Dauphine | PSL UNIVERSITÉ PARIS CEREMADE LAMSADE

Workshop
ADVANCES IN DATA SCIENCE FOR BIG AND COMPLEX DATA

UNIVERSITY PARIS-DAUPHINE, PSL
January 23-24, 2020

WORKSHOP DATA SCIENCE'2020 Paris Dauphine, PSL University January 23-24

Massive and complex data are taking an increasing lying role in human activity. In Symbolic Data Analysis, one looks at “complex objects” described by classes of different statistical units. For example, in Official Statistics, “regions” described by variables characteristic of classes of hospitals, schools, inhabitants, or in industry, nuclear power plant “towers” described by variables characteristic of classes of cracks, corrosions, positioning. In these cases, the famous $x \in \mathbb{R}^p$ does not apply because the p variables are unpaired (i.e. not defined on the same statistical units). To describe the internal variability of such classes, aggregated data in the form of explanatory "symbols" (intervals, distributions, lists, and so on) are used because description by conventional variables (only numerical or qualitative) does not allow this variability to be incorporated. The Symbolic Data Analysis (SDA) aims to extend the tools of Statistics and AI considered in the broadest sense (Villani Report) to this new type of data and is enjoying increasing success around the world.

Guest speakers will present: Likelihood-based modelling for symbolic data, Logistic regression models for aggregate data and its application to the improvement of standard estimate in the case of massive and complex data (S. Sisson), object oriented statistics and application in geostatistical kriging (A. Menafoglio), "Clusterwise" methods that transform the usual models into classes, making them more efficient (S. Bougeard), Co-clustering extended to symbolic data (R. Verde), hypothesis testing with interval-valued data (A. Roy), Modeling complex objects using dependent Dirichlet models (R. Emilion), Economic symbolic data applied on effects of retirement on health outcomes (A. Srakar), The ranking of complex objects and variables applied to the study of causes of death in European countries (S. Korenjak Cerne), Assessment of soil erosion by quantile estimates for European regions (D. Desbois), Hierarchical Clustering of Symbolic Objects Using Quantile (K. Umbleja), Improving the Relational Data Base model (W. Litwin), Analysis of improving TF-IDF and LDA (Latent Dirichlet Allocation) by the SDA framework (E. Diday), Symbolic Temporal Data: Multivariate Time Series Clustering (P. Brito), and so on…

Call for communications: Advances in this context or open problems posed by laboratories or industrialists are welcome.

Submission: abstract including references in one Word page before the 5/01/2020
Registration: with name, institution and position

Registration and submission at: datascience23242020@gmail.com

Web site: http://vladowiki.fmf.uni-lj.si/doku.php?id=sda:meet:pa20
Workshop Venue: University Paris-Dauphine, PSL. Place du Maréchal de Lattre de Tassigny 757016 PARIS. ROOM P302 the 23th, A711 the 24th.

Steering Committee

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

Scientific Committee

V. Batagelj	(University of Ljubljana, Slovenia)
L. Billard	(University of Georgia, USA)
S. Bougeard	(ANSES, French Agency for Food, France)
P. Brito	(University of Porto, Portugal)
P. Cazes	(CEREMADE, University Paris-Dauphine, PSL)
D. Colazzo	(LAMSADE, University Paris-Dauphine, PSL)
F. De Carvalho	(Cin-UFPE, Recife, Brazil)
D. Desbois	(INRAE-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, PSL)
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	(Ministère de l'Écologie et du Développement Durable - Paris la Défense, France)
R. Verde	(University of Campania "Luigi Vanvitelli", Caserta, Italy)
H. Wang	(School of Economics and Management, Beihang University, Beijing, China)

Workshop supported by:

Société Francophone de Classification

SFdS Société Française de Statistique

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

CATI CITISES

PROGRAMME

Thursday, January 23rd 2020, ROOM P302

8h45 – 9h00 **Welcome Session**

SESSION 1: ***Opening Session***

9h00 – 9h45 **Scott Sisson**
New Models for Distributional-Based Data

SESSION 2: ***Clustering***

9h45 – 10h15 **R. Verde**
Co-clustering Algorithms for Distributional Data Based on Adaptive-Distances

10h15 – 10h45 ***Coffee Break***

10h45 – 11h15 **Z. Wang**
Convex Clustering Method for Compositional Data

11h15 – 11h45 **Kadri Umbleja**
Hierarchical Clustering of Symbolic Objects Using Quantile

12h00 – 14h00 ***LUNCH***

SESSION 3: ***Ranking and Economy***

14h00 – 14h30 **M. Pelka**
Linear Ordering Method for Symbolic Data

14h30 – 15h00 **S. Korenjak Černe**
Ranking Complex Objects and Variables Applied to the Study of Mortality Data in EU Countries

15h00 – 15h30 **D. Desbois**
Assessment of Soil Erosion by Quantile Estimates for European Regions

15h30 – 16h00 ***Coffee Break***

16h00 – 16h30 **Andrej Srakar**
Program Evaluation and Causal Inference for Histogram Data: Estimation of the Effects of Retirement on Health Outcomes

SESSION 4: ***Anomaly Detection***

16h30 – 17h00 **U. Maurras**
Classification of Anomaly Detection Methods with a Focus on Isolation Forest

17h00 – 17h45 ***Round Table***

18h00 Welcome Buffet !

Friday, January 24th, ROOM A711

SESSION1: ***Clustering, Symbolic and Object Oriented Data***

9h00 – 9h30 **P. Brito, M.E. Silva, M. Dainovich**
Analysis of Symbolic Temporal Data: Multivariate Time Series Clustering

9h30 – 10h00 **A. Roy**
Analysis of Interval-Valued Data Using Patterned Covariance Structures

10h00 – 10h30 ***Coffee Break***

10h30 – 11h00 **F. Gioia**
The Fundamental Theorem of Asset Pricing Under Uncertainty

11h00 – 11h30 **A. Menafoglio**
Object Oriented Statistics and Application in Geostatistical Kriging

11h30 – 12h00 **S. Bougeard, N. Niang**
Clusterwise Regression for Multiblock High-Dimensional Data

12h00 – 14h00 ***LUNCH***

SESSION 2: ***Likelihood Regression Models, Unpaired variables, Symbolic extension of TF-IDF and LDA***

14h00 – 14h30 **Scott Sisson**
Composite Likelihood and Logistic Regression Models for Aggregated Data

14h30 – 15h00 **R. Emilion**
Modeling Complex Objects Using Dependent Dirichlet Models

15h00 – 15h30 **E. Diday**
Improving TF-IDF and LDA (Latent Dirichlet Allocation) by the SDA Framework

15h30 – 16h00 ***Coffee Break***

SESSION 3: ***Data Bases and Software***

16h00 – 16h30 **W. Litwin**
SQL for Stored and Inherited Relations
Manifesto for Improved Foundations of Relational Model

16h30 – 17h00 **Round Table and Conclusion**

ABSTRACTS

SESSION 1: ***Opening Session***

Thursday January 23rd, 9h00 – 9h45 – Scott Sisson

New Models for Distributional-Based Data

Scott A. Sisson, Boris Beranger and Jaslene Lin

School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

There has been much recent interest in developing statistical methods that can handle large-scale and complex data. One such approach is based on the idea of reducing the data to a smaller number of summary distributions – such as random histograms, or random intervals – that describe where the data generally reside, at the loss of some information about where each data point is precisely located. These distributions are then used as summary statistics in a standard analysis, with the benefit of large computational savings.

However there has been limited work in developing likelihood-based inference for random intervals and random histograms. This talk will outline a new approach of constructing likelihood functions for symbolic data (Beranger et al., 2018).

The method involves first constructing a standard statistical model for the full dataset. This is typically easy to do. Given this model, and the known form of the distributional summary (random histogram or random rectangle), we may then derive the likelihood function of the observed distributional data.. We will derive several forms of likelihood for (non-parametric) distributional-valued data, and illustrate its inferential performance in several worked analyses.

We will also specify alternative ways of constructing random rectangles (in d-dimensions) that contain more information than just using marginal quantiles. In this manner, we are able to estimate dependence parameters (e.g. correlations) between two margins given only the information within a random rectangle.

As a result we can conclude that:

- It is easy to specify statistical models for the underlying micro-data based on the observed distributional-based data;
- There is no requirement to assume micro-data uniformity within intervals (compared to many existing methods);
- There is value in constructing and analysing multivariate symbols;
- There is value in considering alternative mechanisms for constructing random rectangles.
- More consideration should be given to how to construct even univariate random intervals, if the goal is parametric inference of the micro-data model.

References

- Beranger B., H. Lin and S. A. Sisson (2018). New models for symbolic data analysis. https://arxiv.org/abs/1809.03659

SESSION 2: ***Clustering***

Thursday January 23rd, 9h45 – 10h15 – R. Verde

Co-clustering Algorithms for Distributional Data Based on Adaptive-Distances

Francisco De Carvalho[a], Rosanna Verde[b], Antonio Balzanella[b], Antonio Irpino[b]

[a]Centro de Informatica, Universidade Federal de Pernambuco, Recife, Brazil

[b]University of Campania "Luigi Vanvitelli", Dept. of Mathematics and Physics, Caserta, Italy

Our proposal deals with co-clustering algorithms of distribution-valued data. We aim of partitioning of the rows and the columns of an input data table, whose elements are aggregated data, represented by distributions, or histograms. The proposed procedure is based on an extension of the double *K-means* algorithm to distributional data (DDK). Due to the nature of the data, a suitable metric to compare distributions is used: the L_2 Wasserstein distance. In order to point out the different role of the variables in the analysis, related to their relevance in the characterizing the clusters, we introduce a system of relevance weights. That is performed by introducing adaptive distances in the algorithm which is so denominated Adaptive Distributional Double K-means (ADDK). The strategy achieves a co-clustering of the elements and of the variables, and achieves suitable weights for the variable, simultaneously. Those are automatically computed in an additional step of the optimization process.

Especially, we propose four algorithms in order to provide: i.) a system of weights for the variables, ii.) different systems of weights for the variables, one for each cluster (cluster-wise) and, according to a decomposition of the L_2 Wasserstein distance, iii.) a double system of weights on two distributions' components, for the variables and iv.) for the variables, for each cluster (cluster-wise).

Applications on simulated and real data sets show the effectiveness of the proposed algorithms and the relevance of some variables according to the structure of the datasets.

References

- E. Diday and G. Govaert (1977) Classification automatique avec distances adaptatives. R.A.I.R.O. Informatique Computer Science, 11(4):329–349
- G. Govaert and M. Nadif (2015). Co-Clustering: Models, Algorithms and Applications. Wiley, New York.
- A. Irpino, R. Verde, and F. A. T. De Carvalho (2014). Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. Expert Systems with Applications, 41(7):3351 – 3366.
- A. Irpino and R. Verde (2015). Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification, 9(2):143–175, 2015. ISSN 1862-5347

Thursday January 23rd, 10h45 – 11h15 – Z. Wang

Convex Clustering Method for Compositional Data

Xiaokang Wang, Huiwen Wang and Zhichao Wang

School of Economics and Management, Beihang University, Beijing, China

Compositional data, as a type of symbolic data that expresses the structural information of a whole, comprise of parts that are positive and subject to a constant-sum constraint (Pawlowsky-Glahn et al., 2015). Traditional statistical methods are not applicable to compositional data since the parts are correlated, and particularly, the subcompositional dominance property should be satisfied. If one interprets or analyzes compositions in raw form as real numbers, it can lead to misinterpretations and false conclusions.

Clustering analysis is one of the most popular unsupervised learning approaches to studying the grouped structure of data. Many clustering techniques are by definition NP hard problems to solve and the solution can be only approximated with heuristic measures. These methods such as the k-means clustering often converge to a local minimum of the criterion function with an improper choice of the initial points. Recently, Hocking et al. (2011) proposed the convex clustering approach that incorporates a regularization term to leverage the group sparsity of the clustering problem. The additional regularization term ensures a global optimal solution with the convex formulation, which can be efficiently solved with the ADMM (alternating direction method of multipliers) method (Chi and Lange, 2015).

In this paper, we develop a convex clustering method for grouping compositional data, which provides a global optimal solution given the convex relaxations of k-means and hierarchical clustering. We apply the ilr (isometric logratio) transformation to represent compositional data as orthonormal coordinates with respect to the Aitchison geometry. It is then shown that the convex clustering algorithm on the ilr-transformed data performs much more accurately than the direct clustering on the untransformed compositional data. The algorithm is further tested on a real-world dataset to illustrate the interpretability of the results from convex clustering method on compositional data.

References

- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.
- Hocking, T.D., Joulin, A., Bach, F., Vert, J.P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. http://ai.stanford.edu/~ajoulin/article/419_icmlpaper.pdf
- Chi, E.C., Lange, K. (2015). Splitting methods for convex clustering. Journal of Computational and Graphical Statistics, 24(4), 994-1013.

Thursday January 23rd, 11h15 – 11 h 45 – K. Umbleja

Hierarchical Clustering of Symbolic Objects Using Quantile

Kadri Umbleja

Tallinn University of Technology, Estonia

Histograms are the most general case of symbolic data. As complex distributions, they contain a lot of details and information that is vital when comparing different objects. Clustering is one of the vital data mining tasks. There have been many attempts to find suitable dissimilarity measure for histograms. A thorough survey of symbolic data clustering methods has been compiled by de Carvalho and de Souza. Generalized Minkowski metrics for mixed feature types based on the Cartesian system model has been defined in Ichino and Yaguchi. Different dissimilarity measures for symbolic data clustering are covered by Billard and Diday. SODAS software project produced much research into different aspects of symbolic data clustering like. Most of those methods are developed from similar approach to intervals and do not optimally consider the complex distributions like histograms. Therefore, a (dis)similarity measure using quantile values for histogram is proposed.

Idea of using quantiles is not totally new. Irpino and Verde proposed clustering approach based on Wasserstein-Kantorovich metric that also used quantile functions. Currently proposed method overcomes the drawback of complex histogram merger operation. Histograms are transformed into "quantile rectangles". At every chosen quantile point, for every feature, values at selected point are stored as intervals, becoming multidimensional "rectangles". The merger in that case is simple list merging operation that is computationally much cheaper than other type of histogram merger operations that additionally tend to generate new bins. Furthermore, only minimum and maximum value at every quantile rectangle and feature is actually needed.

Dissimilarity is found between two histograms, over every feature and every pre-selected quantile by finding the average size of quantile rectangle is those two histograms are to be merged, reflecting how similar they would be if they would form a cluster. Choice of quantiles can be used to guide the clustering – using truncated data, lowers the impact of outliers. Having more quantiles, makes the comparison more detailed but lowers the impact of difference along any one single quantile. Recommended number of quantiles is between 3 to 11.

The proposed method benefits from using computationally easy operations, no additional space requirements, comparing dissimilarity along different point of distribution and offers possibility of guiding the clustering by choice of quantiles. This kind of approach allows us to follow small microscopic details in distribution and form cluster according to those small similarities than are mostly overlooked in other approaches. In addition, the proposed method has monotone property.

References

- Billard L, Diday E (2006) Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley & Sons
- Brito P, De Carvalho FdA (2008) Hierarchical And Pyramidal Clustering. Symbolic Data Analysis and the Sodas Software pp 157-180
- De Carvalho FdA, Lechevallier Y, Verde R (2008) Clustering Methods In Symbolic Data Analysis. Symbolic Data Analysis and the Sodas Software pp 181-204
- de Carvalho FdA, de Souza RM (2010) Unsupervised Pattern Recognition Models For Mixed Feature-Type Symbolic Data. Pattern Recognition Letters 31(5):430-443
- Ichino M, Yaguchi H (1994) Generalized Minkowski Metrics For Mixed Feature-Type Data Analysis. IEEE Transactions on Systems, Man, and Cybernetics 24(4):698-708
- Irpino A, Verde R (2006) A New Wasserstein Based Distance For The Hierarchical Clustering Of Histogram Symbolic Data. In: Data science and classification, Springer, pp 185-192

SESSION 3: ***Ranking and Economy***

Thursday January 23rd, 14h00 – 14h30 – M. Pelka

Linear Ordering Method for Symbolic Data

Marcin Pełka

Wrocław University of Business and Economics, Department of Economics and Finance

In classical data situation objects are described by single-valued (numerical or categorical) variables. This allows to represent each object as a vector of qualitative or quantitative measurements where each column represents a single variable. Such data representation is too restrictive to represent more complex data. If we want to take into consideration the uncertainty and variability of the data we must assume sets of categories or intervals (with frequencies or weights in some cases). Such kind of data representation has been studied in the Symbolic Data Analysis (SDA). SDA provides different methods to deal with such different data types.

Like in the classical data analysis, in the symbolic data analysis we sometimes deal with a complex phenomenon that can't be measured directly (e.g. development of a country, green-growth level in a country, social cohesion of regions, usefulness of a tank in a MMO game). Such complex phenomenon is usually described by a various set of variables (classical, symbolic or both).

To analyze which object (that is a part of that complex phenomenon) is "doing better than others" concerning that complex situation a linear ordering based on the concept of the pattern of development can be used. The idea of the pattern of development was proposed by Hellwig during UNESCO conference in Warsaw in 1967 [Hellwig 1967; Hellwig 1972].

Besides that a visualization method, that utilizes multidimensional scaling, can be applied. In the first step the objects of interest undergo MDS, as a result of which they can be visualised in a two-dimensional space. In the second step the objects are linearly ordered to produce a ranking. A description on the procedure can be found in [Walesiak, 2016; Walesiak and Dehnel, 2018]. This two-step hybrid approach can be also applied to different types of symbolic data and the proposed article enhances this approach by using different symbolic distance measures to take into account different points of view.

References

- Hellwig Z., (1967). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, COM/WS/91, Warsaw, 9 December, 1967 (unpublished UNESCO working paper).
- Hellwig Z., (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, [in:] Gostkowski Z. (ed.), Towards a system of Human Resources Indicators for Less Developed Countries, Papers Prepared for UNESCO Research Project, Ossolineum, The Polish Academy of Sciences Press, Wrocław, pp. 115–134.
- Walesiak, M. (2016). Visualization of linear ordering results for metric data with the application of multidimensional scaling. Ekonometria, 2(52), pp. 9–21.
- Walesiak, M., & Dehnel, G. (2018). Evaluation of Economic Efficiency of Small Manufacturing Enterprises in Districts of Wielkopolska Province Using Interval-Valued Symbolic Data and the Hybrid Approach. In: Papież, M. and Śmiech, S. (Eds.), The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings, Foundation of the Cracow University of Economics, Cracow, pp. 563–572.

Thursday January 23rd, 14h30 – 15h00 – S. Korenjak Černe

Ranking Complex Objects and Variables Applied to the Study of Mortality Data in EU Countries

Simona Korenjak-Černe[a], Edwin Diday[b]

[a] University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia
[b]CEREMADE, University Paris-Dauphine | PSL, Paris, France

Cross-country comparisons of age-gender-cause specific mortality reflect health and demographic policies of the countries. Traditional indicator of country's health status is life expectancy at birth but this indicator captures only overall mortality rate without specification about death cause(s). Since it is well know that some health problems can be well controlled with the appropriate prevention programs and with healthy lifestyle (e.g., circulatory problems), it seems much more informative to include in the study also age-gender-cause specifics (Lotrič Dolinar et al., 2019). This also means that we are dealing with more complex data which include more information. The analysis of such data requires advanced data analysis methods. One of such possibilities if offered in symbolic data analysis (SDA) that we used in our study.

In our study we focus on the age-gender-cause specific mortality data of 28 EU countries for the year 2015 obtained from the EUROSTAT. We use symbolic data analysis methods and tools implemented in SYR software (Afonso et al., 2019). We present data and results with a symbolic data table, where each country (or group of countries) is presented with 72 symbolic variables: 36 bar-charts representing distribution of deaths over 4 death causes (neoplasms, circulatory, respiratory, other) by each age-gender combination, i.e., 18 age-groups x 2 gender); and 36 histograms, one for each age-gender combination (where cut-points differ since they are determined with a specific discretization method based on Fisher's algorithm extended to symbolic data).

Based on such symbolic representation we made further analysis, such as positioning countries on the main component plane obtained with PCA for symbolic data (Diday, 2013) and identifying groups of similar countries with k-means method adopted for symbolic data. We further study methods for automatic detection of the most discordant and most concordant countries and groups of countries with novel methods especially developed for complex data (Diday, 2019). Such ranking is based on the idea that a country c is discordant for a category x if the frequency $f_c(x)$ of this category is high for the country c and the proportion of countries c' having a frequency $f_{c'}(x)$ close from $f_c(x)$ is low.

References

- F. Afonso, A. Lotrič Dolinar, S. Korenjak-Černe, E. Diday (2019). Symbolic data analysis of Gender-Age-Cause Specific Mortality in European Countries (with SYR software). The 16th Conference of the IFCS, August 2019, Thessaloniki, Greece.
- E. Diday (2013). Principal Component Analysis for Bar Charts and Metabins Tables. Stat. Anal. Data Min. 6(5), Wiley, 403-430.
- E. Diday (2019). Placing in Order Classes and Units Described by Complex Data in the Symbolic Data Analysis Framework and Improving Explanatory Power of Machine Learning. Chapter in Advances in Data Science. Eds. E. Diday, R. Guan, G. Saporta, H. Wang. ISTE-Wiley.
- Lotrič Dolinar, J. Sambt, S. Korenjak-Černe (2019). Clustering EU Countries by Causes of Death. Popul. Res. Policy Rev. 38, Springer, 157-172.

Thursday January 23rd, 15h00 – 15h30 – D. Desbois

Towards the Cost Assessment of Soil Erosion: Displaying Quantile Estimates of Fertilizer Costs for European Regions

Dominique Desbois

UMR Economie Publique, INRAE-AgroParisTech, Université Paris-Saclay

The decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation but also based on the performances of the productive systems used by farm holdings, assessing their environmental impacts through the cost of specific resources used. This communication uses the symbolic data analysis tools in order to analyse the conditional quantile estimates of the fertilizer costs of specific productions in agriculture, as a replacement proxy for internal soil erosion costs. After recalling the conceptual framework of the estimation of agricultural production costs, we present the empirical data model, the quantile regression approach and the interval data techniques used as symbolic data analysis tools, mainly symbolic principal component analysis and symbolic clustering of the estimation intervals. The comparative analysis of econometric results for main products between European regions illustrates the relevance of the displays obtained for inter-regional comparisons based on specific productivity.

References

- Afonso F., Diday E. and Toque C. (2018) Data science par analyse des données symboliques, Technip, Paris, 444 p.
- Billard L., Diday E. (2006) Symbolic Data Analysis: Conceptual Statistics and Data Mining, 321 p.
- Cazes P., Chouakria A., Diday E., Schektman Y. (1997) Extensions de l'analyse en composantes principales à des données de type intervalle. Revue de Statistique Appliquée, n°24, pp. 5-24.
- Chavent M., Lechevalier Y., Briant O. (2007) DIVCLUS-T: A monothetic divisive hierarchical clustering method. Computational Statistics & Data Analysis, 52, 2, 687-701.
- Desbois D. (2015) Estimation des coûts de production agricoles : approches économétriques. PhD dissertation directed by J.C. Bureau and Y. Surry, ABIES-AgroParisTech, Paris, 2015.
- Desbois D., Butault J.-P., Surry Y. (2013) Estimation des coûts de production en phytosanitaires pour les grandes cultures. Une approche par la régression quantile, Economie Rurale, n° 333. pp.27 49.
- Desbois, D., Butault J.-P. and Surry Y. (2017). Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile, Économie rurale, 361, 3-22.
- Koenker R. and Bassett G. (1978) Regression quantiles. Econometrica, 46, 33-50, 1978.
- Lauro C.N. and Palumbo F. (2000) Principal component analysis of interval data: a symbolic data analysis approach. Computational Statistics, 15, 1, 73-87.

Thursday January 23rd, 16h00 – 16h30 – A. Srakar

Program Evaluation and Causal Inference for Histogram Data: Estimation of the Effects of Retirement on Health Outcomes

Andrej Srakar[a,b], Valentina Prevolnik Rupel[a], Tjaša Bartolj[a]

[a]Institute for Economic Research (IER), Ljubljana
[b]School of Economics and Business, University of Ljubljana

Statistical analysis of complex, i.e. non-standard data is gaining ground. Analysis of compositions, intervals, histograms, distributions, functions and manifolds has become more and more common in contemporary statistics and econometrics. Despite several types of regressions existing for symbolic data, causal inference in contemporary sense has not been studied so far. Furthermore, only slowly is it gaining ground using functional data. We develop statistical theory for using instrumental variables with symbolic histogram data, related to the three existing regression models: Billard and Diday (2006), Dias and Brito (2011) and Irpino and Verde (2012). We show that causal inference in all three models can be transformed into 2SLS estimation for quantile functions and derive the explicit forms of the estimators and their asymptotic behavior. We demonstrate their performance in Monte Carlo simulation study, comparing them to "regular" histogram regression estimators and functional linear regression with instrumental variables (Florens and Van Bellegem, 2015; Benatia, Carrasco, and Florens, 2017). We apply the findings to a pressing problem in the analysis of the aging process: the effects of retirement on health outcomes. Some authors conclude that retirement may lead to significant health improvements, but other studies find negative retirement effects. We use a panel dataset of Survey of Health, Ageing and Retirement in Europe (SHARE) in Waves 1-6. To address reverse causality in the relationship of retirement and health behaviours we use several different quantile-based instruments. A novelty in the approach is that we treat countries as units and the variables are aggregated over countries. In this manner, we estimate the effect of the empirical distribution of retirement across countries on empirical distribution of health outcomes over countries (the instrumental variable is distributional as well). We are, therefore, able to estimate the causal effect for a group of countries (most commonly, exogenous change is used only to estimate the effect on one treated population). As program evaluation and causal inference has so far not been studied with empirical distributional data (and very seldom with functional data) the article is a significant step ahead in regression analysis for symbolic data.

References

- D. Benatia, M. Carrasco and J.-P. Florens (2017). Functional linear regression with functional response. Journal of Econometrics, 201(2):269-291.
- L. Billard and E. Diday (2006). Symbolic data analysis: conceptual statistics and data mining. Chichester: Wiley and Sons.
- S. Dias and P. Brito (2011). A new linear regression model for histogram-valued variables. In 58th ISI World Statistics Congress, Dublin, Ireland.
- Irpino and R. Verde (2012). Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. arXiv:1202.1436v2
- J.-P. Florens and S. Van Bellegem (2015). Instrumental variable estimation in functional linear models. Journal of Econometrics, 186(2):465-476, 2015.

SESSION 4: ***Anomaly Detection***

Thursday January 23rd, 16h30 – 17h – U. Maurras

Classification of Anomaly Detection Methods with a Focus on Isolation Forest

Maurras U. Togbe[a], Yousra Chabchoub[a], Aliou Boly[b], Raja Chiky[a]

[a] Institut Supérieur d'Electronique de Paris (ISEP), [b] Université Cheikh Anta Diop de Dakar (UCAD)

Anomaly detection is an important issue in many application domains such as transport, health, finance, etc. In fact, a fast anomalies detection can avoid huge economic losses, natural disasters and even save human lives. Many anomalies detection algorithms are proposed in the literature (Chandola et al. [2009]). They generally consider anomalies as observations with a notable deviation from the normal behavior given by most of the observation. Therefore, they first model the normal behavior then they identify the anomaly, using different approaches relative to datamining and machine learning domains.

Anomalies detection has different constraints depending on the considered application. In a data stream context, as an example, a real-time detection is required. So, the considered algorithm must have a short response time and a low complexity to be faster than the data arrival rate, which is constantly increasing nowadays. In sensor networks domain, the detection is sometimes performed in devices with very limited resources implying other constraints for the considered algorithm.

We present in this work a complete state of the art on anomaly detection algorithms. We propose a classification of these methods based on the type of data sets (data streams, time series, graphs, etc.), the application domain and the considered approach (statistics, clustering, nearest neighbors, etc.) (see Figure 1). Then we focus on one of the most recent and most effective anomaly detection methods called Isolation Forest (Liu et al. [2008]). We discuss the advantages and the limits of this algorithm.

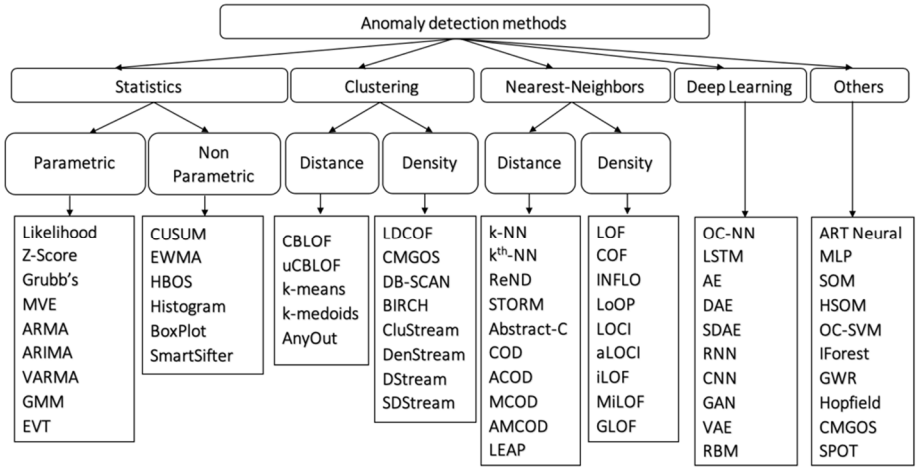


Figure 1. Anomaly detection methods classification.

References

- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3):15, 2009. https://dl.acm.org/doi/10.1145/1541880.1541882
- F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, pages 413–422. IEEE, 2008. https://ieeexplore.ieee.org/document/4781136

SESSION 1: ***Clustering, Symbolic and Object Oriented Data***

Friday, January 24th, 9h00 – 9h30 – P. Brito

Analysis of Symbolic Temporal Data: Multivariate Time Series Clustering

P. Brito[1,2] , M.E. Silva[1,3] , M. Dainovich[1]

[1]FEP, Univ. Porto, Portugal ; [2]LIAAD INESC-TEC, Portugal ; [3]CIDMA, Portugal

Symbolic Data Analysis provides a framework where variability within observations may explicitly be taken into account in the data representation and analysis. Symbolic data often arise from the aggregation of individual records, either registered at a lower granularity level, or gathered for each unit at different moments in time. When data is recorded at different time moments for the same entities, we obtain a time series for each unit and variable.

Time series clustering is an important and dynamic area of research, with application in a wide range of fields. Time series clustering analysis may be achieved by parametric and non-parametric methods. In this work we investigate three different non-parametric approaches for clustering multivariate time series. In a first approach, time series are directly compared using appropriate distance measures, which are then combined, to allow for distance-based clustering. Alternatively, we represent each time series by a set of features, thus transforming the original multivariate time series data set into one cross-sectional data array, to which classical methodologies are then applied. Finally, we interpret the multivariate time series as three-way data structures, and apply a clustering method designed for such three-way data.

The different considered approaches are applied to a multivariate time series comprising the observed production of seven cultures in 94 mesoregions of Brazil, for a period of 43 years. The results show that the three-way data approach is as efficient as clustering methods specific for time series data.

References

- Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: A literature survey. IEEE transactions on knowledge and data engineering, 21(1), 6–20.
- Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering: A decade review. Information Systems, 53, 16–38.
- Hyndman, R. J., Wang, E., & Laptev, N. (2015). Large-scale unusual time series detection. In 2015 IEEE international conference on data mining workshop (pp. 1616–1619).
- Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting, 33(2), 345–358.
- Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857–1874.
- Montero, P., Vilar, J. A., et al. (2014). Tsclust: An R package for time series clustering. Journal of Statistical Software, 62(1), 1–43.
- Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in R using the dtwclust package. Vienna: R Development Core Team.
- Schoonees, P., Groenen, P., & van de Velden, M. (2015). Least-squares bilinear clustering of three-way data (Tech. Rep.).

Friday, January 24th, 9h30 – 10h00 – A. Roy

Analysis of Interval-Valued Data Using Patterned Covariance Structures

Anuradha Roy

The University of Texas at San Antonio, San Antonio, Texas, USA

About the interval-valued data Billard and Diday (2006) once said, "It is the presence of this internal variation which necessitates the need for new techniques for analysis which in general will differ from those for classical data". We take care of this internal variation of the interval-valued data by considering the interval-valued data as two repeated measurements at the lower and upper bounds of an interval, and use a block compound symmetry (BCS) covariance structure $\boldsymbol{\Gamma}_y^{(2)}$ to model the data.

$$\boldsymbol{\Gamma}_y^{(2)} = \begin{bmatrix} \boldsymbol{U}_0 & \boldsymbol{U}_1 \\ \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{bmatrix}$$

We develop a new method to derive principal components (PCs) of interval-valued data, and consider the Fruit Juice data from Giordani and Kiers (2006, Table 4) to show the performance of our new method. This interval-valued data describing 16 fruit juices evaluated by a group of judges on six features, namely, Appearance, Smell, Taste, Naturalness, Sweetness and Density. More specifically, there are eight fruit juices and two brands for each juice. We represent each fruit juice as 12 x 1 dimensional vector $\boldsymbol{y}$ by grouping together first the six lower bounds of the intervals and then the six upper bounds of the intervals for each brand and model the data by using a (12 x 12) BCS structure $\boldsymbol{\Gamma}_y^{(2)}$. The (6 x 6) diagonal blocks $\boldsymbol{U}_0$ represent the variance-covariance matrix of the six features at the lower as well as at the upper bound of the intervals, whereas the (6 x 6) off-diagonal blocks $\boldsymbol{U}_1$ represent the covariance matrix of the six features between the lower and the upper bounds of the intervals.

Giordani and Kiers (2006) did not use the brand information in deriving the PCs in their paper. We use the brand information by combining the two brands together for each fruit juice and represent each fruit juice as 24 x 1 dimensional vector $\boldsymbol{y}$ and then use a (24 x 24) doubly BCS covariance structure $\boldsymbol{\Gamma}_y^{(3)}$ to model the data.

$$\boldsymbol{\Gamma}_y^{(3)} = \left[\begin{array}{cc|cc} \boldsymbol{U}_0 & \boldsymbol{U}_1 & \boldsymbol{W} & \boldsymbol{W} \\ \boldsymbol{U}_1 & \boldsymbol{U}_0 & \boldsymbol{W} & \boldsymbol{W} \\ \hline \boldsymbol{W} & \boldsymbol{W} & \boldsymbol{U}_0 & \boldsymbol{U}_1 \\ \boldsymbol{W} & \boldsymbol{W} & \boldsymbol{U}_1 & \boldsymbol{U}_0 \end{array}\right]$$

The (6 x 6) off-diagonal blocks $\boldsymbol{W}$ represent the covariance matrix of the six features between any two brands and it is assumed the same for any bound (lower or upper) or between the two bounds. We see there is a substantial improvement in the result when we use the brand information in the model by exploiting the structure $\boldsymbol{\Gamma}_y^{(3)}$ and we find the three features Taste, Naturalness and Density are vital in differentiating the two brands.

References

- L. Billard and E. Diday, (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining, John Wiley & Sons Ltd. Chichester, West Sussex, England.
- P. Giordani and H. A. L. Kiers (2006). A comparison of three methods for principal component analysis of fuzzy interval data. Computational Statistics and Data Analysis, 51: 379–397.

Friday, January 24th, 10h30 – 11h00 – F. Gioia

The Fundamental Theorem of Asset Pricing Under Uncertainty

Chiara Donnini, Federica Gioia

Dep. Management Studies and Quantitative Methods, "Parthenope" University of Naples, Italy

In the real life, several important economic decisions are subject to uncertainty, since they involve an element of risk. As in the standard literature the choice under *risky uncertainty* is studied by considering a set in which uncertain outcomes are described by means of objectively known probabilities defined on an abstract set of possible outcomes. In addition to this uncertainty dictated by the risk, many decision-making processes take place in an environment in which the data are affected by another kind of uncertainty, that is they are not known precisely. This is the case of financial data, that are often just estimated or subject to imprecision, incompleteness etc. To distinguish this uncertainty from the one generated by the risk, we will refer to it as *uncertainty in the data*. In this scenario the data may be treated by considering, instead of a single value, the interval of values in which the data may fall: the *interval data*. Interval data may be combined by the interval algebra instruments; a form of interval algebra appeared for the first time in the literature in 1924. Modern developments of such an algebra were started by R.E. Moore (1966); main results may be found in [1]. Interval algebra becomes more and more applied in domains like: economics, statistics, engineering etc. In particular, interval algebra applied to financial data has been the subject of active research over the past two decades as the traditional methods on point data do not handle the uncertainty in the data. In the presented work *risky uncertainty* and *uncertainty in the data* coexist and are studied considering a set of possible states of the world each one with the related lack of information on the data (intervals). We study an economy with initial date t_0 and terminal date t_1, risky uncertainty and uncertainty in the data; by the first one, the economy is characterized by k different possible states of the world at the second period: $\Omega=\{\omega_1, \omega_2,..., \omega_k\}$, at time t_0 investors do not know which state will occur at time t_1. The economy is endowed with n risky securities denoted as S_j for $j = 1,...,n$. and we assume that the investors do not know the exact value of the price of security S_j in the i^{th} state of the world but they bet, at best, on the *interval* of its possible values. We assume even the presence of a bank account process that will be distinguished from the other securities, because its time t_1 price will be assumed to be strictly positive for each state of the world and it is supposed to be deterministic scalar. Let us observe that the bank account is devoid of uncertainty since it is risk free and precisely known. In this framework we consider a notion of *arbitrage*, that is the opportunity to make a profit in a financial market without risk and without net investment of capital. As it is known, in the modern theory of Finance the notion of arbitrage is crucial, it is the cornerstone of the option pricing theory due to F. Black, M. Scholes and to R. Merton for which they received the Nobel prize in Economics in 1997. The principle of no arbitrage states that a mathematical model of a financial market should not allow for arbitrage possibilities. The goal of this work is to extend this principle in an economy with uncertainty, giving an extension of the Fundamental Theorem of Asset Pricing with interval prices, using the existing theory on interval linear systems [2]. In our framework the Fundamental Theorem of Asset Pricing may represent a useful tool for the investors: even if there exists uncertainty in the data joint to the risky uncertainty, the investors may know if there is no way of making arbitrage. Moreover, as consequence of the Fundamental Theorem of Asset Pricing with interval prices, we have that the *interval state price vector* [3] that characterizes the absence of arbitrage, generates infinite probability measures and those are equivalent to the probability distribution describing the risky uncertainty shared by all the investors in the market.

References

- [1] Alefeld G., Herzberger J. (1983). Introduction for interval computations. New York: Accademic Press.
- [2] Rhon J. (2003). Solvability of Systems of Linear Interval Equations, SIAM. J. Matrix Anal.and Appl., vol. 25(1), 237-245.
- [3] Gioia F. (2011). Pricing in Financial Markets with Interval Data: arbitrage and order between intervals. Advances and Applications in Statistical Sciences, vol. 6, p. 597-613.

Friday, January 24th, 11h00 – 11h30 – A. Menafoglio

O2S2 Over Complex Domains: An Approach Based on Random Domain Decompositions

Alessandra Menafoglio[a], Piercesare Secchi[a,b]

[a]MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano,
[b]Center for Analysis Decision and Society, Human Technopole, Milano

The analysis of complex data distributed over large or highly textured regions poses new challenges for spatial statistics. Although methods to deal with spatial object data have been successfully applied in several environmental studies (see, e.g., Menafoglio et al. 2017), they rely upon global models for the spatial dependence of the field, that are rarely usable in the presence of large, textured or convoluted domains, with holes or barriers.

We focus on a novel system of ideas based on Random Domain Decompositions, that allows to decompose a problem whose complexity is untractable at a global scale, in a (random) system of local tractable problems, in a divide-et-impera framework. We shall illustrate our recent methodology presented by Menafoglio et al. (2018), that enables one to perform spatial prediction of object data distributed in highly texture regions. In this framework, we propose to perform repeated Random Domain Decompositions (RDDs) of the study area, each defining a set of homogeneous sub-regions where to perform local object-oriented spatial analyses, under stationarity assumptions. This system of *weak* analyses are then aggregated into a final global analysis, in a *bagging* setting (Breiman, 1996). In this broad framework, the complexity of the domain can be taken into account by defining its partitions through a non-Euclidean metric that properly represents the adjacency relationships among the observations over the domain.

The method we propose is entirely general, and prone to be used with numerous types of object data (e.g., functional data, density data or manifold data), being grounded upon the theory of Object Oriented Spatial Statistics (O2S2). In this vein, we shall first describe the case of functional data embedded into a Hilbert space, whose geometry allows to use linear geostatistical methods in each subregion defined by the RDD. We shall finally provide insights on a recent extension of the method that allows for the analysis of data belonging to a Riemannian (Menafoglio et al., 2018). Here, the RDD may not only be used to better describe the adjacency relation among data when these are distributed a textured domain, but also to provide a system of linear approximations (tangent spaces) of the Riemannian manifold and allow for the application of linear geostatistical methods in this case too.

As an insightful illustration of the potential of the methodology, we shall consider the spatial prediction of aquatic variables in estuarine systems, that are non-convex and very irregularly shaped regions where the narrow areas of land between adjacent tributaries act as barriers. Here, we focus on the analysis and spatial prediction of distributional data (density functions and covariance matrices) relevant to the study of dissolved oxygen depletion in the Chesapeake Bay (US).

References

- Breiman, L. (1996). Bagging predictors. Machine Learning 24, 123–149.
- Menafoglio, A., G. Gaetani, and P. Secchi (2018). Random Domain Decompositions for object-oriented Kriging over complex domains. Stochastic Environmental Research and Risk Assessment 32(12), 3421–3437
- Menafoglio, A., D. Pigoli and P. Secchi (2018). Kriging Riemannian Data via Random Domain Decompositions MOX-report 64/18,Politecnico di Milano.
- Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. European Journal of Operational Research 258(2), 401–410.

Friday, January 24th, 11h30 – 12h00 – S. Bougeard

Clusterwise Regression for Multiblock High-Dimensional Data
Current Questions, Available Solutions, Developments & Perspectives

S. Bougeard[(1)], V. Cariou[(2)], H. Abdi[(3)], G. Saporta[(4)], N. Niang[(4)]

[(1)] Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail (Anses)
[(2)] Ecole Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes-Atlantique (Oniris), France
[(3)] University of Texas, Dallas, USA, [(4)] CEDRIC CNAM, Paris, France

When observations do not come from a homogenous population, global regression methods are sub-optimal. Mixture model (based on likelihood) or clusterwise regression (based on least squares) are useful when the sub-populations are unknown beforehand: these techniques simultaneously provide homogenous clusters and local regressions through the optimization of a well-defined criterion. However, in many fields, mixture models have two main limitations: (i) the number of observations in a sub-population must be greater than the number of variables; (ii) the variables must have a multi-normal distribution, but both assumptions are rarely met in practice.

First, we review and discuss the literature concerning four current questions in clusterwise regression. (i) How to deal with (multiblock) high-dimensional data? (ii) How to determine unknown clusterwise parameters? (iii) How to predict new observations? (iv) How to apply clusterwise methods in practice?

Then, responses to these questions are proposed: (i) An extension of clusterwise regression for high-dimensional data that do not follow a pre-specified distribution is presented. The variables also have the particularity of being organized into thematic blocks. The proposed method is called regularized clusterwise multiblock regression. It combines the simultaneous search for sub-populations within the observations, as well as local (multiblock) regularized regression models associated with each of these sub-populations. (ii) A test, based on the minimization of the prediction error, provides the user the unknown parameters (i.e., the optimal number of sub-populations), the optimal number of components and the optimal value of the regularization parameter. (iii) We propose to investigate a key feature generally neglected in clusterwise regression: the prediction of new observations based on a SIMCA-based procedure. It results that our method improves the quality of the prediction and facilitates the interpretation of multiblock ill-conditioned data. (iv) In practice, the proposed method is available for users though the "mbclusterwise" R package. The proposed method is illustrated on a retrospective survey conducted in 2010 in 113 French rabbit farms, which aims to identify risk indicators for antibiotic consumption.

References

- Bougeard, S. - Package R mbclusterwise (https://cran,r-project,org/web/packages/mbclusterwise /index,html), 2016.
- Bougeard S., Abdi H., Saporta G., Niang N. - Clusterwise analysis for multiblock component methods, *Adv Data Anal Classif*, 2018a, **12**(2):285-313.
- Bougeard S., Cariou V., Saporta G., Niang N. - Prediction for regularized clusterwise multiblock regression, *Appl Stoch Models Bus Ind*, 2018b, **34**(6), 852-867.
- DeSarbo W.S., Cron W.L. - A maximum likelihood methodology for clusterwise linear regression, *J Classif*, 1988, **5**, 249-282.
- Hwang H., DeSarbo W.S., Takane Y. - Fuzzy clusterwise generalized structured component analysis, *Psychometrika*, 2007, **72**, 181-198.
- Späth H. - Clusterwise linear regression, *Computing*, 1979, **22**, 367-373.

SESSION 2: Likelihood Regression Models, Unpaired variables, Symbolic Extension of TF-IDF and LDA

Friday, January 24th, 14h00 – 14h30 – Scott Sisson

Composite Likelihood and Logistic Regression Models for Aggregated Data

Scott A. Sisson, Boris Beranger and Tom Whitaker

School of Mathematics and Statistics, University of New South Wales, Sydney, Australia

Symbolic data analysis has been proposed as a technique for summarising large and complex datasets into a much smaller and tractable number of distributions – such as random rectangles or histograms – each describing a portion of the larger dataset. Recent work has developed likelihood-based methods that permit fitting models for the underlying data while only observing the distributional summaries (Beranger et al., 2018). However, while powerful, when working with random histograms this approach rapidly becomes computationally intractable as the dimension of the underlying data increases. We introduce a composite-likelihood variation of this likelihood-based approach for the analysis of random histograms in K dimensions, through the construction of lower-dimensional marginal histograms (Whitaker et al., 2019a). The performance of this approach is examined through simulated and real data analysis of max-stable models for spatial extremes using millions of observed datapoints in more than K=100 dimensions. Large computational savings are available compared to existing model fitting approaches.

Logistic regression models are a popular and effective method to predict the probability of categorical response data. However inference for these models can become computationally prohibitive for large datasets. Here we adapt ideas from symbolic data analysis to summarise the collection of predictor variables into histogram form, and perform inference on this summary dataset. We develop ideas based on composite likelihoods to derive an efficient one-versus-rest approximate composite likelihood model for histogram-based random variables, constructed from low-dimensional marginal histograms obtained from the full histogram (Whitaker et al., 2019). We demonstrate that this procedure can achieve comparable classification rates compared to the standard full data multinomial analysis and against state-of-the-art subsampling algorithms for logistic regression, but at a substantially lower computational cost. Performance is explored through simulated examples, and analyses of large supersymmetry and satellite crop classification datasets.

References

- Beranger B., H. Lin and S. A. Sisson (2018). New models for symbolic data analysis. https://arxiv.org/abs/1809.03659
- Whitaker T., B. Beranger and S. A. Sisson (2019a). Composite likelihood functions for histogram-valued random variables. https://arxiv.org/abs/1908.11548
- Whitaker T., B. Beranger and S. A. Sisson (2019b). Logistic regression models for aggregated data. https://arxiv.org/abs/1912.03805

Friday, January 24th, 14h30 – 15h00 – R. Emilion

Dependent Symbolic Variables, Testing

Richard Emilion

Denis Poisson Institute, University of Orléans, France

Our aim is to propose some models of distributions that can be fit to correlated histogram valued variables. a way of testing whether such variables are correlated or not.

Let k denote any fixed integer, $k \geq 2$. Any normalized histogram h with k bins will be represented by a vector $h = (l, r_1, \ldots, r_k, p_1, \ldots, p_k)$, where l is a real number denoting the left endpoint, r_j denotes the logarithm of the j-th bin length and p_j the area of the j-th rectangle erected over bin j, $j = 1, \ldots, k$.

1. *One variable* [2]

Let $h_1, \ldots, h_n$ be a sample of n histograms. If the two vectors $(l, r_1, \ldots, r_k)$ and $(p_1, \ldots, p_k)$ are considered as independent then a product $N(\mu,\Sigma) \otimes \text{Dirichlet}(\lambda_1, \ldots, \lambda_k)$ of a multi-variate normal distribution $N(\mu,\Sigma)$ in dimension $k+1$ and a Dirichlet distribution can be fit.

If the two vectors are dependent then start with $k + 1 + k$ normal distributions such that the $k + 1$ first ones follow a $N(\mu,\Sigma)$ distribution, and the k last ones are i.i.d. $N(0,1)$, the two sub-vectors being dependent. Now transform the last k components into independent $\Gamma(\lambda_1,1; 1), \ldots, \Gamma(\lambda_k,1; 1)$ and divide each of these k Gamma variables by their sum to get the required Dirichlet$(\lambda_1, \ldots, \lambda_k)$ which is correlated to the $N(\mu,\Sigma)$. Note that a mixture of such distributions can also be fit as in [5]. This generalizes the model proposed in [1] for intervals.

2. *More than one correlated variables*

Starting with multivariate dependent normal distributions, transform them as above to get Dependent Dirichlet distributions.

3. *Testing*

Consider two vectors having Dirichlet distributions. To testing their independence, we can first transform them by using as in [3], the so-called neutrality property, and getting vectors with independent components as in [4]. Then we can proceed to multi-testing.

References

- Brito P., Duarte Silva, A.P. (2012). Modelling Interval Data with Normal and Skew-Normal Distributions. Journal of Applied Statistics 39,1, 3–20.
- Emilion R. (2018). Dirichlet Process Mixture models for Distributional Data. SDA 2018.
- Li, Y. (2015). Goodness-of-Fit Tests For Dirichlet Distributions With Applications
  http://rave.ohiolink.edu/etdc/view?acc_num=bgsu1435003723
- Roy A., Klein D. (2018). Hypothesis Testing of Equality of Two Mean Intervals SDA 2018.
- Xia B., Wang H., Emilion R., Diday E. (2017). EM algorithm for Dirichlet Samples and its Application to Movie Data. IEEE Workshop on Analytics and Risk (Beijing, China).

Friday, January 24th, 15h00 – 15h30 – E. Diday

Improving TF-IDF and LDA (Latent Dirichlet Allocation) in the SDA Framework

E. Diday

Paris Dauphine University /PSL, France

First we introduce the density function of a class c for a category x denoted $f_c(x)$ and the density of the $f_c(x)$ when x is fixed and c varies denoted $g_x(c)$. A class is concordant with the other classes for the category x if x is frequent in the class c (i.e. $f_c(x)$ is high) and a big proportion of classes have this high frequency (i.e. $g_x(c)$ is high). A class is discordant for a category x if x is frequent in the class but few classes have this frequency for x (i.e. $g_x(c)$ is low).

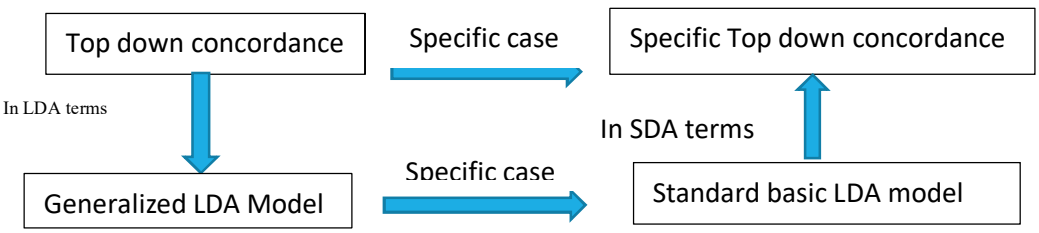
The TF-IDF is a popular index in text mining. Its basic idea is to characterize a category of a class by its relevance in comparison with the other classes. It is high if this category is frequent inside the class and rare are the given classes where it appears. We show that the discordance enhance the TF-IDF and contains it as a case.

As the TF-IDF takes only care on the existence or not of a category in the classes it cannot detect some relevance of some categories that the symbolic discordance can do in the case where their frequency in a class rarely appears in the other classes.

The LDA model (Blei et al. (2003)), is a generative probabilistic semantic able to answer questions such as "what is the probability of this previously unseen document?". The LDA model posits that each document is generated as a mixture of topics, where the continuous-valued mixture proportions are distributed as a latent Dirichlet random variable.

More precisely, the top down LDA model aims to provide an unseen document of N words and to give its probability. It starts from a vector Θ defining the probability of K topics sampled from a given Dirichlet model $p(\Theta; \alpha)$. A topic z_n is sampled from Θ. From this topic, a word w_n is sampled from a given multinomial $p(w_n/\Theta_{ik}; \beta)$. This process is repeated until a document c of N words is obtained. Then, the probability of c can be expressed in term of the standard LDA as follows:

$p(c) = \int_\Theta (\prod_{n=1,N} \sum_{z_n} p(w_n/ z_n; \beta)\, p(z_n/ \Theta)) p(\Theta; \alpha)) d\Theta$ which is the basic equation of the LDA model defined in Blei et al. (2003). In comparison with the LDA model, the concordance approach can be considered as a bottom up model obtained by building $f_c(x)$ and $g_x(c)$ from the individuals of a given class c described by the symbol f_c defined by the frequency of each of the given categories in c. A bottom up concordance can be defined which leads to a more accurate LDA basic equation. The following schema summarize these results:



References

- Blei D., Ng A.,. Jordan M. "Latent Dirichlet Allocation" .JMRL 3(Jan): 993-1022, 2003.
- Diday E. Explanatory Tools for Machine Learning in the Symbolic Data Analysis Framework. Chap 1 in Diday E., Rong G., Saporta G., Wang H., (editors)(2019 or 2020). Advances in Data Science (Symbolic, Complex and Network Data). ISTE WILEY Science Publishing Ltd.

SESSION 3: ***Data Bases and Software***

Friday, January 24th, 16h00 – 16h30 – W. Litwin

SQL for Stored and Inherited Relations
Manifesto for Improved Foundations of Relational Model

Witold Litwin

Université Paris Dauphine, PSL

Normalized base relations extended with inherited attributes may be more faithful to reality. Typical queries may become logical navigation free, hence less procedural. These properties are offered at present only by specific views. Adding inherited attributes can be nonetheless always less procedural than to define any such views. Present schemes of relations with foreign keys should even typically suffice. One can interpret those as defining inherited attributes as well. Implementing extended relations on popular DBSs should be simple. Relational model should evolve, for benefit of likely millions of DBAs, clients, developers….

References

- Witold Litwin. Manifesto for Improved Foundations of Relational Model. Procedia Computer Science, Vol. 160, 2019, pp. 624-628, Elsevier, (publ.).