitioning, have become established and frequently applied methods for exploring unknown structures in complex data. Despite their popularity, a major drawback of these methods is their instability, since small random changes in the data can cause large changes in the results. For the interpretation in practical applications, however, stability is a crucial requirement to draw consistent conclusions - but currently recursive partitioning methods provide no statistical theory for judging the confidence of their results. We therefore present a new framework for assessing the stability of recursive partitioning results, that is based on a family of distance measures for partitions. The new approach is motivated, illustrated and compared to existing distance measures by means of real and simulated examples.

Keywords

recursive partitioning; stability measuring framework; distance measure

Michel Philipp
University of Zurich, Switzerland, e-mail: m.philipp@psychologie.uzh.ch

Thomas Rusch
Vienna University of Economics and Business, Austria, e-mail: thomas.rusch@wu.ac.at

Kurt Hornik
Vienna University of Economics and Business, Austria, e-mail: kurt.hornik@wu.ac.at

Carolin Strobl
University of Zurich, Switzerland, e-mail: carolin.strobl@psychologie.uzh.ch

Combining model-based recursive partitioning and random-effects estimation for the detection of treatment subgroups

Marjolein Fokkema

Abstract Identification of subgroups of patients for which treatment A is more effective than treatment B, and vice versa, is of key importance to the development of personalized medicine. Several tree-based algorithms have already been developed for the detection of such treatment-subgroup interactions, but none of those algorithms allow for taking into account clustering structures in a dataset. In this paper, we propose an algorithm that allows for detection of treatment-subgroup interactions and estimation of cluster-specific effects. The new algorithm uses model-based recursive partitioning to detect treatment-subgroup interactions, and a linear mixed-effects model for estimation of random-effects parameters. In a simulation study, we evaluate the performance of the new algorithm, and compare it with that of model-based recursive partitioning without random-effects estimation. We will provide an illustration, by applying the algorithm to an existing dataset of treatment outcomes. Finally, we will discuss (dis)advantages of the new algorithm, and some directions for further research.

Keywords

Model-based recursive partitioning; treatment subgroups; generalized linear mixed-effects models

References

BATES, D., MÄCHLER, M., BOLKER, B., and WALKER,. S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Marjolein Fokkema
Universiteit Leiden, Netherlands, e-mail: m.fokkema@fsw.leidenuniv.nl

ZEILEIS, A., HOTHORN, T., and HORNIK, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17, 492–514.

Meta-Cart: Integrating Classification and Regression Trees into Meta-analysis

Xinru Li, Elise Dusseldorp, and Jacqueline Meulman

Abstract Meta-analysis is an important tool to synthesize results from multiple studies in a systematic way. Interaction effects play a central role in assessing conditions under which the relationship between study features and effect size (the outcome variable) changes in strength and/or direction. Within the framework of meta-analysis, when several study features are available, meta-regression lacks sufficient power to detect interactions between them. To overcome this shortcoming, a new approach named "meta-CART" introduced Classification and Regression Trees (CART) in the field of meta-analytic data to identify interactions. The current implementation of meta-CART has its shortcomings: when applying CART, the sample sizes of studies are not taken into account, and the effect size is dichotomized around the median value. In our presentation, we will overcome these shortcomings by 1) weighting the study effect sizes by their accuracy, and 2) using the numerical values of the outcome variable instead of dichotomization. The new methodology will be compared to the current meta-CART in terms of Type I error, Power, and recovery performance in a Monte Carlo simulation study. Our initial results are promising, and an extensive simulation study for different population effect sizes and heterogeneity magnitudes will be presented.

Xinru Li
Mathematical Institute, Leiden University, Netherlands, e-mail: x.li@math.leidenuniv.nl

Elise Dusseldorp
Mathematical Institute, Leiden University, Netherlands, e-mail: elise.dusseldorp@math.leidenuniv.nl

Jacqueline Meulman
Mathematical Institute, Leiden University, Netherlands, e-mail: jmeulman@math.leidenuniv.nl

Keywords

meta-analysis; classification; binary tree; moderator; interaction

References

DUSSELDORP, E., VAN GENUGTEN, L., VAN BUUREN, S., VERHEIJDEN, M. W., and VAN EMPELEN, P. (2014). Combinations of techniques that effectively change health behavior: Evidence from Meta-CART analysis. *Health Psychology*, 33, 1530–1540.

Real-time classification of in-flight aircraft damage

Brenton S. Blair and Herbert K. H. Lee

Abstract When there is a sudden equipment malfunction or damage to an aircraft, it is critical to be able to quickly detect and diagnose the problem, so that the pilot can attempt to maintain control of the aircraft and land safely. Methodology is developed for real-time classification of flight trajectories to be able to distinguish between undamaged aircraft and several different situations of damage. Principal components decomposition allows a lower-dimensional representation of three-dimensional trajectories in time. Classification trees provide a computationally efficient approach with sufficient accuracy to be able to detect and classify the different situations quickly. Results are demonstrated using runs from a flight simulator in collaboration with NASA (the United States National Aeronautics and Space Administration).

Keywords

Classification tree; principal components; flight trajectory

Brenton S. Blair
Department of Applied Math & Statistics University of California, Santa Cruz, United States, e-mail: bsblair@ams.ucsc.edu

Herbert K.H. Lee
Department of Applied Math & Statistics University of California, Santa Cruz, United States, e-mail: herbie@ams.ucsc.edu

The δ-machine

Mark De Rooij

Abstract We introduce the δ-machine, a statistical learning tool for classification based on dissimilarities or distances, δ, between inputs. Compared to other statistical learning tools, which are often black boxes, this machine has a clear interpretation in terms of distances towards a prototype or exemplar. We introduce the machine, discuss its properties, derive variable importance measures and partial dependence plots for the machine, and show toy as well as empirical examples. Detailed interpretations of the machinery will be discussed.

Keywords

Classification; Statistical Learning; Dissimilarities

Mark De Rooij
Leiden University, Institute of Psychology, Netherlands, e-mail: Rooijm@fsw.leidenuniv.nl

The Five Factor Model of personality and evaluation of drug consumption risk

Elaine Fehrman, Awaz K. Muhammad, Evgeny Mirkes, Vincent Egan, and Alexander N. Gorban

Abstract The problem of evaluating an individual's risk of drug consumption and misuse is highly important and novel. An online survey methodology was employed to collect data including personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information. The data set contained information on the consumption of 18 central nervous system psychoactive drugs. Correlation analysis using a relative information gain model demonstrates the existence of a group of drugs (amphetamines, cannabis, cocaine, ecstasy, legal highs, LSD, and magic mushrooms) with strongly correlated consumption. An exhaustive search was performed to select the most effective subset of input features and data mining methods to classify users and non-users for each drug. A number of classification methods were employed (decision tree, random forest, k-nearest neighbors, linear discriminant analysis, Gaussian mix, probability density function estimation, logistic regression and naïve Bayes) and the most effective method selected for each drug. The quality of classification was surprisingly high. The best results with sensitivity and specificity being greater than 75% were achieved for VSA (volatile sub-

Elaine Fehrman
Men's Personality Disorder & National Women's Directorate, Rampton Hospital, Retford, Nottinghamshire, DN22 0PD, United Kingdom, e-mail: Elaine.Fehrman@nottshc.nhs.uk

Awaz K. Muhammad
University of Leicester, Department of Mathematics, Leicester, LE1 7RH, United Kingdom, e-mail: akm40@leicester.ac.uk

Evgeny Mirkes
University of Leicester, Department of Mathematics, Leicester, LE1 7RH, United Kingdom, e-mail: em322@le.ac.uk

Vincent Egan
Department of Psychiatry and Applied Psychology, University of Nottingham, Nottingham, NG8 1BB, United Kingdom, e-mail: Vincent.Egan@nottingham.ac.uk

Alexander N Gorban
University of Leicester, Department of Mathematics, Leicester, LE1 7RH, United Kingdom, e-mail: ag153@le.ac.uk

stance abuse) and methadone. The poorest result was obtained for prediction of alcohol consumption.

Keywords

Risk; decision tree; kNN; radial basis functions; features selection; visualization

K-NN controlled condensation: a new method for data preprocessing in classification tasks

Carmen Villar-Patiño and Carlos Cuevas-Covarrubias

Abstract Accuracy and speed are two very important features for any classification algorithm. Usually in practice, it is difficult to optimize both simultaneously. K-NN methods are accurate an easy to implement; however, this approach implies an intensive computational work that makes it too slow to be applied in some contexts. This problem becomes especially important when the training dataset is large. We propose a new preprocessing method that condensates training datasets in order to make K-NN classification more efficient without a significant loss of precision. Given an initial sample, it defines a small subset of well selected informative observations. Once the sample size is reduced, K-NN classification becomes faster. Contrasting with previous methods reported in the literature, this new idea includes a parameter that helps the user to control the exchange of speed for precision. The performance of this new technique is assessed with several examples from different scientific areas, using the K-NN model based approach as a benchmark. This assessment is limited to two and three category problems. Never the less, the extension to more than three classes is straight forward.

Keywords

Supervised classification; prototype selection; real time processing

Carmen Villar-Patiño
Universidad Anahuac, Facultad de Ingeniería, Mexico, e-mail: maria.villar@anahuac.mx

Carlos Cuevas-Covarrubias
Universidad Anahuac, Facultad de Ciencias Actuariales, Mexico, e-mail: ccuevas@anahuac.mx

The conference is supported by:

CONFCOMMERCIO
IMPRESE PER L'ITALIA
ASCOM PROVINCIA DI BOLOGNA

Ospitalità a Bologna

sas
THE POWER TO KNOW.

Springer