UNIVERSITÀ degli STUDI di CATANIA | DIPARTIMENTO di ECONOMIA e IMPRESA

MBC^2 2018

Model-Based Clustering and Classification

4^{th} International Workshop

Catania, September 5–7, 2018

Book of Abstracts

Editors:
Salvatore Ingrassia and Antonio Punzo

Committees

Scientific

Christophe Biernacki (France), Andrea Cerioli (Italy), Luis Angel Garcìa-Escudero (Spain), Christian Hennig (UK), Salvatore Ingrassia (Italy), Geoffrey McLachlan (Australia), Volodymyr Melnykov (USA), Angela Montanari (Italy), Roberto Rocci (Italy), Maurizio Vichi (Italy).

Organizing

Salvatore Ingrassia (chair), Roberto Di Mari, Angelo Mazza, Antonio Punzo, Salvatore Daniele Tomarchio.

Contents

List of Abstracts

Talk Session 1: Robustness and regularization

5 Sept.
18.15–19.40
TS 1

Wild adaptive trimming for robust estimation and cluster analysis

Andrea Cerioli[1], Alessio Farcomeni[2] and Marco Riani[1]

[1]University of Parma, Italy; [2]University of Rome, Italy

Trimming principles play an important role in robust statistics. However, their use for clustering typically requires some preliminary information about the contamination rate and the number of groups. In this talk we describe a fresh approach to trimming that does not rely on this preliminary knowledge and that proves to be particularly suited for solving problems in robust cluster analysis. Our approach replaces the original K-population (robust) estimation problem with K distinct one-population steps, which take advantage of the good breakdown properties of trimmed estimators when the trimming level exceeds the usual bound of 0.5. In this setting we prove that exact affine equivariance is lost on one hand, but on the other hand an arbitrarily high breakdown point can be achieved by "anchoring" the robust estimator. We also support the use of adaptive trimming schemes, in order to infer the contamination rate from the data. A further bonus of our methodology is its ability to provide a reliable choice of the usually unknown number of groups.

References

Cerioli, A., A. Farcomeni, and M. Riani (in press). Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, DOI: 10.1111/sjos.12349.

---

5 Sept.
18.15–19.40
TS 1

Constrained maximum likelihood estimation of clusterwise linear regression models with unknown number of components

Roberto Di Mari[1], Roberto Rocci[2], and Stefano Antonio Gattone[3]

[1]Department of Economics and Business, University of Catania, Italy; [2]Department of Economics and Finance, University of Rome Tor Vergata, Italy; [3]Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, Italy.

We consider an equivariant approach imposing data-driven bounds for the variances to avoid singular and spurious solutions in maximum likelihood (ML) estimation of clusterwise linear regression models. We investigate its use in the choice of the number of components and we propose a computational shortcut, which significantly reduces the computational time needed to tune the bounds on the data. In the simulation study and the two real-data applications, we show that the proposed methods guarantee a reliable assessment of the number of components compared to standard unconstrained methods, together with accurate model parameters estimation and cluster recovery.

Exploring robust Fuzzy clustering on multivariate skew data

Francesca Greselin[1], Luis Angel García-Escudero[2] and Agustin Mayo-Iscar[2]
[1]University of Milano-Bicocca; [2]University of Valladolid

With the increasing availability of multivariate datasets, asymmetric structures in the data ask for more realistic assumptions, with respect to the incredibly useful paradigm given by the Gaussian distribution. Moreover, in performing ML estimation we know that a few outliers in the data can affect the estimation, hence providing unreliable inference. Challenged by such issues, more flexible and solid tools for modeling heterogeneous skew data are needed. Our fuzzy clustering method is based on mixtures of Skew Gaussian components, endowed by the joint usage of impartial trimming and constrained estimation of scatter matrices, in a modified maximum likelihood approach. The algorithm generates a set of membership values, that are used to fuzzy partition the data set and to contribute to the robust estimates of the mixture parameters. The new methodology has been shown to be resistant to different types of contamination, by applying it on artificial data. A brief discussion on the tuning parameters has been developed, also with the help of some heuristic tools for their choice. Finally, synthetic and real dataset are analyzed, to show how intermediate membership values are estimated for observations lying at cluster overlap, while cluster cores are composed by observations that are assigned to a cluster in a crisp way.

References

Davé R.N., Krishnapuram R. (1997) Robust clustering methods: a unified view, *IEEE Transactions on Fuzzy Systems*, **5**, 270–293.

Dotto F., Farcomeni A., García-Escudero L.A. and Mayo-Iscar A. (2016) A fuzzy approach to robust regression clustering, *Advances in Data Analysis and Classification*, 1–20.

Fritz H., García-Escudero L.A. and Mayo-Iscar A. (2013) Robust constrained fuzzy clustering, *Information Sciences*, **245**, 38–52.

García-Escudero L.A., Greselin F., Mayo-Iscar A. (2018) Robust fuzzy and parsimonious clustering based on mixtures of Factor Analyzers, *International Journal of Approximate Reasoning*, **94**, 60–75.

Gustafson E.E., Kessel W.C. (1979) Fuzzy clustering with a fuzzy covariance matrix, in: *Proceedings of the IEEE lnternational Conference on Fuzzy Systems*, San Diego, 761–766.

Rousseeuw P.J., Trauwaert E. and Kaufman L. (1995) Fuzzy clustering with high contrast. *Journal of Computational and Applied Mathematics,* **64**, 81–90.

Trauwaert E., Kaufman L., Rousseeuw P. (1991) Fuzzy clustering algorithms based on the maximum likelihood principle. *Fuzzy Sets and Systems*, **42**(2), 213–227.

Keynote Lecture 1

6 Sept.
08.30-09.25
KL 1

Model Based Clustering through copulas for high dimensional data

Dimitris Karlis

Dept of Statistics, Athens University of Economics and Business

In a recent paper Kosmidis and Karlis (2016) proposed model based clustering based on multivariate distributions defined through copulas. This approach offers a number of advantages over existing methods mainly due to the flexibility to define appropriate models in certain different circumstances. Some of the existing models can be seen as a special case of this construction. In this talk we exploit the ideas of extending the approach for higher dimensions and different types of data. The central idea is to use a Gaussian copula and implement the correlation matrix of the Gaussian copula through certain parsimonious representations giving rise to models of different complexity. To some extend this is based on existing representations in the MBC literature suitably adapted for the case of Gaussian copulas. We use two different approaches, the first makes use of factor analyzers based on the factor decomposition of the correlation matrix and the second is based on Choleski type decompositions. Application with real and simulated data will be also described.

This is joint work with Ioannis Kosmidis, University of Warwick and Fotini Panagou (AUEB)

References

Kosmidis I and Karlis D (2016). Model-based clustering using copulas with applications. *Statistics and Computing*, **26**(5), 1079–1099

Keynote Lecture 2

Heterogeneity in large-scale data: invariance, causality and novel robustness

6 Sept.
09.25-10.20
KL 2

Peter Bühlmann
ETH Zürich

Heterogeneity in potentially large-scale data can be beneficially exploited for causal inference and novel robustness. The key idea relies on invariance and stability across different heterogeneous regimes or sub-populations (Peters et al., 2016). What we term as "anchor regression" (Rothenhäusler et al., 2018) opens up new insights and connections between causality and protection (robustness) against worst case perturbations in prediction problems. We will discuss the methodology and some applications.

References

Peters, J., Bühlmann, P. and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals *Journal of the Royal Statistical Society, Series B*, **78**, 947–1012.

Rothenhäusler, D., Bühlmann, P., Meinshausen, N. and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. *Preprint arXiv:1801.06229*.

Talk Session 2: Supervised and unsupervised classification

Unobserved Classes and Extra Variables in High-dimensional Discriminant Analysis

6 Sept.
10.50–12.35
TS 2

Michael Fop[1], Pierre-Alexandre Mattei[2], Thomas Brendan Murphy[1] and Charles Bouveyron[3]

[1]School of Mathematics & Statistics, University College Dublin, Ireland.
[2]Department of Computer Science, IT University of Copenhagen, Denmark.
[3] Laboratoire J.A. Dieudonné, UMR CNRS 7351 & Equipe Epione, INRIA Sophia-Antipolis, Université Côte d'Azur, France.

In supervised classification problems, the test set may contain data points belonging to classes not observed in the learning phase. Moreover, the same units in the test data may be measured on a set of additional variables, recorded at a subsequent stage with respect to when the learning sample was collected. In this situation, the classifier built in the learning phase needs to adapt to handle potential unknown classes and the extra dimensions. We introduce a model-based discriminant approach that can detect unobserved classes and adapt to the increasing dimensionality. The method is embedded in a more general framework for adaptive variable selection and classification, developed in application to high-dimensional spectrometry data.

References

Bouveyron, C. (2014). Adaptive mixture discriminant analysis for supervised learning with unobserved classes, *Journal of Classification*, **31**(1), 49–84.

Maugis, C., Celeux, G., and Martin-Magniette, M. L. (2011). Variable selection in model-based discriminant analysis, *Journal of Multivariate Analysis*, **102**(10), 1374–1387.

Maximizing the Usefulness of Statistical Classifiers for Two Populations

6 Sept.
10.50–12.35
TS 2

Daniel. R. Jeske

Department of Statistics University of California, Riverside

The usefulness of two-class statistical classifiers is limited when one or both of the conditional misclassification rates is unacceptably high. Incorporating a neutral zone region into the classifier provides a mechanism to refer ambiguous cases to follow-up where additional information might be obtained to clarify the classification decision. Through the use of the neutral zone region, the conditional misclassification rates can be controlled and the classifier becomes useful. An application to prostate cancer will be used to illustrate how neutral zone regions can extract utility from a potentially disappointing classifier that might otherwise be abandoned.

A multivariate characterisation of some popular cluster analysis methods

6 Sept.
10.50–12.35
TS 2

Christian Hennig
University College London

I have argued (Hennig 2015) that there are various different aims of cluster analysis, for which different clusterings may be optimal even on the same dataset. I present a collection of indexes that measure different aspects of interest in clustering (such as within-cluster homogeneity, between-cluster separation, representation of the underlying distance structure by the clustering, correspondence to high density regions, good representation of clusters by centroids etc.). There are a number of cluster validity indexes proposed in the literature (Valkidi et al. 2015). Most if not all of them attempt to give a one-dimensional assessment of the overall quality of a clustering, which does not provide insight into how the trade-off between the specific characteristics that could be potentially desirable plays out.

The proposed collection of indexes is used to give a multivariate characterisation of the behaviour of some popular clustering methods including Gaussian and skew-t mixtures based on 20 real datasets. The focus here is not on "recovering the true clusters" but rather on elaborating how the methods differ in a data analytic sense. This can help users to choose an appropriate method for a specific clustering task.

References

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters* **64**, 53–62.

Hennig, C. (2017) Cluster validation by measurement of clustering characteristics relevant to the user. arXiv:1703.09282.

Valkidi, M., Vazirgiannis, M. and Hennig, C. (2015) Method-Independent Indices for Cluster Validation and Estimating the Number of Clusters. In: Hennig C., Meila M., Murtagh F. and Rocci R. (eds.) *Handbook of Cluster Analysis*. Chapman and Hall/CRC, USA.

Model-based double hierarchical parsimonious clustering

6 Sept.
10.50–12.35
TS 2

Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria
Department of Statistical Sciences, Sapienza University of Rome

Starting from a dissimilarity data set between n statistical units, the hierarchical classifications produced by clustering algorithms usually comprise partitions into K classes for all values of K between 1 and n, being represented by dendrograms that contain $(n-1)$ internal nodes. Several authors noted that the complete sets of partitions and classes do not appear to be used by investigators, and can hinder interpretation. One approach for resolving this difficulty has involved the construction of *parsimonious trees*, which contain a limited number of internal nodes; some information is discarded in this process, but the main features of the data can be represented more clearly.

In this papers starting from the data matrix X of size $(n \times J)$, corresponding to n statistical units and J quantitative variables, we propose a model for double hierarchical

parsimonious clustering. This is a new methodology for simultaneous hierarchical parsimonious clustering of the units – aggregated around centroids – and of the variables – aggregated around factors. The model is estimated by using the LS method and an efficient coordinate descent algorithm is given. The goodness of fit of the double hierarchical parsimonious trees can be computed to assess the quality of the two hierarchical partitions.

Lightning Talk Session 1

Supervised classification with matrix sketching

6 Sept.
12.35–13.00
LT 1

Laura Anderlucci, Roberta Falcone and Angela Montanari
University of Bologna

Matrix sketching is a data compression technique that has been recently developed in the computer science community. An input matrix A is efficiently approximated with a smaller matrix B, so that B preserves most of the properties of A up to some guaranteed approximation ratio. In so doing numerical operations on big data sets become faster. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. The statistical properties of sketched regression algorithms have been widely studied in Woodruff (2014) and in Ahfock, Astle and Richardson (2017). In this work we study the performances of sketching algorithms in the supervised classification context, both in terms of misclassification rate and of boundary approximation, as the degree of sketching increases. We also address, through sketching, the issue of unbalanced classes, which hampers most of the common classification methods.

References

Ahfock, D., Astle, W.J. and Richardson S. (2017). Statistical properties of sketching algorithms, *https://arxiv.org/abs/1706.03665*.

Jing-Hao, X. and Hall, P. (2015) Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?, *IEEE transactions on pattern analysis and machine intelligence*, **37**(5), 1109–1112.

Jing-Hao, X. and Titterington, D. M. (2008) Do unbalanced data have a negative effect on LDA?, *Pattern Recognition*, **41**(5), 1558–1571.

Woodruff D. P.(2014). *Sketching as a Tool for Numerical Linear Algebra.* Foundations and Trends in Theoretical Computer Science, vol 10, issue 1-2.

Data Units as (Co-)Clustering Model Enlargement

6 Sept.
12.35–13.00
LT 1

Christophe Biernacki[1] and Alexandre Lourme[2]
[1]Inria & University of Lille & CNRS (France); [2]University of Bordeaux (France)

Model-Based-Clustering methods [McLachlan G. and Peel D., 2000] aim at splitting unlabelled data into groups. Model-Based-Co-Clustering methods [Govaert G. and Nadif M., 2013] yields simultaneously groups of data and variables. In both cases, changing the data units may affect the estimated partition(s). But, combining several data units with scale dependent models enables to enlarge the set of competing (Co-)Clustering models since any unit change can be seen as a particular model definition [Biernacki C. and Lourme A., 2018]. Consequently, it raises the following open question: how to select a model when the number of (co-)clustering models explodes?

References

Govaert G. and Nadif M. (2013). *Co-Clustering*. Wiley.

McLachlan G. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

Biernacki C. and Lourme A. (2018). Unifying data units and models in (co-)clustering. *Advances in Data Analysis and Classification*.

Robust Adaptive Eigenvalue Decomposition Discriminant Analysis: supervised learning in presence of outliers, label noise and unobserved classes

6 Sept.
12.35–13.00
LT 1

Andrea Cappozzo, Francesca Greselin[1] and Thomas Brendan Murphy[2]

[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 20126, Milano, Italy; [2]School Of Mathematics & Statistics and Insight Research Centre, University College Dublin, Belfield, Dublin 4, Ireland

In a standard classification framework a set of trustworthy learning data are employed to build a decision rule, with the final aim of classifying unlabelled units belonging to the test set. Therefore, unreliable learning observations can strongly undermine the classifier performance, especially if the training size is small. Additionally, the test set may include classes not previously encountered in the learning phase. The present work introduces a robust adaptive model-based discriminant analysis (RAEDDA) capable of handling situations in which one or more of the following problems occur: outliers both in the training and in the test set, label noise in the training set and extra classes in the test not observed in the learning phase. An inductive EM-based procedure is employed for robust parameter estimation, making use of impartial trimming for identifying possible outliers and data with uncertain labels. Experiments on real data, artificially adulterated, are provided to underline the benefits of the proposed method.

References

Bouveyron, C. (2014). Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes, *Journal of Classification*, **31**(1), 49–84.

Bouveyron, C. and Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels, *Pattern Recognition*, **42**, 2649–2658.

Dean, N., Murphy, T. B. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, **55**(1), 1–14.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **957**(458), 611–631.

García-Escudero, L. A. and Gordaliza, A. and Mayo-Iscar, A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions, *Advances in Data Analysis and Classification*, **8**(1), 27–43.

McLachlan G. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

Neykov, N. and Filzmoser, P. and Dimova, R. and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator, *Computational Statistics & Data Analysis*, **52**(1), 299–308.

A thinned-trimmed CEM algorithm for robust clustering around regression lines

6 Sept.
12.35–13.00
LT 1

Andrea Cerioli[1], Luis Angel Garcia-Escudero[2], Agustin Mayo-Iscar[2], Domenico Perrotta[3] and Francesca Torti[3]

[1]University of Parma, Italy; [2]University of Valladolid, Spain; [3]European Commission, Joint Research Centre

Robust techniques typically allow for a contamination rate of at most 50% and robust clustering is no exception to this rule (García-Escudero et. al, 2008, and Ritter, 2014). The motivation lies in the (sometimes implicit) assumption that the "good" population should correspond to the majority of data. However, there may be instances where the relevant structure only concerns a possibly very small portion of the data, while the remaining observations do not contribute any meaningful information. In these situations robust methodologies clearly fail to unveil the relevant data structure even if they are tuned to ensure the supposedly maximum value of their breakdown point, i.e. 50% (see Cerioli et al., 2017, and the related discussions).

One way to attack the problem in a multivariate framework is to adopt robust estimation procedures based on trimming with arbitrarily high breakdown (Cerioli et al., 2018). Instead, in this work we follow a different approach which proves to be preferable when the data have a cluster-wise regression structure and the largest portion of contaminated observations can be interpreted as concentrated noise. This is precisely the context of our motivating application field: fraud detection in international trade data, where traded values and quantities are recorded. These variables are linked by a functional linear relationship and anomalous transactions often stand out as outliers from the regression clusters defined by genuine trading behaviour. Noise corresponds to the presence of a possibly very dense "small trade area", which is often concentrated close to the origin of the coordinate axes but which can also span along one or more regression lines.

Our approach combines the usual levels of trimming (i.e., up to 50%) with thinning, a denoising procedure based on density estimation. Thinning is able to remove a very high fraction of the observations that contribute to noise, thus allowing robust cluster-wise regression methods, such as TCLUST-REG (García-Escudero et al., 2010), to work on data with "standard" (i.e., lower than 50%) contamination rates. For this purpose, we develop a modified version of the Classification EM (CEM) algorithm of García-Escudero et al. (2010), where thinning weights are computed and applied before each maximization step. An additional problem that we address is the comparison of alternative solutions arising from different random starting points of the CEM algorithm, which may be based on a different number of observations retained after thinning.

We show the potential of our method through simulated data and through applications to real anti-fraud scenarios in international trade. We also

- compare the performance of our proposal with the tandem approach developed by Cerioli and Perrotta (2014), in which denoising is performed before robust cluster-wise regression and thus outside the robust CEM algorithm;
- investigate its performance under alternative attitudes towards robustness to high-leverage points, a critical issue in robust regression methods (Huber and Ronchetti, 2009 and García-Escudero et al., 2010).

References

Cerioli, A., A. Farcomeni, and M. Riani (2018). Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics in press*, DOI: 10.1111/sjos.12349.

Cerioli, A. and D. Perrotta (2014). Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification* **8**, 5–26.

Cerioli, A., M. Riani, A. C. Atkinson, and A. Corbellini (2017). The power of monitoring: how to make the most of a contaminated multivariate sample (with discussion). *Statistical Methods and Applications in press*, DOI: 10.1007/s10260-017-0409-8.

García-Escudero, L., A. Gordaliza, A. Mayo-Iscar, and R. San Martin (2010). Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis* **54**(12), 3057–3069.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics* **36**, 1324–1345.

Huber, P. J. and E. M. Ronchetti (2009). *Robust Statistics. Second Edition*. Hoboken: Wiley.

Ritter, G. (2014). *Robust Cluster Analysis and Variable Selection*. Boca Raton: Chapman and Hall/CRC.

Critical Discussions of Selected Robust Clustering Procedures and Their Applications in Economics

6 Sept.
12.35–13.00
LT 1

Przemysław Jaśko, Daniel Kosiorowski and Ewa Szlachtowska

Cracow University of Economics

In this paper we critically discusses advantages and disandvantages of the selected robust clustering methods known from literature. We among others study TCLUST algorithm proposed by García-Escudero et al. (2008) and compare it with selected model-based procedures appealing to Bayesian Clustering and Mixture Clustering.

References

Coretto P. and Hennig C. (2010). A simulation study to compare robust clustering methods based on mixtures, *Advances in Data Analysis and Classification*, **4**(2–3), 111–135.

García-Escudero L. A., Gordaliza A., Matrán C. and Mayo-Iscar A. (2008). A general

trimming approach to robust cluster analysis, *The Annals of Statistics*, **36**(3), 1324–1345.

Liverani S., Hastie D. I., Azizi L., Papathomas M. and Richardson S. (2015). **PReMiuM**: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. R package Version 3.1.7, available at `https://CRAN.R-project.org/package=PReMiuM`.

Nia V. P. and Davison A. C. (2012). **bclust**: An R Package for High-Dimensional Bayesian Clustering with Variable Selection. R package Version 1.5, available at `https://CRAN.R-project.org/package=bclust`.

Szlachtowska E., Kosiorowski D. and Mielczarek D. (2016). Ocena jakości aplikacyjnej odpornego algorytmu analizy skupień TCLUST na przykładzie zbioru danych dotyczacych jakości powietrza w Krakowie, *Przeglad Statystyczny*, **63**(1), 67–80.

Robust clustering in the presence of skewed data groups

6 Sept.
12.35–13.00
LT 1

Yana Melnykov[1], Volodymyr Melnykov[1] and Xuwen Zhu[2]

[1]The University of Alabama; [2]University of Louisville

The performance of model-based clustering depends on the presence of noise or outlying observations severely. Such observations might ruin the systematic structure of the groups leading to incorrect or misleading results. Among the most famous approaches taking into account the potential presence of outliers, there are finite mixture models with t, skew-t, or contaminated normal components. We propose an alternative to the traditional approaches that is capable of modeling skewed heavy-tailed data groups effectively. A novel approach to identifying noise observations is introduced. The procedure is illustrated on simulated and real-life data sets with good results.

References

Lin T.I., Lee J.C., Hsieh W.J. (2007) Robust Mixture Modeling Using the Skew t Distribution, *Statistics and Computing*, **17**(2), 81–92.

Peel D., McLachlan G.J. (2000) Robust Mixture Modeling Using the t Distribution, *Statistics and Computing*, **10**(4), 339–348.

Punzo A., McNicholas P.D. (2017) Robust Clustering in Regression Analysis via the Contaminated Gaussian Cluster-Weighted Model, *Journal of Classification*, **34**(2), 249–293.

---

6 Sept.
12.35–13.00
LT 1

Gaussian mixture modeling and model-based clustering under measurement inconsistency

Volodymyr Melnykov[1], Shuchismita Sarkar[1] and Rong Zheng[2]
[1]University of Alabama; [2]Western Illinois University

Finite mixtures present a powerful tool for modeling complex heterogeneous data. One of their most important applications is model-based clustering. It assumes that each data group can be reasonably described by one of mixture model components. This establishes a one-to-one relationship between mixture components and clusters. In some cases, however, this relationship can be broken due to the presence of observations from the same class recorded in different ways. This effect can occur because of recording inconsistencies due to the use of different scales, operator errors, or simply various recording styles. The idea presented in this paper aims to alleviate this issue through modifications incorporated into mixture models. While the proposed methodology is applicable to a broad class of mixture models, in this paper it is illustrated on Gaussian mixtures. Several simulation studies and an application to a real-life data set are considered, yielding promising results.

References

Alimoglu, F. and Alpaydin, E. (1996). Methods of Combining Multiple Classifiers Based on Different Representations for Pen-based Handwriting Recognition. *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Baudry, J. P., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, **19**, 332–353.

Celeux, G. and Govaert (1995). Gaussian parsimonious clustering models. *Computational Statistics and Data Analysis*, **28**, 781–93.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, **14**, 315–332.

Dasgupta, S. (1999). Learning mixtures of Gaussians. In: Proc. IEEE *Symposium on Foundations of Computer Science*, pp. 633–644 New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Jounal of the Royal Statistical Society*, Series B, **39**, 1–38.

Di Zio, M., Guarnera, U., and Rocci, R. (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics and Data Analysis*, **51**(5), 2573–2585.

Fisher, P. (1999). Models of uncertainty in spatial data. *Geographical Information Systems*, **1**, 191–205.

Fop, M., Murphy, T. B., and Hanlon, L. (2017). Model-based Clustering of Data with Measurement Errors. In: *CLADAG 2017*, 23, Italy.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, **7**, 385–405.

Han, J., Kamber, M., and Pei, J. (eds.) (2012).*Data mining: concepts and techniques*, 3rd ed, Elsevier, New York.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Just, B. H., Marc, D., Munns, M., and Sandefer, R. (2016). Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields. *Perspectives in Health Information Management*, 13.

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

Kumar, M. and Patel, N. (2007). Clustering data with measurement errors. *Computational Statistics and Data Analysis*, **51**(12), 6084–6101.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium*, **1**, 281–297.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models New York*: Wiley, New York.

Melnykov, V. (2013). Finite mixture modelling in mass spectrometry analysis. *Journal of the Royal Statistical Society* Series C, **62**, 573–592.

Melnykov, V. (2016), Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics*, **25**, 66–90.

Pankove, J. I. (2012). *Optical processes in semiconductors*. Courier Corporation.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, **185**, 71–110.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. In: *IEEE Data Eng. Bull.*, **23**(4), 3–13.

Schlattmann, P. (2009). *Medical applications of finite mixture models*, Springer.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, **6**, 461–464. 24

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). **mclust** 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models *R Journal*, **8**, 289–317.

Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, **17**, 201–226.

Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**, 1409–1438.

Tjaden, B. (2006). An approach for clustering gene expression data with error information. *Bmc Bioinformatics*, **7**(1), 17.

Ullrich, B., Antilln, A., Bhowmick, M., Wang, J., and Xi, H. (2014). Atomic transition region at the crossover between quantum dots to molecules. *Physica Scriptal*, **89**(2), 025801.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Young, W. C., Raftery, A. E., and Yeung, K. Y. (2017). Model-based clustering with data correction for removing artifacts in gene expression data. *Annals of Applied Statistics*, **11**(4), 1998–2026.

Zhu, X. and Melnykov, V. (2016). Manly transformation in finite mixture modeling. *Computational Statistics and Data Analysis*, **121**, 190–208.

Keynote Lecture 3

Classification, clustering and co-clustering for ordinal data

6 Sept.
14.00–14.55
KL 3

Julien Jacques[1], Margot Selosse[1] and Christophe Biernacki[2]
[1]UniversitÃl' de Lyon, Lyon 2 ERIC EA3083; [2]UniversitÃl' de Lille & Inria

A model-based approach for analyzing and modeling ordinal data is presented. This model relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Classification, clustering and co-clustering algorithms are derived from the proposed model. Model inference relies on a Stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of clusters in clustering, the number of co-clusters in co-clustering, and the level of parsimony in classification. The main advantages of these ordinal dedicated models are their parsimony, the interpretability of the parameters (mode, precision) and the possibility to take into account missing data. All these algorithms are available in the ordinalClust package for R. The usefulness of the proposed algorithms are illustrated by analyzing a psychological survey on women affected by a breast tumor.

References

Biernacki C. and Jacques J. (2016), Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm, *Statistics and Computing*, 26 [5], 929-943.

Jacques J. and Biernacki C.(2018), Model-based co-clustering for ordinal data, *Computational Statistics and Data Analysis*, 123, 101-115.

Selosse M., Jacques J., Biernacki C. (2017). ordinalClust: a package for analyzing ordinal data, *Preprint HAL n°01678800*.

Selosse M., Jacques J., Biernacki C. (2017). **ordinalClust**. R package Version 1.2, available at `https://CRAN.R-project.org/package=ordinalClust`.

Selosse M., Jacques J., Biernacki C. and Cousson-GÃl'lie F. (2017). Analyzing health quality survey using constrained co-clustering model for ordinal data and some dynamic implication, *Preprint HAL n°01643910*.

Keynote Lecture 4

6 Sept.
14.55-15.50
KL 4

Unifying robust clustering aggregation based on optimal transportation

Eustasio del Barrio
IMUVA, Universidad de Valladolid

A robust clustering method for probabilities in Wasserstein space is introduced. This new "trimmed k-barycenters" approach relies on recent results on barycenters in Wasserstein space that allow intensive computation, as required by clustering algorithms to be feasible. The possibility of trimming the most discrepant distributions results in a gain in stability and robustness, highly convenient in this setting. As a remarkable application, we consider a parallelized clustering setup in which each of m units processes a portion of the data, producing a clustering report, encoded as k probabilities. We prove that the trimmed k-barycenter of the $m \times k$ reports produces a consistent aggregation which we consider the result of a "wide consensus". We also prove that a weighted version of trimmed k-means algorithms based on k-barycenters in the space of Wasserstein keeps the descending character of the concentration step, guaranteeing convergence to local minima. We illustrate the methodology with simulated and real data examples. These include clustering populations by age distributions and analysis of cytometric data.

References

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Mayo-Iscar, A. (2018). Robust clustering tools based on optimal transportation. *Statistics and Computing*, DOI: 10.1007/s11222-018-9800-z.

Talk Session 3: Model-based clustering of complex data

Clustering for multidimensional networks via infinite mixture models

6 Sept. 16.20–17.45 TS 3

Silvia D'Angelo[1], Michael Fop[2] and Marco Alfò[1]

[1]Sapienza, University of Rome; [2]University College Dublin

Network data are relational data where the presence of a given relation between any two units is expressed by an edge connecting them. When multiple relations are observed among the same group of nodes, a collection of networks is available. This collection takes the name of multidimensional network, or multiplex. Latent space models for network data describe the observed structure in the multidimensional network by means of an unobserved latent space, in which the main assumption is that units close in the latent space are more likely to be connected. In many data applications, units have the propensity to aggregate into communities. This characteristic has been modelled in the context of single and dynamic networks. We propose a clustering framework for multidimensional networks based on infinite mixtures of Gaussian distributions. The model is estimated within a hierarchical Bayesian framework. Moreover, a single latent space is employed to describe the multiplex, both for representation and computational efficiency purposes.

References

Hoff P. D., Raftery A. E. and Handcock, M. S. (2002), Latent Space Approaches to Social Network Analysis, *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Handcock M., Raftery A. and Tantrum J. (2007), Model-based clustering for social networks (with discussion), *Journal of the Royal Statistical Society: Series A*, **170**(2), 1–22.

Sewell D. K. and Chen Y. (2017), Latent space approaches to community detection in dynamic networks, *Bayesian Analysis*, **12**(2), 351–377.

On modelling multivariate high-dimensional time series: a factorial hidden Markov model

6 Sept. 16.20–17.45 TS 3

Antonello Maruotti[1], Antonio Punzo[2] and Jan Bulla[3]

[1]Libera UniversitÃă Maria Ss. Assunta; [2]UniversitÃă di Catania; [3]University of Bergen

We introduce multivariate models for the analysis of stock market returns. Our models are developed within the hidden Markov framework to describe the temporal evolution of the returns, whereas the marginal distribution of the returns is described by a mixture of multivariate contaminated-normal distributions. The contaminated-normal distribution represents an elliptical generalization of the Gaussian distribution allowing for automatic outlier/extreme value detection in the same natural way as observations are typically assigned to the latent states in the hidden Markov model (HMM) context (Punzo and Maruotti, 2016). The proposed HMMs account for three major dependency structures in multivariate time series data, including the correlation among multiple series,

temporal dependence, and heterogeneity. Following the HMMs literature, we assume that the hidden structure underlying the observed data is a first-order Markov chain, and that returns can be modeled as a multivariate process conditioning on the sequence of hidden states. The challenge of modeling multiple stock series and their interactions is fairly common to all analyses of high dimensional data with many variable of interest. Dimensionality-related aspects present a challenge because these series could potentially be highly correlated. Therefore, estimation and interpretation of model parameters may become nontrivial. In order to examine the interrelationships between series to perform dimensionality reduction in the variable space simultaneously (allowing for an easy interpretation of model parameters), we propose the use of a latent factor model (Maruotti et al., 2017). Accordingly, we define a general class of parsimonious HMMs by imposing a factor decomposition on state-specific covariance matrices. The loadings and noise terms of the covariance matrix may be constrained to be equal or unequal across latent states. In addition, the noise term may be subject to further restrictions, resulting in a set of eight parsimonious covariance structures (McNicholas and Murphy, 2008). This model structure allows for the accounting of dependence between series, and provides a clear interpretation of the (latent) association structure between series. Even in this relatively general framework, the parameters of the proposed parsimonious HMMs can be estimated using the method of maximum likelihood based on the Alternating Expectation Conditional Maximization (AECM) algorithm.

References

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296.

Maruotti, A., Bulla, J., Lagona, F., Picone, M. and Martella, F. (2017). Dynamic mixture of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, **11**, 1617–1648.

Punzo, A. and Maruotti, A. (2016). Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. *Journal of Computational and Graphical Statistics*, **25**, 1097–1116.

---

6 Sept.
16.20–17.45
TS 3

Clustering of spatially dependent functional data

Vincent Vandewalle[1], Cristian Preda[2] and Sophie Dabo[3]

[1]Université de Lille, EA 2694 & Inria; [2]Université de Lille, Laboratoire Paul Painlevé, UMR 8524 & Inria; [3]Université de Lille, Laboratoire LEM UMR CNRS 9221 & Inria

Two approaches for clustering spatial functional data are presented. The first one is the model-based clustering that uses the concept of density for functional random variables and logistic weights on the prior cluster probabilities depending on spatial coordinates. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean, and includes spatial weights in the distances computation. These two approaches take into account the spatial features of the functional data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data.

References

Dabo-Niang, S., Yao, A. F., Pischedda, L., Cuny, P., and Gilbert, F. (2010) Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, **24**(4), 487–497.

Delaigle, A. and Hall, P. (2010) Defining probability density for a distribution of random functions. *The Annals of Statistics*, pp. 1171–1193.

Jacques, J. and Preda, C. (2014) Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 231–255.

Cheam, A., Marbac, M., and McNicholas, P. (2017) Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics*, **28**(3), 1–11.

Talk Session 4: Developments in modeling high-dimensional data

7 Sept.
8.45–10.10
TS 4

High-dimensional clustering with Random Projections

Laura Anderlucci, Francesca Fortunato and Angela Montanari
University of Bologna

Random projections (RPs) have shown to provide promising results for high-dimensional classification. In this work, we address the issue of high-dimensional clustering by exploiting the general idea of RP ensemble to perform unsupervised classification. Specifically, we generate a set of low dimensional independent random projections and we perform a model-based clustering on each of them. The top B_1 projections, i.e. the ones showing the best grouping structure according to different cluster quality measures, are then selected. The final partition is obtained by aggregating, via consensus, the chosen classifiers. The performances of the method are assessed on a set of both real and simulated data.

References

Cannings, T.I. and Samworth, R.J. (2017). Random projection ensemble classification, *Journal of the Royal Statistical Society - Series B*, **79**, 1–38.

Dimitriadou, E., Weingessel, A. and Hornik, K. (2002). A combination scheme for fuzzy clustering, *International Journal of Pattern Recognition and Artificial Intelligence*, **16**, 901–912.

Hennig, C. and Liao, T.F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, *Journal of the Royal Statistical Society - Series C*, **62**, 309–369.

McLachlan G. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

---

7 Sept.
8.45–10.10
TS 4

Robust patients sub-typing with noisy high-dimensional gene expression data

Pietro Coretto[1], Angela Serra[2] and Roberto Tagliaferri[1]
[1]University of Salerno (Italy); [2]University of Tampere (Finland)

One of the most important research areas in personalised medicine is the discovery of disease sub-types with relevance in clinical applications. This is usually accomplished by exploring gene expression data with unsupervised clustering methodologies. However, microarray data sampling is terribly noisy, and this undermines the possibility to reach scientific consensus on the empirical evidence. This is recognized as a crucial issues, and the research concentrated on the improvement of sampling, and data acquisition techniques. However, even modern microarrays still remain noisy. In this work we propose a new methodology under the commitment to be robust to noise in every step. The proposed

method first computes a robust and sparse correlation matrix (called RSC) of the genes, then decomposes it and projects the patient data onto the first *m* spectral components. After that, a robust and adaptive to noise clustering algorithm (called OTRIMLE) is applied. The algorithm is set up to optimise the separation between survival curves estimated cluster-wise. The method is able to identify clusters that have different omics signatures, and statistically significant differences in survival times. The proposed method obtains a competitive performance in terms of survival separability, even if it uses a single gene expression view compared to the multi-view approach of the state-of-the-art SNF method. Finally, the method is able to find meaningful and interesting biological pathways within groups.

References

Coretto, P., and Hennig, C. (2016). Robust Improper Maximum Likelihood: Tuning, Computation, and a Comparison With Other Methods for Robust Gaussian Clustering, *Journal of the American Statistical Association*, **111**(516), 1648–1659.

Coretto, P. and Hennig, C. (2017). Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering, *Journal of Machine Learning Research*, **18**(142), 1–39.

Coretto, P. and Hennig, C. (2017). **otrimle**: Robust Model-Based Clustering. R package Version 1.1, available at `https://CRAN.R-project.org/package=otrimle`.

Coretto P., Serra A., and Tagliaferri R. (2018). Robust clustering of noisy high-dimensional gene expression data for patients subtyping, to appear on *Bioinformatics*. DOI: `10.1093/bioinformatics/bty502`.

Serra, A., Coretto, P., Fratello, M., and Tagliaferri, R. (2018). Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data, *Bioinformatics*, **34**(4), 625–634.

Infinite Mixtures of Infinite Factor Analysers

7 Sept.
8.45–10.10
TS 4

Keefe Murphy[1], Isobel Claire Gormley[1] and Cinzia Viroli[2]

[1]University College Dublin, Ireland; [2]Università di Bologna, Italy.

Factor-analytic Gaussian mixture models are often employed as a model-based approach to clustering high-dimensional data. Typically, the numbers of clusters and latent factors must be specified in advance of model fitting, and remain fixed. The pair which optimises some model selection criterion is then chosen. For computational reasons, models in which the number of latent factors is common across clusters are generally considered.

Here the infinite mixture of infinite factor analysers (IMIFA) model is introduced. IMIFA employs a Poisson-Dirichlet process prior to facilitate automatic inference of the number of clusters using the stick-breaking construction and a slice sampler. Furthermore, IMIFA employs shrinkage priors to allow cluster specific numbers of factors, automatically inferred via an adaptive Gibbs sampler. IMIFA is presented as the flagship of a family of factor-analytic mixture models, providing flexible approaches to clustering high-dimensional data.

Applications to the benchmark olive oil data set and a manifold learning handwritten digit example illustrate the IMIFA model and its advantageous features: IMIFA obviates the need for model selection criteria, reduces model search and the associated computational burden, improves clustering performance by allowing cluster-specific numbers of factors, and quantifies uncertainty in the numbers of clusters and cluster-specific factors.

References

Murphy K., Gormley I. C., and Viroli C. (2018). Infinite Mixtures of Infinite Factor Analysers, *arXiv preprint*. `https://arxiv.org/pdf/1701.07010.pdf`

Murphy K., Gormley I. C., and Viroli C. (2018). **IMIFA**: Infinite Mixtures of Infinite Factor Analysers and Related Models. R package Version 2.0.0,
available at `https://cran.r-project.org/package=IMIFA`.

Lightning Talk Session 2

Generalized Additive Cluster-Weighted Model

7 Sept.
10.10-10.40
LT 2

Stefano Barberis[1], Salvatore Ingrassia[2] and Giorgio Vittadini[1]

[1]University of Milano Bicocca; [2]University of Catania

An extension of mixture models with random covariates related to the Cluster Weighted Model (Ingrassia et al., 2012) is presented for model-based clustering applications. The Generalized Additive Cluster Weighted Model (GAM-CWM) is a very flexible model, able to capture complex relations between a response variable and a set of covariates in each mixture component. The main difference between models related to the CWM and other mixture models is that in CWM the joint probability $p(x,y)$ of a response variable y and a set of explanatory variables x is modelled in each mixture component rather than the conditional $p(y|x)$. Different extensions of the basic CWM have been proposed including the student-t distribution (Ingrassia et al., 2012), generalized linear mixed CWM (Ingrassia et al., 2015) and the polynomial gaussian CWM (Punzo, 2014).

The theory of Generalized Additive Model (Hastie et al., 1986) extends the generalized linear model precisely with the aim of making it more flexible introducing a sum of smooth functions of covariates in the linear predictor. In the same way the GAM-CWM extends the generalized linear CWM and the polynomial CWM defining a new powerful and very general class of models where the principles of CWM model and the GAM model are combined together.

Maximum likelihood estimates are provided via EM algorithm and model selection is carried out using Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL). With simulated and real data are investigated performances, limits and benefits comparing this model with other mixture models related to it.

References

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions, *Journal of Classification*, **29**(2), 363–401.

Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S. C. (2015). The generalized linear mixed cluster-weighted-model, *Journal of Classification*, **32**(1), 85–113.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**(3), 297–310.

Punzo, A. (2014). Flexible mixture modelling with the polynomial Gaussian cluster-weighted model, *Statistical Modelling*, **14**(3), 257–291.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia, SIAM.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. CRC Press.

7 Sept.
10.10-10.40
LT 2

Averaging via stacking in model-based clustering

Alessandro Casa[1], Luca Scrucca[2] and Giovanna Menardi[1]

[1]Università degli Studi di Padova; [2]Università degli Studi di Perugia

In the framework of density-based clustering, regardless of the specific paradigm chosen, the first step requires to obtain an estimate of the density assumed to describe the data generating mechanism. In the parametric paradigm the estimation task is carried out using mixture models with a clear predominant position taken by Gaussian components. Operationally several models, usually having different number of components and possibly different parametrizations, are estimated and the single best one among them is chosen according to an information criterion. The final partition of the data points is then obtained exploiting the one-to-one correspondence between the groups and the components of the single chosen model. Nevertheless throwing away all the fitted models except for the best one could be sub-optimal since useful information could be lost in the process, especially if the values of the information criterion for the discarded models are close to the one of the selected model. Furthermore, from an inferential point of view, not taking into account the selection step could lead to anti-conservative statements since a source of uncertainty is disregarded. A viable solution to workaround this issue consists in appropriately weight a subset of the top estimated models, as in Bayesian Model Averaging approaches proposed by Wei and McNicholas (2015) and Russell et al (2015). In this work, taking our step from Smyth and Wolpert (1999), we adapt the ideas on which stacking is based on in an unsupervised framework. We consider a density estimator being a convex linear combination of a suitably chosen subset of the fitted models where the weights of the combination are estimated via maximum likelihood. In order to avoid weighting excessively highly parametrized models, possibly prone to overfitting the data, we follow a penalized likelihood approach and examine different penalizations. Since by averaging over different models the correspondence between groups and mixture components is lost, we discuss possible solutions, coherent with the density-based framework, to obtain partitions starting from the estimated density.

References

Russell, N., Murphy T. B., and Raftery A. E. (2015). Bayesian model averaging in model-based clustering and density estimation, *arXiv preprint arXiv:1506.09035.*

Smyth, P. and Wolpert, D. (1999). Linearly combining density estimators via stacking, *Machine Learning*, **36**, 59–83.

Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering, *Advances in Data Analysis and Classification*, **9**(2), 197–217.

Subspace Clustering for the Finite Mixture of Generalized Hyperbolic Distributions

7 Sept.
10.10-10.40
LT 2

Nam-Hwui Kim, Ryan P. Browne
University of Waterloo

The finite mixture of Generalized Hyperbolic distributions is a flexible model for clustering, but its large number of parameters for estimation, especially in high dimensions, can make it computationally expensive to work with. In light of this issue, we provide an extension of the subspace clustering technique developed for finite Gaussian mixtures to that of Generalized Hyperbolic distribution. The methodology will be demonstrated with numerical experiments.

References

Kim N., Browne R. P. (2018). Subspace Clustering for the Finite Mixture of Generalized Hyperbolic Distributions, *Unpublished Paper*.

Learning the number of components and data clusters in Bayesian finite mixture models

7 Sept.
10.10-10.40
LT 2

Gertraud Malsiner-Walli[1], Sylvia Frühwirth-Schnatter[1] and Bettina Grün[2]
[1]WU Vienna University of Business and Economics; [2]Johannes Kepler Universität Linz

Bayesian cluster analysis aims at inferring the number of data clusters present in a data set using either finite or infinite mixture models. In Bayesian finite mixture models usually a one-to-one relationship between components and data clusters is assumed. The number of components can be determined by comparing the marginal likelihoods of the potential models (Frühwirth-Schnatter, 2006) or by approximating the posterior of the number of components using different methods, e.g., reversible jump Markov chain Monte Carlo (Richardson and Green, 1997), Markov birth-and-death process sampling (Stephens, 2000) or the Jain-Neal split-merge sampler (Miller and Harrison, 2018).

We propose to explicitly distinguish between the number of data clusters and components and purposely allow for more components than data clusters. We extend the standard approach by including priors on the number of components and on the parameters of the Dirichlet distribution for the mixture weights. This allows us to approximate the posteriors of the number of components as well as data clusters using Gibbs sampling techniques. The performance of the proposed sampling technique is compared to previously proposed approaches. The additional flexibility gained by suitably selecting the parameters of the hyperpriors is highlighted and guidance for their choice provided.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

Miller, J.W. and Harrison, M.T. (2018) Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 1–17. DOI: 10.1080/

01621459.2016.1255636.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, **59**(4), 731–792.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *The Annals of Statistics*, **28**(1), 40–74.

---

7 Sept.
10.10-10.40
LT 2

Constraining kernel estimators in semiparametric copula-based mixture models

Gildas Mazo[1] and Yaroslav Averyanov[2]

[1]MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France;
[2]MODAL, Inria Lille Nord Europe, Lille, France

We present a novel algorithm for performing inference and/or clustering in semiparametric copula-based mixture models. The algorithm replaces the standard kernel density estimator by a weighted version that permits to take into account the constraints put on the underlying marginal densities. Lower misclassification error rates and better estimates are obtained on simulations. The pointwise consistency of the weighted kernel density estimator is established under an assumption on the rate of convergence of the sample maximum.

References

Mazo G. and Averyanov Y. (2018). Constraining kernel estimators in semiparametric copula mixture models, *Submitted*.

Mazo G. (2017). A semiparametric and location-shift copula-based mixture model, *Journal of Classification*, **34**(3), 444–464.

---

7 Sept.
10.10-10.40
LT 2

Gaussian Parsimonious Clustering Models with Covariates

Keefe Murphy and T. Brendan Murphy

School of Mathematics and Statistics, University College Dublin, Ireland

In model-based clustering methods using finite Gaussian mixture models, clustering is typically implemented on response variables only and reference is not made to associated covariates until the structure of the clustering is investigated in light of information present in the covariates. It is desirable to have covariates incorporated into the clustering process and not only into the interpretation of the clustering structure and model parameters, in order to exploit clustering capabilities and provide richer insight into the type of observation which characterises each cluster.

The mixture of experts model provides one such framework: it extends the mixture model to accommodate the presence of covariates by modelling the parameters of the

mixture model as functions of fixed covariates. However, for Gaussian responses, the flexibility afforded by parsimonious parameterisations of the component covariance matrices have to date been lacking in the mixture of experts context.

We consider model-based clustering methods with constrained covariance structures that account for external information available in the presence of covariates by proposing the general MoEClust family of models; these models allow the distribution of the latent cluster membership variable and/or the distribution of the response variable to depend on covariates. This family of models address the aim of including covariates in Gaussian parsimonious clustering models or, equivalently, the aim of incorporating parsimonious covariance structures into the framework of Gaussian mixtures of experts. The MoEClust models demonstrate significant improvement from both perspectives in applications to univariate and multivariate data sets. A software implementation for the full suite of MoEClust models is also introduced.

References

Murphy K. and Murphy T. B. (2017). Parsimonious Model-Based Clustering with Covariates, *arXiv preprint*. `https://arxiv.org/pdf/1711.05632.pdf`

Murphy K. and Murphy T. B. (2017). **MoEClust**: Parsimonious Model-Based Clustering with Covariates. R package Version 1.1.0,
available at `https://cran.r-project.org/package=MoEClust`

Model-based Clustering with R-vine copulas

7 Sept.
10.10-10.40
LT 2

Marta Nai Ruscone[1] and Thomas Brendan Murphy[2]

[1]LIUC Università Cattaneo; [2]University College Dublin

Finite mixtures are applied to perform model-based clustering of multivariate data. Existing models are not flexible enough for modeling the dependence of multivariate data since they rely on potentially undesirable correlation restrictions to be computationally tractable. We discuss a model-based clustering method via R-vine copula to understand the complex and hidden dependence patterns in correlated multivariate data. One of the advantages of this approach is that it accounts for the tail asymmetry of the data by using blocks of asymmetric bivariate copulas. We use real datasets to illustrate the proposed procedure.

References

Banfield J. and Raftery A.(1993). *Model-based Gaussian an non-Gaussian clustering*. Biometrics, **49**, 803-821

Kosmidis I. and Karlis D. (2016). *Model-based clustering using copulas with applications*. Statistics and Computing, **26**, 1079–1099.

McLachlan G. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

Keynote Lecture 5

7 Sept.
11.10-12.05
KL 5

Deep Gaussian Mixture Models

Cinzia Viroli[1] and Geoffrey J. McLachlan[2]
[1]University of Bologna; [2]University of Queensland

In the recent years, there has been an increasing interest on deep learning for classification and clustering tasks. Deep learning is a hierarchical inference method formed by subsequent multiple layers of learning able to more efficiently describe complex relationships. In this work, Deep Gaussian Mixture Models are introduced and discussed. A Deep Gaussian Mixture model is a network of multiple layers of latent variables, where, at each layer, the variables follow a mixture of Gaussian distributions. Thus, the deep mixture model consists of a set of nested mixtures of linear models, which globally provides a nonlinear model able to describe the data in a very flexible way. In order to avoid overparameterized solutions, dimension reduction by factorial models can be applied at each layer of the architecture thus resulting in deep mixtures of factor analysers.

References

Baek, J., McLachlan, G., Flack, L. (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(7), 1298–1309.

McLachlan G.J. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

Tang, Y., Hinton, G. E., Salakhutdinov, R. (2012). Deep mixtures of factor analysers. In Langford, J., Pineau, J. (eds.) Proceedings of the 29th International Conference on Machine Learning (ICML-12), New York, NY, USA, pp. 505–512.

Viroli, C. and McLachlan, G. J. (2018). *Statistics and Computing*, forthcoming.

Keynote Lecture 6

Artificial Intelligence and Media Content

7 Sept.
12.05-13.00
KL 6

Nello Cristianini
University of Bristol

It is easy to focus on methods and accuracy when designing classification and clustering algorithms, but as soon as we apply them to people, there is a range of other considerations that we need to keep in mind.

The combination of machine learning and big data has enabled us to create a new generation of artificial intelligence (AI), and now we interact with its applications daily. The strategic position occupied by AI agents within our global information infrastructure means that they are in the position to observe a large portion of our activities, learning from them, but also creates a new type of risk, including the possibility of surveillance and manipulation of user behaviour, unintended bias, lack of transparency, etc.

Based on the details of how AI has emerged from the combination of machine learning and this unified data infrastructure, we can understand recent reports that have raised concerns, from fake news to psychometric election targeting, from criminal justice applications to dynamic pricing in insurance based on social media content. By reviewing the way this technology works, we can put these reports in a context, and look ahead at forthcoming challenges. Importantly, we can plan future regulation of this strategic sector.

Talk Session 5: Issues in hidden Markov models

Consistent estimation of the filtering and smoothing probabilities in non parametric hidden Markov models

7 Sept.
14.00–15.25
TS 5

Yohann De Castro, Sylvain Le Corff and Elisabeth Gassiat

Université Paris-Sud

We consider the filtering and smoothing recursions in nonparametric finite state space hidden Markov models (HMMs) when the parameters of the model are unknown and replaced by estimators. We provide an explicit and time uniform control of the filtering and smoothing errors in total variation norm as a function of the parameter estimation errors. We prove that the risk for the filtering and smoothing errors may be uniformly upper bounded by the L^1-risk of the estimators. It has been proved very recently that statistical inference for finite state space nonparametric HMMs is possible. We study how the recent spectral methods developed in the parametric setting may be extended to the nonparametric framework and we give explicit upper bounds for the L^2-risk of the nonparametric spectral estimators. In the case where the observation space is compact, this provides explicit rates for the filtering and smoothing errors in total variation norm. The performance of the spectral method is assessed with simulated data for both the estimation of the (nonparametric) conditional distribution of the observations and the estimation of the marginal smoothing distributions.

References

Gassiat E., Cleynen A. , Robin S. (2016). Inference in finite state space non parametric hidden Markov models and applications, *Statistics and Computing*, **26**(1-2), 61-71.

De Castro Y., Gassiat E., Lacour C. (2017). Minimax adaptive estimation of non-parametric hidden Markov models, *Journal of Machine Learning Research*, **17**, 111 (43p).

De Castro Y., Gassiat E., Le Corff S. (2017). Consistent estimation of the filtering and smoothing distributions in non-parametric hidden Markov models. *IEEE Trans. Info. th*, **63** (8), 4758-4777.

Time-dependent nonparametric latent variable modeling

7 Sept.
14.00–15.25
TS 5

Hajo Holzmann, Anna Leister, Grigory Alexandrovich and Ann-Kristin Becker
Philipps-Universität Marburg, Germany

We consider identification and estimation in dynamic latent variable models with state-dependent distributions from a nonparametric class.

First, we investigate general finite-state hidden Markov models and obtain nonparametric identification of the parameters as well as the order of the Markov chain if the transition probability matrices have full-rank and are ergodic, and if the state-dependent distributions are all distinct, but not necessarily linearly independent. Based on this identification result, we develop nonparametric maximum likelihood estimation theory. Numerical properties of the estimates as well as of nonparametric goodness of fit tests are investigated in a simulation study.

Second, we consider the dynamic stochastic block model as recently introduced in Matias and Miele (2017), and obtain nonparametric identification in case of binary, finitely weighted and general edge states. We formulate conditions on the true parameters which guarantee actual point identification instead of mere generic identification, and which also lead to novel conclusions in the static case. We also give numerical illustrations via the variational EM algorithm in simulation settings covered by our identification analysis.

References

Alexandrovich, G., Holzmann, H., Leister, A. (2016) Nonparametric identification and maximum likelihood estimation of hidden Markov models. *Biometrika*, **103**, 423âĂŞ434.

Becker, A.-K, Holzmann, H. (2018) Nonparametric identification in the dynamic stochastic block model. *Preprint*

Matias, C. and V. Miele (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, **79**, 1119-1141.

Time-specific clustering via rectangular latent Markov models, with an analysis of the well being of nations

7 Sept.
14.00–15.25
TS 5

Alessio Farcomeni[1], Gordon Anderson[2], Maria Grazia Pittau[1] and Roberto Zelli[1]
[1]Sapienza - University of Rome; [2]University of Toronto

One limitation of several longitudinal model-based clustering methods is that the number of groups is fixed over time, apart from (mostly) heuristic approaches. In this work we propose a latent Markov model admitting variation in the number of latent states at each time period. The consequence is that (i) subjects can switch from one group to another at each time period and (ii) the number of groups can change at each time period. Clusters can merge, split, or be re-arranged. For a fixed sequence of the number of groups, inference is carried out through maximum likelihood, using forward-backward recursions taken from the hidden Markov literature once an assumption of local independence is granted. A penalized likelihood form is introduced to simultaneously choose an

optimal sequence for the number of groups and cluster subjects. The penalized likelihood is optimized through a novel Expectation-Maximization-Markov-Metropolis algorithm. The work is motivated by an analysis of the progress of well-being of nations, as measured by the three dimensions of the Human Development Index over the last 25 years. The main findings are that (i) transitions among nation clubs are scarce, and mostly linked to historical events (like dissolution of USSR or war in Syria) and (ii) there is mild evidence that the number of clubs has shrunk over time, where we have four clusters before 2005 and three afterwards. In a sense, nations are getting more and more polarized with respect to standards of well-being. Non-optimized R code for general implementation of the method, data used for the application, code and instructions for replicating the simulation studies and the real data analysis are available at `https://github.com/afarcome/LMrectangular`.

Lightning Talk Session 3

The analysis of high frequency financial price changes

7 Sept.
15.25–15.50
LT 3

Leopoldo Catania[1], Roberto Di Mari[2] and Paolo Santucci de Magistris[3]
[1]Aarhus University and CREATES; [2]University of Catania; [3]Luiss University and CREATES

High frequency price changes of financial assets are usually assumed to follow a distribution defined over a continuous support, with timeâĂŤvarying parameters. However, in the real world high frequency prices, and thus their changes, are intrinsically discrete variable. We start from this empirical evidence to develop a new model able to describe the dynamic properties of a multivariate time–series of high frequency price changes. Emphasis is given to the large presence of zeroes that characterize these series. We assume the existence of two independent unobserved latent processes which determine the price changes' dynamic properties and the zeroes occurrences. Given the probabilistic structure embedded in our modelling framework we analyze the different sources of this large amount of zeroes as for example: absence of news, same magnitude of positive and negative news, and periods of market illiquidity. Furthermore, multivariate dynamic properties driving the different sources of zeroes between several assets are analyzed.

Multi-Resolution Bagging for Ensemble Classification

7 Sept.
15.25–15.50
LT 3

Majed El Helou, Rawan Chanouha, Hazem Hajj
Faculty of Engineering and Architecture, AUB, Beirut, Lebanon

Recent years have witnessed significant advances in the performance of classification algorithms, where ensemble techniques have had their fair share of success. However, some problems are yet to be solved, one of which is known as the imbalanced dataset classification problem [1]. Another challenge for ensemble techniques is the need for richer and larger ensembles (different classifier model definitions, different training strategies etc.). In this paper, we propose a novel approach to tackle the difficulties faced with imbalanced datasets. The solution we present is inspired by the well-known bagging approach [2]. Our algorithm leverages a hierarchically-structured bagging technique on the training data, forming models that are experts on different degrees of resolution and in different subregions of the feature space. This strategy is able to solve ambiguities at different levels of scale in the feature space. It is therefore able to correctly focus on minority classes. Very importantly, the underlying architecture permits the creation of a rich set of classifiers from a single model, which can then be integrated into any large ensemble. Our approach improves classification on various datasets used in the literature [1, 3]. In addition, we compare our ensemble results to the ones obtained with our base classifier as a an illustration of the ensembling capabilities.

References

S. Visa and A. Ralescu (2005). Issues in mining imbalanced data sets - a review paper. In Proc. of Midwest Artificial Intelligence and Cognitive Science Conference, 67–73.

M. Galar, A. FernÃąndez, E. Barrenechea, H. Bustince, F. Herrera. (2012). A review on ensembles for class imbalance problem: bagging, boosting and hybrid based approaches, *IEEE Transactions on Systems, Man, and Cybernetics - part C: Applications and Reviews*, **42**(4), 463–484.

Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. (2007). Cost-sensitive boosting for classification of imbalanced data, *Pattern Recognition*, **40**(12), 3358–3378.

---

7 Sept.
15.25–15.50
LT 3

Improving clustering assessment through supervised classification modeling

Mario Fordellone[1] and Maurizio Vichi[1]

[1]Sapienza University of Rome

In the unsupervised classification techniques, clusters of homogeneous objects are detected by means of a set of features measured (observed) on a set of objects without knowing the membership of objects to clusters. In these applications the aim is to discover the heterogeneity structure of the data. Often, techniques based on separability and homogeneity criteria of the groups are used, giving *a priori* the number of groups. In the unsupervised classification models the principal approaches of cluster analysis are: Connectivity-based clustering better known as hierarchical clustering, Centroid-based clustering, Distribution-based clustering, Density-based clustering, and many other parametric and non-parametric techniques.

Conversely, supervised classification is based on the idea to forecast the membership of new objects (output) based on a set of features (inputs) measured on a training set of objects for which the membership to clusters is known. Therefore, in these applications the aim is to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs. Usually, a subsample (training) that is representative of specific groups is selected and then this model is used as references for the classification of new (unobserved) other objects. Training sets are selected based on the knowledge of the user. In the supervised classification models we have Artificial neural network, Naive Bayes classifier, Nearest Neighbor Algorithm, decision trees, logistic regression, generalized linear models, and many other parametric and non-parametric techniques are included.

Now, we know that in the unsupervised classification field, we have important issues that could drastically influence results: the unknown number of clusters, the selection of variables that most contribute to clustering, the final assessment of clusters. In other words, all the taken decisions to address the study can lead different results and then, each single decision become crucial on the aim of our study and should be tested.

In this work, we propose an algorithm based on the use of supervised classification modeling. In particular, we will prove that by using of supervised classification techniques we have effective inferential tools for choosing the number of clusters, selecting the most important variables for the clustering and assessing the quality of clusters. An application on both toy data and real data will be discussed.

References

Agresti, A., Kateri, M. (2011). Categorical data analysis. In International Encyclopedia of Statistical Science, pp. 206–208.

Aloise, D., Deshpande, A., Hansen, P., Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, **75**(2), pp. 245-248.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, **3**(1), 1–27.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, **10**(7), 1895-1923.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Miscellaneous clustering methods. Cluster Analysis, 5th Edition, pp. 215–255.

Filipovych, R., Resnick, S. M., & Davatzikos, C. (2011). Semi-supervised cluster analysis of imaging data. *NeuroImage*, **54**(3), pp. 2185-2197.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, **1**(14), 281–297.

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, **44**(1), 23–34.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 411–423.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *SIAM Review*, **37**(4), 531–551.

Simulating mixtures of non-normal multivariate data with fixed cluster overlap in FSDA.

7 Sept.
15.25–15.50
LT 3

Marco Riani[2], Francesca Torti[1] and Domenico Perrotta[1]

[1] European Commission, Joint Research Centre [2] University of Parma, Italy

The linear combination or product of independent random variables with known distributions and characteristic functions, in general cannot be derived in closed form. Witkovsky (2018) proposes to compute the characteristic function of the combination/product through the numerical, rather than analytical, inversion of the characteristic function of the random variables. To this end, he developed the MATLAB toolbox `CharFunTool`. A similar library, **CharFun**, is also available in R (Simkova, 2017).

In the case of linear combinations of non central χ^2 random variables, Davies (1980) also proposed a numerical solution based on the numerical inversion of the characteristic

function. This solution was adopted by Melnykov et al. (2012) in his `MixSim` framework for mixtures generation, which we have ported and extended in our MATLAB `FSDA` toolbox (Riani et al., 2015).

We now propose to extend the current `FSDA` implementation of `MixSim` to the linear combinations of other relevant distributions through the use of the `CharFunTool` framework.

References

Davies, R. B. (1980). Numerical Computation Cumulative Distribution Function and Probability Density Function from Characteristic Function, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29**(3), 323-333.

Melnykov V., Chen Chen W. and Maitra R.(2012). **MixSim**: An R Package for Simulating Data to Study Performance of Clustering Algorithms, *Journal of Statistical Software*, **51**(12), 1–25.

Riani M., Cerioli A., Perrotta D. and Torti F. (2015). Simulating mixtures of multivariate data with fixed cluster overlap in FSDA, *Advances in Data Analysis and Classification*, **9**(4), 461–481.

Simkova L. (2017). Numerical Computation Cumulative Distribution Function and Probability Density Function from Characteristic Function. `https://cran.r-project.org/web/packages/CharFun/CharFun.pdf`.

Witkovsky V. (2018). Exact distribution of selected multivariate test criteria by numerical inversion of their characteristic functions, *To appear*.

Parsimonious models in matrix data mixture modeling

Shuchismita Sarkar[1], Xuwen Zhu[2], Volodymyr Melnykov[1] and Salvatore Ingrassia[3]

[1]University of Alabama; [2]University of Louisville; [3] UniversitÃă di Catania

Finite mixture modeling is a popular technique for capturing heterogeneity in data. Although the vast majority of the theory developed in this area up to date deals with vector-valued data, some recent advancements have been made to expand the concept to matrix-valued data, for example, by means of matrix Gaussian mixture models (Viroli 2011; Melnykov and Zhu 2018). Unfortunately, matrix mixtures tend to suffer from the overparameterization issue due to a high number of parameters involved in the model. As a result, this may lead to problems such as overfitting and mixture order underestimation. One possible approach of addressing the overparameterization issue that has proven to be effective in the vector-valued framework is to consider various parsimonious models. One of the most popular classes of parsimonious models is based on the spectral decomposition of covariance matrices (Bandield and Raftery 1993; Celeux and Govaert 1995). In this study, an attempt to generalize this class and make it applicable in the matrix setting is made. Estimation procedures are thoroughly discussed for all models considered. The application of the proposed methodology is illustrated on a real-life data set.

References

Viroli C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data, *Statistics and Computing*, **21**, 511–522.

Melnykov, V. and Zhu, X. (2018). On model-based clustering of skewed matrix data, under review at *Journal of Multivariate Analysis*.

Banfield J. D. and Raftery A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Celeux G. and Govaert G. (1995). Gaussian parsimonious clustering models, *Computational Statistics and Data Analysis*, **28**, 781–793.

The Delta Machine: Binary Data Classification

7 Sept.
15.25–15.50
LT 3

Zdenek Sulc[1], Beibei Juan[2] and Mark de Rooij[2]

[1]University of Economics, Prague, Czechia; [2] Leiden University, The Netherlands

The Delta Machine is a statistical tool for the supervised classification with similar aims as the logistic regression or the support vector machines. Based on (dis)similarities between profiles of the observations to profiles of a representation set consisting of prototypes, it predicts the value of a class variable. One can choose one from four similarity measures for quantitative data (the Euclidean the squared Euclidean distances, the Exponential decay and the Gaussian decay functions) and the Gower coefficient for the mixed-type data. However, when dealing with the binary data, one can only use the Gower coefficient in the R function daisy from the cluster package (Maechler et al., 2018). In order to improve the binary data analysis, 15 similarity measures for binary data, such as the Jaccard, Dice or Hamann coefficients, see, e.g. Warrens (2008), were added to the Delta Machine. They offer different ways to express the similarity between the categories. The aim is to compare the classification performance of the newly added similarity measures with the original ones in binary-coded datasets. The results are also compared with commonly used classification methods, namely, the logistic regression and support vector machines described e.g. in James et al., (2013). For the comparison, the accuracy and the ROC analysis statistics obtained by the repeated cross-validations are used. The variability of these statistics is also analyzed using box-plots. The preliminary results show that the different similarity measures perform similarly concerning the classification performance, but the numbers of the used parameters in the models differ substantially.

References

James, G., Witten, D. Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag, New York.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2018). **cluster**: Cluster Analysis Basics and Extensions. R package version 2.0.7-1, available at `https://CRAN.R-project.org/package=cluster`.

Warrens, M.J. (2008). Similarity Coefficients for Binary Data. Ph.D. thesis, University of Leiden.

---

7 Sept.
15.25–15.50
LT 3

Zero-and-one inflated mixtures for loss given default

Salvatore D. Tomarchio and Antonio Punzo
University of Catania

The peculiar characteristics of the empirical loss given default distribution (LGD) call for more flexible models. In fact, it has support $[0,1]$, exhibit an excess of zeros and ones, and it is generally multimodal on $(0,1)$. Thus, we introduce a zero-and-one inflated mixture where a three level multinomial model is considered for the membership of the LGD values to the sets $\{0\}$, $(0,1)$ and $\{1\}$, while a finite mixture of distributions is used on $(0,1)$. To allow for more flexible shapes on $(0,1)$, besides considering distributions already defined on $(0,1)$, we used distributions defined on $(-\infty,\infty)$ mapped on $(0,1)$ via the inverse-logit transformation. The models are then fitted on two real LGDs datasets, one from an European Bank and the other from the Bank of Italy, and compared according to information criteria and goodness-of-fit.

References

Gurtler, M., & Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*, **37**(7), 2354–2366.

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, **51**(1), 111–126.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: using gamlss in R.* CRC Press.

Schuermann, T. (2004). What do we know about Loss Given Default?.

De Oliveira Jr, M. R., Louzada, F., de Araujo Pereira, G. H., Moreira, F. F., & Calabrese, R. (2015). Inflated mixture models: Applications to multimodality in loss given default.

Talk Session 6: Recent developments in clustering of matrix data

Matrix Transformation Mixture Modeling

7 Sept.
16.50–18.15
TS 6

Volodymyr Melnykov and Xuwen Zhu
University of Alabama, University of Louisville

The existing finite mixture modeling and model-based clustering literature focuses primarily on the analysis of multivariate data observed in the form of vectors, with each element representing a specific feature. In this setting, multivariate Gaussian mixture models have been the most commonly used. Due to severe modeling issues observed when normal components cannot provide adequate fit to groups, much attention has been paid to developing models capable of accounting for skewness in data. In our work, we target the problem of mixture modeling with components that can handle skewness in matrix-valued data. The proposed developments open a wide range of possible modeling capabilities, with numerous applications, as illustrated in this paper. A novel matrix mixture model is proposed. Its skewness parameters enjoy appealing interpretability. The corresponding estimation procedure and various ways of parameterization are discussed. Comprehensive simulation studies and applications to real-life datasets illustrate the efficiency of the proposed developments, supported by good results.

References

Melnykov V., Zhu X. (2018). On model-based clustering of skewed matrix data, *Journal of Multivariate Analysis*.

McLachlan G. and Peel D. (2000). *Finite Mixture Models*. Wiley, New York.

Mixtures of Matrix Variate Bilinear Factor Analyzers

7 Sept.
16.50–18.15
TS 6

Michael P. B. Gallaugher, Paul D. McNicholas
McMaster University, Ontario, Canada

Over the years data has become increasingly higher dimensional, which has prompted an increased need for dimension reduction techniques. This is perhaps especially true for clustering (unsupervised classification) as well as semi-supervised and supervised classification. Although dimension reduction in the area of clustering for multivariate data has been quite thoroughly discussed in the literature, there is relatively little work in the area of three way, or matrix variate, data. A mixture of matrix variate bilinear factor analyzers (MMVBFA) model is developed for use in clustering high-dimensional matrix variate data. This work can be considered both the first matrix variate bilinear factor analyzers model as well as the first MMVBFA model. Parameter estimation is discussed, and the MMVBFA model is illustrated using simulated and real data.

---

7 Sept.
16.50–18.15
TS 6

Model-based clustering of tensor data

Shuchismita Sarkar[1], Volodymyr Melnykov[1] and Xuwen Zhu[2]
[1]The University of Alabama; [2]University of Louisville

The majority of existing mixture modeling and model-based clustering techniques are designed for the analysis of multivariate data. However, there are situations when the data are presented in more complex forms such as matrices or tensors. We extend the recent work in the matrix-valued setting by developing a methodological framework for mixture modeling and model-based clustering of tensor-valued observations and illustrate it on the data set containing self-reported salaries of faculty employed in American institutions. The information is available over thirteen academic years and grouped by the faculty rank and gender. The illustrative study aims at identifying salary patterns at American universities and investigating the potential sources of the variability in salary.

References

Gallaugher M.P.B., McNicholas P.D. (2018). Finite Mixtures of Skewed Matrix Variate Distributions, *Pattern Recognition*, **80**, 83–93.

Melnykov V., Zhu X. (2018). On Model-Based Clustering of Skewed Matrix Data, *Journal of Multivariate Analysis*, **176**, 181–194.

Viroli, C. (2011) Finite Mixtures of Matrix Normal Distributions for Classifying Three-Way Data, *Statistics and Computing*, **21**(4), 511–522.

Author Index