DAUPHINE
UNIVERSITÉ PARIS
CEREMADE LAMSADE

Workshop

ADVANCES IN DATA SCIENCE FOR BIG AND COMPLEX DATA

From data to classes and classes as new statistical units

UNIVERSITY PARIS-DAUPHINE

January 10-11, 2019

ROOM C108 First floor

In all domains of human activity, we are more and more faced with the problem of understanding and extracting knowledge from standard, big and complex data, often multi-sources (as mixture of numerical, textual, image, social networks data). New tools are needed to transform huge data bases intended for management to data bases usable for Data Science tools. This transformation leads to the construction of new statistical units described by aggregate data in term of symbols (intervals, distributions, list, etc.) as single-valued data are not suitable because they cannot incorporate the additional information on data structure available in symbolic data. Data Science, considered as a science by itself, consists in general terms, in the extraction of knowledge from data. Symbolic data analysis (SDA) provides a new way of thinking in Data Science by extending the standard input to a set of classes of individual entities. SDA is an emerging area of Data Science based on aggregating individual level data into group-based summarized by symbols and developing Data Science methods to analyze them. At the crossroad of statistics, mathematics and computer science, it is ideal for the analysis of large and complex datasets, and has immense potential to become a standard methodology in the near future.

The aim of this workshop is to present recent advances in reasoning from data to classes and considering classes as new statistical units, to academic researchers and industrials in all domains where data are obtained and need to be analyzed for understanding them and improving decisions. After the lectures, you will be able to participate in recent software training, illustrated with simple examples, using your personal laptops.

Academics with mathematical, statistical and computer solutions to the problems raised by standard, complex and (or) massive data as well as industrials or laboratories wishing to present open problems concerning this type of data, are welcome to participate.

Web site: http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:pa19

Steering Committee

P. Bertrand	(CEREMADE, Université Paris-Dauphine)
E. Diday	(CEREMADE, Université Paris-Dauphine)
W. Litwin	(LAMSADE, Université Paris-Dauphine)

Scientific Committee

V. Batagelj	(University of Ljubljana, Slovenia)
L. Billard	(University of Georgia, USA)
P. Brito	(University of Porto, Portugal)
V. Cariou	(ONIRIS, INRA, France)
P. Cazes	(CEREMADE, University Paris-Dauphine)
D. Colazzo	(LAMSADE, University Paris-Dauphine)
F. De Carvalho	(CIn-UFPE, Recife, Brazil)
D. Desbois	(INRA-AgroParisTech, University Paris-Saclay)
R. Emilion	(University of Orléans, France)
C. Guinot	(University of Tours & PhG-Bioconsulting, France)
M. Ichino	(College of Science and Engineering, Tokyo Denki University, Japan)
M. Mizuta	(Hokkaido University, Japan)
F. Lebaron	(ENS Paris-Saclay – Cachan, France)
M. Noirhomme	(FUNDP Namur University, Belgium)
Y. Lechevallier	(Directeur de recherche honoraire, INRIA, France)
M. Nadif	(University Paris-Descartes, France)
S. Pinson	(LAMSADE, University Paris-Dauphine)
O. Rodriguez	(University of Costa-Rica, San José, Costa Rica)
G. Saporta	(Conservatoire National des Arts et Métiers, Paris)
S. Sisson	(University of New South Wales, Sydney, Australia)
C. Toque	(Min.de l'Écologie et du Développement Durable - Paris la Défense, France)
R. Verde	(University of Campania "Luigi Vanvitelli", Caserta, Italy)
H. Wang	(Beihang University, Beijing, China)

Workshop supported by:

Société Francophone de Classification

Société Française de Statistique

Association EGC
Association Internationale Francophone d'Extraction et de Gestion des Connaissances

CATI CITISES

PROGRAMME

Thursday, January 10th 2019

8h45 – 9h00 WELCOME SESSION E. Diday, C. Guinot

SESSION 1: *Opening session*

9h00 – 9h45 P. Brito
SDA: Past, present and future

SESSION 2: *Prediction*

9h45 – 10h15 A. Balzanella, R. Verde, A. Irpino
Spatial prediction on data streams represented by histogram

10h15 – 10h45 *Coffee Break*

10h45 – 11h15 N.Niang, S. Bougeard, V. Cariou, G. Saporta
Prediction in clusterwise multiblock PLS
11h15 – 11h45 H. Wang, S. Lua, J. Zhaoa
Aggregating multiple types of complex data in stock market prediction: A model-independent framework
11h45 – 12h15 H. Wang, T. Huang, and S. Wang
Artificial neural network method incorporating spatial structure with its Application in Prediction of PM2.5

12h15 – 14h00 *LUNCH*

SESSION 3: *Interval Data*

14h00 – 14h30 F. A. T. De Carvalho, E. de A. Lima Neto, U. da Nobrega
A new exponential-type kernel robust regression for interval-valued data
14h30 – 15h00 Y. Sun, X. Zhang, A. T.K. Wan, S. Wang
Model Averaging for Interval-valued Data.
15h00 – 15h30 D. Desbois
Exploring the distribution of conditional quantile estimate ranges: an application to the estimation of specific production costs of pig in the European Union.

15h30 – 16h00 *Coffee Break*

SESSION 4: *Bibliographic Analysis and Economy*

16h00 – 16h30 V. Batagelj, D. Maltseva
Temporal bibliographic analysis
16h30 – 17h00 Andrej Srakar,
Symbolic input-output analysis: a harmonic analysis approach to combining statistical distributions
17h00 – 17h30 M. Febrissy and M. Nadif
Co-clustering via Nonnegative Matrix Tri-Factorization: A comparative study

SESSION 5: *Software*

17h30 – 18h30 F. Afonso
Software for Symbolic Data Analysis through Industrial Applications

Friday, January 11th

SESSION1: *Clustering*

9h00 – 9h30 Y. Lechevallier, F. A. T. de Carvalho
Weighted multi-view partitioning of time series
9h30 – 10h00 Ch. Biernacki, M. Marbac, V. Vandewalle
Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering
10h00 – 10h30. P. Bertrand, J. Diatta
Multilevel clusterings as abstract convexities

10h30 – 11h00 *Coffee Break*

11h00 – 11h30 L. Billard
Clustering of Symbolic Data
11h30 – 12h00 V. Cariou, T. F. Wilderjans
Extending the clustering of variables around latent components approach to three-way data by means of clusterwise Parafac
12h00 – 12h30 A. Gloaguen, A. Tenenhaus
Regularized Generalized Canonical Correlation Analysis extended to multiway data with a medical application

12h30 – 14h00 *LUNCH*

SESSION 2: *Basic formalism and examples*

14h00- 14h30 L. Billard
A Brief Overview of Symbolic Data in the Statistical Framework.
14h30 – 15h00 R. Emilion
Likelihood in the symbolic context with examples
15h00 – 15h30 E. Diday
Concordance and discordance between classes of complex data

15h30 – 16h00 *Coffee Break*

SESSION 3: *Data Bases and Software*

16h00 – 16h30 W. Litwin
SQL for Stored and Inherited Relations
16h30 – 17h00 Ch. Bienarcki, F. Afonso
Software presentation and discussion

17h00 – 17h30 Round Table

ABSTRACTS

Thursday, January 10th 2019

8h45 - 9h00 WELCOME ADDRESS E. Diday, C. Guinot

SESSION 1 - Opening session

9h00 – 9h45 P. Brito

University of Porto & LIAAD-INESC TEC, Portugal

Symbolic Data Analysis: past, present and future

Abstract: Since its introduction by Diday in the eighties of last century, Symbolic Data Analysis has known a considerable development. It emerged from the need to consider data that contain information which cannot be represented within the classical data models, together with the objective of designing methods that produce results directly interpretable in terms of the input descriptive variables. The "model" for data representation should allow taking into account intrinsic variability, therefore allowing representing with a same language, e.g., elements and clusters of a given set.
In this talk we recall the first logical-rooted models for symbolic data representation, and the data analysis approaches that resulted from such models. We then see how the community moved from Symbolic Data-Analysis (a different approach to analyse data) to Symbolic-Data Analysis (the analysis of Symbolic Data), defining new variable types and adopting a tabular-based representation. We present the main approaches for the analysis of different types of symbolic data, trying to put in evidence their specificities and connections. Finally we refer current developments and discuss possible avenues for future research.

References

Bertrand, P. and Goupil, F.(2000). Descriptive statistics for symbolic data. In: Analysis of Symbolic Data, H.-H. Bock and E. Diday (Eds.), 106-124, Springer-Verlag, Berlin-Heidelberg.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. Journal of the American Statistical Association 98(462), 470-487.

Bock, H.-H. and Diday, E. (Editors) (2000). Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Berlin-Heidelberg.

Brito P. and Duarte Silva A.P. (2012). Modelling interval data with Normal and Skew-Normal distributions. Journal of Applied Statistics, 39(1), 3_20.

Dias, S. and Brito, P. (2015). Linear Regression model with histogram-valued variables. Statistical Analysis and Data Mining, 8(2), 75-113.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: the basic choices. In: Classification and Related Methods of Data Analysis, Proc. of IFCS'87, H.-H. Bock (Ed.), 673-684, North Holland, Amsterdam.

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

Le-Rademacher, J. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. Journal of Statistical Planning and Inference, 141, 1593-1602

Neto, E.A.L. and De Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. Computational Statistics & Data Analysis, 52(3),1500-1515.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic Data Analysis. Statistical Analysis and Data Mining, 4(2), 157_170.

Verde, R. and Irpino, A. (2007). Dynamic clustering of histogram data: using the right metric. In: Selected Contributions in Data Analysis and Classification, P. Brito (Ed.), 123-134, Springer, Berlin, Heidelberg.

SESSION 2 - Prediction

9h45 – 10h15 A. Balzanella, R. Verde, A. Irpino

Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Italy

Spatial prediction on data streams represented by histogram

Abstract: Our proposal consists in a strategy of analysis of data streams recorded by geo-referenced sensors. We focus on the problem of measuring the spatial dependence among the observations recorded over time and with the prediction of the data distribution, where no sensor record is available.
The proposed strategy is made by two main steps:
- an online step, where the incoming data records are summarized by histograms;
- an off line step, that performs the measurement of the spatial dependence and the spatial prediction.
The main novelties are the introduction of the variogram and the kriging for histogram data. Through these new tools we can monitor the spatial dependence and to perform the prediction starting from histogram data, rather than from the sensor records, using L2 Wasserstein metric.
The effectiveness of the proposal is evaluated on real and simulated data.

References

Aggarwal, C. C., Han, J., Wang, J., Yu, P. Clustream: a framework for clustering evolving data streams. In Very Large Data Bases, 2003

Arroyo J., Maté C. Forecasting histogram time series with k-nearest neighbours methods, In International Journal of Forecasting, Volume 25, Issue 1, 2009, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2008.07.003.

Bock, H.-H., Diday, E. (Eds.). Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin-Heidelberg, 2000

Balzanella, A., Rivoli, L., Verde, R. Data Stream Summarization by Histograms Clustering. In P. Giudici, S. Ingrassia, M. Vichi (Eds.), Statistical Models for Data Analysis (pp. 27–35). Springer International Publishing, 2013

Caballero, W., Giraldo, R., and Mateu, J. A universal kriging approach for spatial functional data. Stochastic Environmental Research and Risk Assessment 27, 1553–1563, 2013

Chiles, J. P., Delfiner, P. Geostatististics, Modelling Spatial Uncertainty 2nd Edition (2nd ed.). Wiley-Interscience,2012.

Clayton, V. D., Correcting for negative weights in ordinary kriging, Computers & Geosciences, 22:7, 765-773, 1996.

Cressie N, Wikle C.K. Statistics for Spatio-Temporal Data. Wiley \& Sons, Inc, 2011.

Giraldo, R., Delicado, P., Mateu, J. Ordinary kriging for function-valued spatial data. Environmental and Ecological Statistics, 18(3), 411–426, 2011.

Irpino A., Romano E. Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. Noirhomme-Fraiture M, Venturini G (eds) EGC Revue des Nouvelles Technologies de l'Information, vol RNTI-E-9, pp 99–110, 2007

Menafoglio, A., Secchi, P., Dalla Rosa, M. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electronic Journal of Statistics, 7, 2209–2240, 2013.

Rushendorff L. Wasserstein metric. In Encyclopedia of Mathematics. Springer, 2001.

++

10h15 – 10h45 ***COFFEE BREAK***

++

10h45 – 11h15 N. Niang[1], S. Bougeard[2], V. Cariou[3], G. Saporta[1]

[1] CNAM-CEDRIC, Paris, France
[2]Anses, Département d'Epidémiologie, Ploufragan, France
[3] StatSC, ONIRIS, INRA, Nantes, France

Prediction in clusterwise multiblock PLS

Abstract: Multiblock Partial Least Squares aims at exploring and modeling the relationships between several variables to be predicted from several other explanatory ones organized into meaningful blocks. As an extension of PLS regression it inherits its advantages and handles situations where the variable number is higher than the observation number as well as multicollinearity within explanatory blocks. However in many applications, observations do not come from a single homogeneous population but they rather correspond to an unknown grouping structure associated with different regression models. It follows that the estimation of a single set of regression coefficients for the whole dataset may mask the variables relationships and be misleading. Clusterwise PLS regression is proposed to overcome this drawback. Initially considered with unsupervised models (Diday 1974), such an approach consists herein of simultaneously look- ing for a partition of the observations into clusters and their associated PLS regression model by minimizing sum of squared errors computed over all the clusters. We present clusterwise multiblock PLS: an extension of clusterwise PLS regression to multiresponse variables and independent variables organized in blocks. This new method provides a partition combining the description of the independent variables with the prediction of the set of response variables. Each cluster of the partition is associated with its own multiblock PLS model (e.g., components, set of coefficients), which is then used to improve the overall _t of the prediction step. Thereafter, several rules are considered to assign a new observation to one of the obtained clusters. Finally, the prediction is performed using the local model of the assigned cluster (Bougeard et al. 2018). As our strategy is general and based on a clear criterion to minimize, the proposed approach can be directly extended to other multiblock regression methods. The properties of clusterwise multiblock PLS will be evaluated with a simulation study and will also be illustrated with a real data example.

References

Bougeard S., Cariou V, Saporta G., Niang N., Prediction for regularized clusterwise multiblock regression. Appl Stochastic Models Bus Ind. 34,6, 2018, pp 852-867.

Diday E., Introduction à l'analyse factorielle typologique. Revue de Statistique Appliquée, 22, 1974, pp.29-38.

11h15 – 11h45 H. Wang[1,2], S. Lua[1,3], J. Zhaoa[1,2]

[1] School of Economics and Management, Beihang University
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing
[3] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations

Aggregating multiple types of complex data in stock market prediction: A model-independent framework

Abstract: Recent years have witnessed the increasing volumes and types of data in the stock market. And being typical examples of complex data, these data are generally formatted as lists, symbolic data (Diday, 2016) or functional data, instead of singular values only. On one hand, these data provide unprecedented opportunities for understanding the stock market more comprehensively and makes price predictions more accurate than before (Zhou et al., 2018). On the other hand, however, these data also bring challenges to classic statistical approaches since existing models might be constrained to a certain type of data.

Aiming to aggregate data from different sources and to offer type-free capability to existing models, a framework for predicting the stock market in scenarios with mixed complex data, including scalar data, symbolic data (pie-like) and functional data (curve-like), is established as shown in Figure 1. The presented framework is model-independent because it serves as an interface to multiple types of data and can be combined with various prediction models. Moreover, the framework is proven to be effective through numerical simulations. For price prediction, we incorporate the trading volume (scalar data), intraday return series (functional data), and investors' emotions from social media (symbolic data) through the framework to competently forecast the market trend at opening on the next day. The strong explanatory power of the framework is further demonstrated. Specifically, the intra- day returns are found to impact the following opening prices differently between a bearish market and a bullish market. Additionally, it is not at the beginning of the bearish market but rather the subsequent period in which the investors' "fear" becomes indicative. This framework would help to easily extend existing prediction models to scenarios with multiple types of data and to provide a more systemic understanding of the stock market.

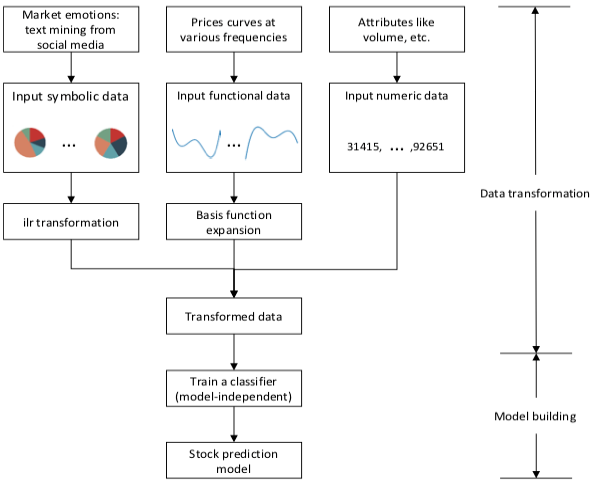


Figure 1 : The framework for stock prediction using mixed types of complex data.

References

E. Diday (2016). Thinking by classes in Data Science: the symbolic data analysis paradigm. WIREs Comput Stat, 8:172–205.

Z. Zhou, K. Xu and J. Zhao (2018). Tales of Emotion and Stock in China: Volatility, Causality and Prediction. World Wide Web, 21:1093–1116, 2018.

11h45 – 12h15 H. Wang[1,2], T. Huang[1,3], and S. Wang[1,3]

[1] School of Economics and Management, Beihang University
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing
[3] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations

Artificial neural network method incorporating spatial structure with its Application in Prediction of PM2.5

Abstract: Efficient statistical analysis with spatial correlated data is of great practical importance, as exemplified by applications in socio-economic and environment analysis. A common approach to accommodate the spatial dependence is to incorporate a spatial autoregressive parameter and a spatial weight matrix into the classical model to improve prediction accuracy (Anselin, 1988; Cressie 2015). For example, the spatial autoregressive model can help us understand the influence of spatial distance by considering the spatial correlation of dependent variables and error items. However, the traditional spatial autoregressive models, including both spatial lag model and spatial error model, are formulated in the framework of linearity, which will perform poorly if there is a non-linear relationship between dependent variables and independent variables in empirical research. Furthermore, this non-linear relationship occurs commonly in the age of big data.

Fortunately, various methods are proposed to deal with this issue. Among them, neural network algorithm is a machine learning method with the advantages of self-organization, self-adaptation, self-learning and strong ability of non-linear mapping (Wang and Wang, 2004). In this paper, an artificial neural network method incorporating spatial structure is proposed, which can improve the ability of the model to deal with non-linear problems. Moreover, the explanatory variables of the model include different types of data, including numerical data and symbolic data. Finally, the simulation results show that the fitting precision of the model can be significantly improved by considering the autocorrelation and non-linear characteristics of variables.

In the empirical study, we discussed the PM2.5 prediction of prefecture-level cities in China. PM2.5 concentration in a region is affected by multiple factors. In addition to the influence of economic development, industrial structure, population structure and urbanization construction in the region, the spread of air pollution between adjacent areas will also significantly affect the level of pollution(Wang et al. 2017). There are 24 economic indicators of 285 prefecture-level cities in China in 2015 used as explanatory variable in this study. Employment structure and value-added data of three industries are symbolic data, so the independent variables of the model include mixed variables, namely ordinary numeric-value data and symbolic data. After bringing these data and spatial weight matrix into operation, the neural network with spatial structure can obviously improve the prediction ability of the model.

References

L. Anselin. Spatial econometrics: methods and models. Vol. 4. Springer Science & Business Media, 2013.

N. A. C. Cressie. Statistics for Spatial Data. Statistics for spatial data. 1993.

Y. Wang, L. Wang. The Nonlinear Regression Based on BP Artificial Neural Network. Computer Engineering and Applications, 40.12(2004):79-82.

H. Wang, J. Gu, H. Wang and Z. Meng. The Study on Main Influence Factors of the Serious Atmosphere Pollution in Beijing-Tianjin-Hebei Region. Athematics in Practice and T

++

12h15 – 14h00 *LUNCH*

++

SESSION 3: Interval data

14h00 – 14h30 F. de A. T. de Carvalho[1], E. de A. Lima Neto[2], U. da Nobrega Rosendo[2]

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, Brazil
[2] Universidade Federal da Paraíba, Dep. de Estatística, Cidade Universitária, João Pessoa, Brazil

A new exponential-type kernel robust regression for interval-valued data

Abstract: Lima Neto and de Carvalho (2018) provided a robust regression methods for interval-valued variables that penalize the presence of outliers in the midpoints and/or in the ranges of interval-valued observations through the use of exponential-type kernel functions. In Lima Neto and de Carvalho (2018), the weight given to the midpoint and to the range of each interval-valued observation are updated at each iteration in order to optimize a suitable objective function. Moreover, the computation of the weight given to the midpoint (respectively, to the range) depends only on the mid-point (respectively, on the range) of the response and of the explanatory variables. It means that the mid- points (respectively, the range) outliers are penalized on the mid-point (respectively, the range) regression. This presentation provides a variant of the robust regression methods for interval-valued variables where the computation of the weight given to the midpoint or to the range depends on both the mid-point and the range of the response and of the explanatory variables. Therefore, in this variant the presence of outliers in the midpoints penalize both the mid-point and the range regressions. Moreover, the observations with outliers on both mid-point and range are more penalized than those observations with outliers only in the mid-point or only in the range. This presentation also provides the application of the proposed robust method, as well as the application of the previous robust methods of Domingues et al (2010), Fagundes et al (2010), Fagundes et al (2016), Lima Neto and de Carvalho (2018), on real valued data sets. Preliminary results are shown and commented.

References

Lima Neto, E. A., de Carvalho, F. A. T. (2018). An exponential-type kernel robust regression model for interval-valued variables. *Information Sciences 454*, 419–442.

Domingues, M. A. O., de Souza, R. M. C. R., Cysneiros, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters 31*, 1991–1996.

Fagundes, R. A. A., de Souza, R. M. C. R., Cysneiros, F. J. A. (2010). Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence 26*, 563–573.

Fagundes, R. A. A., de Souza, R. M. C. R., Soares, Y. M. G. (2016). Quantile regression of interval-valued data. In *Proceedings of the 23rd International Conference on Pattern Recognition (Cancun, Mexico)*, pp. 2586–2591, IEEE.

14h30 – 15h00 Y. Sun[1], X. Zhang[1] A.T.K. Wan[2], S. Wang[1]

[1] Chinese Academy of Sciences
[2] City University of Hong Kong

Model averaging for interval-valued data

Abstract: The technology progress has inspired the growing popularity in statistics of modelling new types of data, especially interval-valued time series (ITS) data. Recently, significant theoretical advances have been made in ITS analysis; however, the majority of the existing work has emphasized single model setups.

Frequentist model averaging (FMA) is an alternative to model selection, where estimates or predictions are obtained from competing models, weighted according to their reliability as determined by the data. A large part of the FMA literature is about finding optimal model weights oriented towards the achievement of some form of optimality (Hansen, 2007; Wan et al., 2010; Zhu et al., 2018).

While FMA has been extensively researched, all results are developed under point- valued data and no study of FMA has considered interval-valued or other forms of symbolic data. The purpose of this paper is to fill this void. Here, we consider a FMA procedure that combines models with different lag lengths and also allow the lag lengths to differ across the centre and range equations. In this framework, we develop a Mallows-type criterion for midpoints of ITS and a new weight choice criterion for ranges of ITS. We prove that the corresponding model averaging estimators for midpoints and ranges are asymptotically optimal under certain assumptions. We further extend our concern to the corrected ranges and establish associated theory. Numerical simulation and empirical applications highlight the merits of the proposed method.

References

Hansen, B. E. (2007). Least squares model averaging. Econometrica 75, 1175–1189.

Wan, A. T. K., Zhang, X. & Zou, G. (2010). Least squares model averaging by Mallows criterion. Journal of Econometrics 156, 277–283.

Zhu, R., Wan, A. T. K., Zhang, X. & Zou, G. (2018). A Mallows-type model averaging estimator for the varying-coefficient partially linear model. Journal of the American Statistical Association In Press. https://www.tandfonline.com/doi/abs/10.1080/ 01621459.2018.1456936.

15h00 – 15h30 D. Desbois

UMR Economie publique, Inra-AgroParisTech, Université Paris-Saclay, France

Exploring the distribution of conditional quantile estimate ranges: an application to the estimation of specific production costs of pig in the European Union

Abstract: This communication uses symbolic data analysis tools to visualize the distribution of conditional quantile estimate ranges, applying them to the problem of cost allocation in agriculture. After recalling the conceptual framework of the estimation of agricultural production costs, the first part presents the empirical model, the quantile regression approach and the interval data processing techniques used as symbolic data analysis tools. The second part presents the comparative analysis of the econometric results of the specific cost of pig production and the resulting gross margin estimates for twelve European Member States in order to discuss the relevance of the exploratory graphs obtained for international comparisons, using principal component analysis and hierarchical grouping of estimation interval distributions.

References

H. Bock, E. Diday. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin, Springer-Verlag, 2000.

A. C. Cameron, P. K. Trivedi. *Microeconometrics. Methods and Applications*. Cambridge, Cambridge University Press, 2005.

P. Cazes, A. Chouakria, E. Diday, Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de statistique appliquée*, 45(3): 5-24, 1997.

M. Chavent, H.H. Bock. *Clustering Methods for Symbolic Objects*. In Bock HH., Diday E. (eds) Analysis of Symbolic Data, pp. 294-341. Berlin, Springer, 2000.

D. Desbois. *Estimation des coûts de production agricoles : approches économétriques*. Thèse de doctorat ABIES-AgroParisTech, dirigée par J.C. Bureau et Y. Surry, 249 p., 2015.

J.F. Divay, F. Meunier. Deux méthodes de confection du tableau entrées-sorties. *Annales de l'INSEE*, 37:59-109, 1980.

X. He, F. Hu. Markov Chain Marginal Bootstrap, *Journal of the American Statistical Association*, 97:783-795, 2002.

R. Koenker, G. Bassett. Regression quantiles. *Econometrica*, 46 : 33-50, 1978.

R. Koenker, Q. Zhao. L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, 3: 223-235, 1994.

C.N. Lauro, F. Palumbo. Principal component analysis of interval data: a symbolic data analysis approach, *Computational Statistics*, 15(1): 73-87, 2000.

D. Desbois, J.-P.,Butault, Y. Surry. Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, *Economie rurale*, nº361, pp. 3-22, 2017.

++

15h30 – 16h00 *COFFEE BREAK*

++

SESSION 4: Bibliographic analysis and Economy

16h00 – 16h30 V. Batagelj[1,2,3,*] , D. Maltseva[3]

[1] IMFM Ljubljana
[2] UP IAM Koper
[3] NRU HSE, Moscow
*Contact author: vladimir.batagelj@fmf.uni-lj.si

Temporal bibliographic analysis

Abstract: In our paper Batagelj and Praprotnik (2016) we proposed temporal quantities (a kind of symbolic data) as a basis for longitudinal approach for temporal network analysis as an alternative to the traditional approach based on time slices (Batagelj et al., 2014). In this paper we show that methods for bibliometric network analysis from Batagelj and Cerinsek (2013) can be extended to temporal bibliographic networks based on temporal quantities. We first show how to convert an ordinary bibliographic network to its temporal counterpart – we can use either the instantaneous or the cumulative description. Using network normalization and network multiplication we can obtain different derived temporal networks providing us with interesting views on the evolution of the studied bibliography. The proposed approach is supported by a Python package for (temporal) network analysis Nets (Batagelj, 2017, 2018) and will be illustrated on some real-life bibliometric networks data sets.

References

Batagelj, V, Cerinšek, M (2013). On bibliographic networks. Scientometrics 96(3), 845-864.

Batagelj, V, Doreian, P, Ferligoj, A, Kejžar, N (2014). Understanding Large Temporal Networks and Spatial Networks: Exploration, Pattern Searching, Visualization and Network Evolution. Wiley.

Batagelj, V, Praprotnik, S (2016): An algebraic approach to temporal network analysis based on temporal quantities. Social Network Analysis and Mining, 6(1), 1-22.

Batagelj, V (2017): Nets – a Python package for network analysis. https://github.com/bavla/Nets

Batagelj, V (2018): Python Packages for Networks. In: Alhajj R., Rokne J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY

16h30 – 17h00 A. Srakar

Institute for Economic Research, Ljubljana & Faculty of Economics, Univ. Ljubljana, Slovenia

Symbolic input-output analysis: a harmonic analysis approach to combining statistical distributions

Abstract: We provide a new, stochastic, approach to study input-output analysis and calculation of multipliers. We apply the findings to the calculation of production and employment multipliers for selected European countries. Input-output (IO) analysis is, in principle, one of the most commonly used, but a non-stochastic approach to national accounts. Yet, it suffers from several common critiques, not least being fixed input structure in each industry; all products of an industry are identical or are made in fixed proportions to each other; and each industry exhibits constant returns to scale in production. To this end, we use symbolic data analysis (following e.g. Michalski, Diday and Stepp, 1981; Brito, 1995; 2000; Billard and Diday, 2000; 2002; 2006) to construct distributions (symbolic modal variables) in the cells of IO tables instead of previously used numerical aggregated values. Using such approach, we are able to include the stochastic component in the modelling with IO tables in a novel way. To combine the cells (i.e. combining/adding-multiplying distributions), we provide foundations of a symbolic Leontief distribution calculus, based on harmonic analysis (in particular, the concept of convolutions). We use concepts from the algebra of random variables (Springer, 1979) and four distributional operations: addition/convolution; distributional difference; product distribution; and ratio (including inverse) distribution. We derive new Leontief formulas for two broad cases: when the total output is fixed or is a distribution itself. Finally, the problem of inverting a large random matrix is solved using existing results from RMT (random matrix theory) and numerical methods. We are able to derive the confidence intervals of production and employment multipliers, calculated in a novel way and apply the methodology to derive the sectorial multipliers for the selected EU countries in the period 2008-2010 and study the performance of new method compared to "classically" used IO analysis to demonstrate the advantages of the new approach. Preliminary results confirm the validity of the approach and show important advantages of taking into account the stochastic component of the IO analysis in a manner as proposed in the paper. Scientific relevance of the paper is clear and large: a) Completely new way of approaching stochastic possibilities of input-output analysis; b) Significant gain in information, the gain in accuracy and predictability still under test; c) Foundation of a new calculus of distributions, to form also the foundation of the work in symbolic data analysis and the analysis of complex data (e.g. SDA, CoDA, FDA) in future.

References

Billard, L., Diday, E. (2000). Regression analysis for interval-valued data. In: *Proc. of IFCS'00*, Belgium, pp. 369-374, Berlin, New York: Springer.

Billard, L., Diday, E. (2002). Symbolic Regression Analysis. In: *Proc. IFCS'02*, Poland, pp. 281-288, Berlin, New York: Springer.

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

Brito, P. (1995). Symbolic Objects: Order Structure and Pyramidal Clustering. *Annals of Operations Research*, 55, 277–297.

Brito, P. (2000). Hierarchical and Pyramidal Clustering With Complete Symbolic Objects. In *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information From Complex Data*, eds. H.-H. Bock and E. Diday, Berlin: Springer-Verlag, pp. 312–324.

Michalski, R. S., Diday, E., Stepp, R. E. (1981). A Recent Advance in Data Analysis: Clustering Objects Into Classes Characterized by Conjunctive Concepts. In *Progress in Pattern Recognition*, eds. L. Kanal and A. Rosenfeld, Amsterdam: North-Holland, pp. 33–56.

Springer, M. D. (1979). *The Algebra of Random Variables*. New York: Wiley.

17h00 – 17h30 M. Febrissy and M. Nadif

LIPADE, Paris Descartes University, France

Co-clustering via Nonnegative Matrix Tri-Factorization: A comparative study

Abstract: Non-negative Matrix Tri-Factorization (NMTF) consists in approximating a matrix **X** with positive or null values by the product of three matrices. Dealing with $\mathbf{X} \approx \mathbf{ZSW}^{\mathbf{T}}$, NMTF provides a good framework for co-clustering; **Z** and **W** can be considered as indicators of classes in rows and columns respectively while S as a compact representation of X or then as an absorption matrix having fewer constraints than **Z** and **W**. In this talk, we propose a fair comparison between several NMTF algorithms embedding different constraints such orthogonality constraints, clusters indicators constraints or graph regularization to eventually classify items such as text documents arising from softly or hardly mixture with balanced or unbalanced collections.

References

Boutsidis, C., Gallopoulos, E.: Svd based initialization: A head start for nonnegative matrix factorization. Pattern Recognition 41(4), 1350-1362 (2008).

Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 126-135. ACM (2006).

Gu, Q., Zhou, J.: Co-clustering on manifolds. In: Proceedings of the 15th ACM SIGKDD. pp. 359-368 (2009).

Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788 (1999).

Long, B., Zhang, Z.M., Yu, P.S.: Co-clustering by block value decomposition. In: ACM SIGKDD. pp. 635-640 (2005).

Wang, H., Nie, F., Huang, H., Ding, C.: Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In: ICDM. pp. 774-783 (2011).

Wang, H., Nie, F., Huang, H., Makedon, F.: Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: IJCAI. vol. 22, p. 1553 (2011).

Wild, S., Curry, J., Dougherty, A.: Improving non-negative matrix factorizations through structured initialization. Pattern recognition 37(11), 2217-2232 (2004).

Yoo, J., Choi, S.: Orthogonal nonnegative matrix factorization: Multiplicative up-

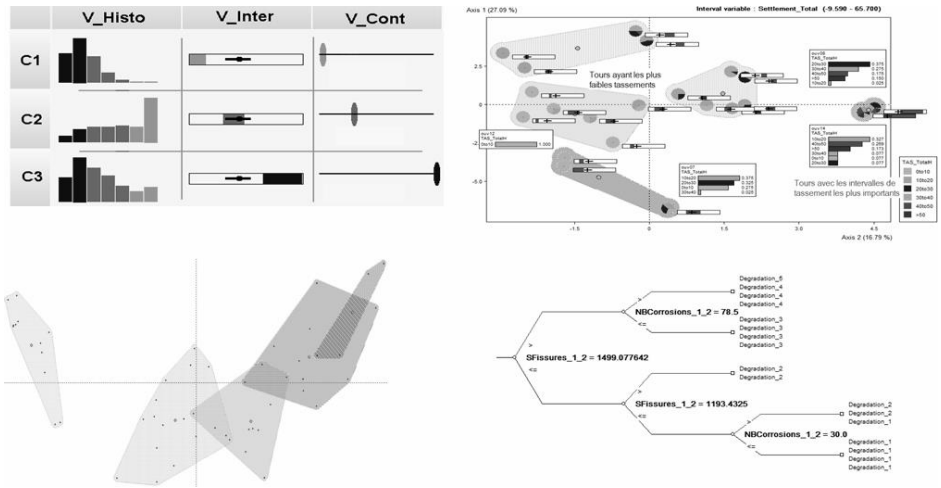
dates on stiefel manifolds. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 140-147 (2008).

SESSION 5: Software

17h30 – 18h00 F. Afonso
Symbad – Symbolic data Lab, Roissy, France

Software for Symbolic Data Analysis through Industrial Applications

Abstract: Three software solutions for Symbolic Data Analysis (SDA) are presented: SYR (www.symbad.co, Afonso et. al 2018), SODAS (www.info.fundp.ac.be/asso/sodaslink.htm, Diday & Noirhomme 2008) and RSDA (cran.r-project.org, Rodriguez 2013). These solutions enable the two stages of an automatic data processing by SDA: the first step involves the fusion and the reduction of the data into classes described by symbolic data. The tools are able to merge and aggregate data from multiple databases into a single data table of symbolic data. The data to be merged and reduced can be massive or not, heterogeneous (quantitative + qualitative + text + temporal data + etc.) and multi-source (one or more databases + web data + sensor readings + environmental data, etc.). The second step is then the analysis of the obtained symbolic data by specific advanced statistical methods (dissimilarities, clustering, decision trees, factorial analysis, regressions, etc.) of the reduced and merged data table.
The different tools and algorithms are illustrated through various applications in different sectors of activity. SDA is applied in epidemiology to measure the link between a disease and the characteristics of the patient and his environment; in civil engineering to detect defects on structures and discover the factors generating these defects; in cybersecurity to characterize the intrusions; in asset management to build up financial portfolios and discover abnormalities in fund management, in official statistics to compare socio-economic developments in different countries, etc.
Finally, the interaction between the different tools is discussed especially the ongoing developments of SYR with the projects R and Python.



Visualizations from the SYR software

References

Afonso, F., Diday, E., Toque, C. (2018). Data Science par Analyse des Données Symboliques. Technip. 448 pages, ISBN : 9782710811817.

Diday E., Noirhomme-Fraiture M. editors & co-autors (2008). Symbolic Data Analysis and the SODAS Software. Wiley. ISBN 978-0-470-01883-5.

Rodriguez O. (2013). Latest developments of the RSDA: An R package for Symbolic Data Analysis, conference of the International Federation of Classification Societies (IFCS), July 2013, At Tilburg, Volume: Volume 1 DOI: 10.13140/2.1.4539.5209

Friday, January 11th

SESSION 1: Clustering

9h00 – 9h30 Y. Lechevallier[1], F.A.T. De Carvalho[2]
[1] INRIA, France
[2] Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, Brazil

Weighted multi-view partitioning of time series

Abstract: Multi-View Clustering models can be viewed as a way to extract information from different data representations to improve the clustering accuracy.
My talk proposes a multi-view partitioning algorithm with automated computation of weights for both views and variables in such a way that the relevant views as well as the relevant variables in each view are selected for clustering. The aim is to obtain a collaborative role of the different tables in order to obtain a final consensus partition.
These methods are designed to build a partition and a prototype for each cluster as well as to learn a relevance weight for each tables by optimizing an adequacy criterion that measures the fitting between the clusters and their representatives.
Experiments with data sets: synthetic data from UCI machine learning repository and real data from Reunion Island (from December 2008 to March 2012, daily solar radiation measurements have been collected, every minute) show the usefulness of the proposed methods.

References

De Carvalho, F. A. T., De Melo, F. M. and Lechevallier, Y.(2015). A multi-view relational fuzzy c-medoid vectors clustering algorithm In Neurocomputing 163, 115–123.

De Araujo, R.C., De Carvalho, F. A.T. and Lechevallier, Y.(2017). Multi-View Hard C-Means with Automated Weighting of Views and Variables In ICJNN 2017,2792–2799.

Pedrycz, W. (2002). Collaborative fuzzy clustering. In Pattern Recognition Letters 23, 675–686.

Frigui, H., Hwanga, C. and Rhee, F.C.-H. (2007). Clustering and aggregation of relational data with applications to image database categorization. In Pattern Recognition 40, 3053–3068.

Jeanty P., Delsaut M., Trovalet L., Ralambondrainy H., Lan-Sun-Luk J.D., Bessafi M., Charton P. and Chabriat J.P.(2013). Clustering daily solar radiation from Reunion Island using data analysis methods. In Int. Conf. on Renewable Energies and Power Quality, Bilbao, Spain.

9h30 – 10h00 C. Biernacki[1], M. Marbac[2] and V. Vandewalle[3]

[1] Inria, University of Lille, CNRS (France)
[2] CREST and ENSAI (France)
[3] EA 2694 University of Lille, Inria (France)

Gaussian Based Visualization of Gaussian and Non-Gaussian Based Clustering

Abstract: A generic method is introduced to visualize in a "Gaussian-like way", and onto R2, results of Gaussian or non-Gaussian based clustering. The key point is to explicitly force a visualization based on a spherical Gaussian mixture to inherit from the within cluster overlap that is present in the initial clustering mixture. The result is a particularly user-friendly drawing of the clusters, providing any practitioner with an overview of the potentially complex clustering result. An entropic measure provides information about the quality of the drawn overlap compared to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the r package ClusVis.

References

Biernacki, C. (2017), Mixture models, in J.-J. Droesbeke, G. Saporta & C. Thomas-Agnan, eds, Choix de modèles et agrégation, Technip.

Bouveyron, C., Côme, E. & Jacques, J. (2015), «The discriminative functional mixture model for a Comparative analysis of bike sharing systems', Ann. Appl. Stat. 9(4), 1726–1760.

Daudin, J.-J., Picard, F. & Robin, S. (2008), A mixture model for random graphs, Stat. Comput. 18(2), 173–183.

Gollini, I. & Murphy, T. (2014), Mixture of latent trait analyzers for model-based clustering of categorical data, Statistics and Computing 24(4), 569–588.

Jacques, J. & Preda, C. (2014), Model-based clustering for multivariate functional data, Comput. Statist. Data Anal. 71, 92–106.

Kosmidis, I. & Karlis, D. (2015), Model-based clustering using copulas with applications, Statistics and Computing pp. 1–21.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. & Govaert, G. (2015), Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library, Journal of Statistical Software 67(6), 241–270.

Scrucca, L. (2010), Dimension reduction for model-based clustering, Statistics and Computing 20(4), 471–484.

10h00 – 10h30 P. Bertrand[1], J. Diatta[2]

[1] Université Paris-Dauphine, PSL Research University, Ceremade, Paris, France
[2] LIM-EA2525, Université de La Réunion, Saint-Denis, France

Multilevel clusterings as abstract convexities

Abstract: A multilevel clustering is a clustering system in which at least one cluster is included in another one. During the 1980's, multilevel clustering models were theoretically investigated with the aim to extend the well-known Benzécri-Johnson bijection. In this presentation, we present multilevel clustering models, characterized as collections of convex sets of some given interval operator satisfying a specific condition. We propose a way to build such multilevel clusterings from a dissimilarity, using interval operators that depend on both the dissimilarity and a parameter that tunes the covering rate of the generated clusters.

References

H.-J. Bandelt, A.W.M. Dress, An order theoretic framework for overlapping clustering, Discrete Math. 136 (1994) 21{37.
J.-P. Barthélemy, F. Brucker, Binary clustering, Discrete Appl. Math. 156 (2008), 1237-1250.
J.-P. Benzécri, L'Analyse des Données : la Taxinomie, Vol. 1, Dunod, Paris, 1973.
P. Bertrand, J. Diatta, Multilevel clustering models and interval convexities. Discrete Appl. Math. 222 (2017), 54-66.
J. Diatta, B. Fichet, Quasi-ultrametrics and their 2-Ball Hypergraphs, Discrete Math. 192 (1998) 87-102.
E. Diday, Orders and overlapping clusters in pyramids, in : J. De Leeuw, et al. (Eds.), Multidimensional Data Analysis, DSWO Press, Leiden, 1986, pp. 201-234.
M. Van de Vel, Theory of Convex Structures, first ed., North-Holland, 1993.

++

10h30 – 11h00 ***COFFEE BREAK***

++

11h00 – 11h30 L. Billard

University of Georgia

Clustering of Symbolic Data

Abstract: The concept of symbolic data originates in Diday (1987). We consider two aspects of cluster methodology. First, while there has been a lot of activity in using regression-based algorithms to partition a data set into clusters for classical data, no such algorithms have been developed for a set of interval-valued observations. A new algorithm is proposed based on the k-means algorithm of MacQueen (1967) and the dynamical partitioning method of Diday (1973) and Diday and Simon (1976), with the partitioning criteria being based on establishing regression models for each sub-cluster. Second, we extend the Kim (2009) and Brito and Chavent (2012) (both of which extended the Chavent, 1998, work on intervals) divisive clustering for histograms based on the data midpoints to a double divisive monothetic method based on both the histogram means and their variances; see Kim and Billard (2018).

References

Chavent, M. (2000). Criterion-based divisive clustering for symbolic data. In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, 299-311.

Brito, P. and Chavent, M. (2012). Divisive monothetic clustering for interval and histogram-valued data. In: Proceedings ICPRAM 2012-1st International Conference on Pattern Recognition Applications and Methods, Vilamoura, Portugal.

Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. International Journal of Computer and Information Sciences 2, 61-88.

Diday, E. (1987). Introduction _a l'approche symbolique en analyse des données. Premier Jounées Symbolique-Numérique, CEREMADE, Université Paris - Dauphine, 21-56.

Kim, J. (2009). Dissimilarity Measures for Histogram-valued Data and Divisive Clustering for Symbolic Objects. Doctoral Dissertation, University of Georgia.

Kim, J. and Billard, L. (2018). Double monothetic clustering for histogram-valued data. Communications for Statistical Applications and Methods 25, 263-274.

11h30 – 12h00 V. Cariou[1], T. F. Wilderjans[2]

[1] StatSC, ONIRIS, INRA, Nantes, France

[2] Methodology and Statistics Research Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University, The Netherlands

Extending the clustering of variables around latent components approach to three-way data by means of clusterwise Parafac

Abstract: In many research areas, such as sensometrics and chemometrics, the dataset at hand is a priori partitioned into meaningful blocks of variables measured on the same set of observations. Moreover, some situations also lead to consider the same variables measured on the same set of observations according to a third parameter (e.g. longitudinal) corresponding to a three-way three-mode data structure. In this scope, the Three-Way Cluster analysis around Latent Variables (CLV3W) approach (Wilderjans & Cariou, 2016; Cariou & Wilderjans, 2018) is proposed to extend the clustering around latent variables analysis (Vigneau & Qannari, 2003) to three-way data. The CLV3W method groups the 2nd mode elements – typically the set of variables – into Q clusters and estimates for each cluster an associated latent component such that the variables within each cluster are as much related (i.e., highest squared covariance) as possible with the latent component. Simultaneously, for each latent component separately, a system of weights is estimated that yields information regarding the 3rd mode. We show that CLV3W turns out to seek a clusterwise model (Diday, 1978) which corresponds to a clusterwise PARAFAC one applied on the 2nd mode of the three-way data array. This approach is illustrated on the basis of two datasets pertaining to Quantitative Descriptive Analysis and Consumer Studies either to detect sensory latent components or to perform consumer' segmentation.

References

Cariou, V., & Wilderjans, T. F. (2018). Consumer segmentation in multi-attribute product evaluation by means of non-negatively constrained CLV3W. *Food Quality and Preference*, *67*, 18-26.

Diday, E. (1978). Analyse canonique du point de vue de la classification automatique. Rapport de Recherche IRIA, n° 293.

Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, *32*(4), 1131-1150.

Wilderjans, T. F., & Cariou, V. (2016). CLV3W: A clustering around latent variables approach t detect panel disagreement in three-way conventional sensory profiling data. *Food quality and preference*, *47*, 45-53.

12h00 – 12h30 A. Gloaguen[1], C. Philippe[2], V. Frouin[3], G. Dehaene-Lambert[4], L. Le Brusquet[5], A. Tenenhaus[6]

[1] Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec, Université Paris-Saclay, France et UNATI, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France
[2, 3] UNATI, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France
[4] INSERM, UMR992, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, France
[5] Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec, Université Paris-Saclay, France
[6] Laboratoire des Signaux et Systèmes (L2S), CentraleSupelec, Université Paris-Saclay, et Brain and Spine Institute, Bioinformatics and Biostatistics platform, Paris France.

Regularized Generalized Canonical Correlation Analysis extended to multiway data with a medical application

Abstract: Regularized generalized canonical correlation analysis (RGCCA) [Tenenhaus & Tenenhaus, 2011, Tenenhaus et al., 2017] is a general framework that encompasses several important multivariate analysis methods such as principal component analysis, partial least squares regression and multiblock component methods as generalized canonical correlation analysis and consensus PCA. RGCCA is currently geared for the analysis two-way data matrices. It frequently occurs to encounter data combining multiway and multiblock structures and in this work, we propose Multiway RGCCA (MGCCA) an extension of RGCCA to the case where blocks have a tensor structure. This multiway structure is fully considered by adding appropriate Kronecker constraints on the RGCCA optimization problem. The usefulness of MGCCA is evaluated on a simulated data and on a cognitive study in human infants using high-density electro-encephalography (EEG).

References

Tenenhaus A., Tenenhaus M., (2011). Regularized Generalized Canonical Correlation Analysis, Psychometrika, 76 (2), 257-284.

Tenenhaus, M., Tenenhaus, A., Groenen, P. J. (2017). Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. Psychometrika, 82(3), 737-777.

++

12h30 – 14h00 *LUNCH*

++

SESSION 2: Basic formalism

14h00 – 14h30 L. Billard

University of Georgia

A Brief Overview of Symbolic Data in the Statistical Framework.

Abstract: The concept of symbolic data originates in Diday (1987). Symbolic data take various forms, but in general are hypercubes or Cartesian products of distributions in R^p, in contrast to classical data which are points in R^p. Since 1987, there has been a lot of work done, with a primary focus on methodology. We try to give a brief overview of some aspects of these accomplishments with a focus on how they fit into the statistical framework. More detailed overviews can be found in Noirhomme-Fraiture and Brito (2011) and Diday (2016), with non-technical introductions in Billard (2011, 2014). We close with some suggestions for where new approaches could be developed.

References

Billard, L. (2011). Brief overview of symbolic data and analytic issues. Statistical Analysis and Data Mining 4, 149-156.

Billard, L. (2014). The past's present is now. What will the present's future bring? In: Past, Present, and Future of Statistical Science (eds. X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott and J.-L. Wang). Chapman and Hall, New York, 323-334.

Diday, E. (1987). The symbolic approach in clustering and related methods of data analysis. In: Classification and Related Methods of Data Analysis (ed. H.-H. Bock). North-Holland, Amsterdam, 673-684.

Diday, E. (2016). Thinking by classes in data science: the symbolic data analysis paradigm. WIREs Computational Statistics 8, 172-205.

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

Le-Rademacher, J. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. Journal of Statistical Planning and Inference 141, 1593-1602.

Le-Rademacher, J. and Billard, L. (2013). Principal component histograms from interval-valued observations. Computational Statistics 28, 2117-2138.

14h30 – 15h0 R. Emilion

Denis Poisson Institute, Université d'Orléans, France

Likelihood in the symbolic context with examples

Abstract: Let X be a random variable describing a statistical population and C be a class random variable which assigns to each individual a class label. All the variability of X within a class having label c is expressed by the conditional distribution $P(X|C=c)$ of X given $(C=c)$. A symbol $S(c)$ of class label c is defined as a function of $P(X|C=c)$, shortly $S(c)=f(P(X|C=c))$ [3].
Since $P(X|C=c)$ can be estimated using an observed sample $(\big(x(1),c(1)\big),\dots,\big(x(k),c(k)\big))$ of the pair (X,C) such that $c(j)=c$, a symbol $S(c)$ can be estimated by a function of a sample of (X,C).
The variable S, called symbolic variable, is a random variable if a probability measure is defined on the space of class labels. The distribution of S can have a density w.r.t. a counting measure in the discrete case or a Lebesgue measure in the continuous case and such a density yields what is called a symbolic likelihood w.r.t. some parametric models.
The symbolic likelihood is specified in three interesting models: Latent Dirichlet Allocation in text mining [2], BLS model in big data paradigm [1], and Hierarchical Dirichlet Model in nonparametric Bayesian statistics [4].

References

[1] Beranger B., Lin H., Scott A. S., New models for symbolic data analysis, ArXiv e-prints (2018).
[2] Blei D.A., Ng A. Y., Jordan M. I., Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022.
[3] Emilion R., Diday E., Likelihood in the Symbolic context. Chapter in a book to appear. Eds. G. Saporta.
[4] Mueller, P., Quintana, F. and Rosner, G., A Method for Combining Inference over Related Nonparametric Bayesian Models. Journal of the Royal Statistical Society, Series B, 66 (2004): 735-749.

15h00 – 15h30 E. Diday

CEREMADE, Université Paris-Dauphine, France

Concordance and discordance between classes of complex data

Abstract: We say that the description of a class is much more explanatory when it is described by symbolic variables (closer from the natural language of the users), then by its usual analytical multidimensional description. The explanatory and characteristic power of classes can then be measured by criteria based on the symbolic data description of these classes and induce a way for comparing clustering methods by their explanatory power. These criteria are defined in a Symbolic Data Analysis framework for categorical variables, based on three random variables defined on the ground population. Tools are then given for ranking individuals, classes and their symbolic descriptive variables from the more towards the less characteristic. These characteristics are not only explanatory but can also express the concordance or the discordance of a class with the other classes. We suggest several directions of research mainly on parametric aspects of these criteria and on improving the explanatory power of Machine Learning tools. We finally present the conclusion and the wide domain of potential applications in socio demography, medicine, web security, etc.

References

Diday E. (2016) Thinking by classes in data science: symbolic data analysis. WIREs *Computational Statistics Symbolic Data Analysis Volume 8, September/October 2016. Wiley Periodicals, Inc. 191.*

Emilion R., Diday E. (2018). Symbolic Data Analysis Basic theory . Chapter in *Advances in Data Sciences,* eds. Saporta, Wang, Diday, Rong Guan, ISTE-Wiley.

++

15h30 – 16h00 ***COFFEE BREAK***

++

SESSION 3: Databases and Software

16h00 – 16h30 W. Litwin

Université Paris-Dauphine PSL, Paris, France

SQL for Stored and Inherited Relations

Abstract: A stored and inherited relation (SIR) is a stored relation (SR) extended with inherited attributes, (IAs), calculated as for a view. IAs can make queries to a SIR free of the logical navigation or of selected value expressions without affecting the normal form of the SR. A specific view of the SR can offer the same advantages. The so-called virtual (dynamic, computed…) attributes (columns), possibly extending SRs at some DBSs, can do as well for supported value expressions, being then less procedural to define than every alternate view. We propose extensions to SQL providing for the IAs always less procedurally defined than any alternate view. Likewise, every altering of a SIR leading otherwise to a view altering would be less procedural. Finally, to define IAs instead of virtual attributes would be at most as procedural. We motivate our proposals through the biblical Supplier-Part DB. We show how to implement SIRs with negligible operational overhead. We postulate SIRs standard on every popular SQL DBS.

References

SQL for Stored and Inherited Relations. Lamsade Technical Report, Mars 2016, updated Dec. 2018. http://www.lamsade.dauphine.fr/~litwin/Relations%20with%20Inherited%20Attributes%20Revisited.pdf

16h30 – 17h00 C. Biernacki [1], F. Afonso[2]

[1] Inria, University of Lille, CNRS, France
[2] Symbad – Symbolic data Lab, Roissy, France

Software presentation and discussion

Presentation 1: MASSICCC (Massive Clustering on Cloud Computing, https://massiccc.lille.inria.fr) is a demonstration platform giving access through a SaaS (service as a software) concept to data analysis libraries developed at Inria. It allows to obtain results either directly through a website specific display (specific and interactive visual outputs) or through an R data object download. Three packages have been integrated: Mixmod and MixtComp for clustering; BlockCluster for co-clustering. A particular attention has been put to provide also meaningful graphical outputs directly in the web platform itself has led to some specific developments.

Presentation 2: SYR (www.symbad.co, Afonso et. al 2018), and RSDA (cran.r-project.org, Rodriguez 2013). The different SDA tools and algorithms are illustrated through various applications in different sectors of activity. SDA is applied in epidemiology to measure the link between a disease and the characteristics of the patient and his environment; in civil engineering to detect defects on structures and discover the factors generating these defects; in cybersecurity to characterize the intrusions; in asset management to build up financial portfolios and discover abnormalities in fund management, in official statistics to compare socio-economic developments in different countries, etc. Finally, the interaction between the different tools is discussed especially the ongoing developments of SYR with the projects R and Python.

Discussion Possibly about different solutions for managing efficiently big and complex data (cloud computing, R, Python, others).

17h00 – 17h30 Round Table