

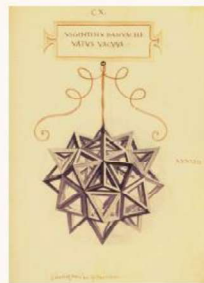
DIPARTIMENTO DI ECONOMIA E GIURISPRUDENZA
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE



CLADAG 2019

11-13 SEPTEMBER 2019
CASSINO

```
def business_model()  
  arr=[ ]  
  items="a,b,c"  
  items>>arr  
  return arr  
end
```



Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors

12-TH SCIENTIFIC MEETING
CLASSIFICATION AND DATA ANALYSIS



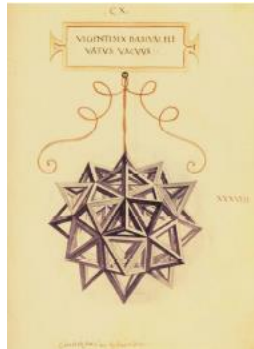
Società
Italiana di
Statistica

© CC – Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6



CLADAG 2019
Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors

2019

Contents

Keynotes lectures

Unifying data units and models in (co-)clustering <i>Christophe Biernacki</i>	3
Statistics with a human face <i>Adrian Bowman</i>	4
Bayesian model-based clustering with flexible and sparse priors <i>Bettina Grün</i>	5
Grinding massive information into feasible statistics: current challenges and opportunities for data scientists <i>Francesco Mola</i>	6
Statistical challenges in the analysis of complex responses in biomedicine <i>Sylvia Richardson</i>	7

Invited and contributed sessions

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models <i>Timo Adam, Roland Langrock, Thomas Kneib</i>	8
Some issues in generalized linear modeling <i>Alan Agresti</i>	12
Assessing social interest in burnout using functional data analysis through google trends <i>Ana M. Aguilera, Francesca Fortuna, Manuel Escabias</i>	16
Measuring equitable and sustainable well-being in Italian regions: a non- aggregative approach <i>Leonardo Salvatore Alaimo, Filomena Maggino</i>	20
Bootstrap inference for missing data reconstruction <i>Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna</i>	22
Archetypal contour shapes <i>Aleix Alcacer, Irene Epifanio, M. Victoria Ibáñez, Amelia Simó</i>	26

Random projections of variables and units <i>Laura Anderlucci, Roberta Falcone, Angela Montanari</i>	30
Sparse linear regression via random projections ensembles <i>Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, Angela Montanari</i>	34
High-dimensional model-based clustering via random projections <i>Laura Anderlucci, Francesca Fortunato, Angela Montanari</i>	38
Multivariate outlier detection in high reliability standards fields using ICS <i>Aurore Archimbaud, Klaus Nordhausen, Anne Ruiz-Gazen</i>	42
Evaluating the school effect: adjusting for pre-test or using gain scores? <i>Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini</i>	45
ACE, AVAS and robust data transformations <i>Anthony Atkinson</i>	49
Mixtures of multivariate leptokurtic Normal distributions <i>Luca Bagnato, Antonio Punzo, Maria Grazia Zoia</i>	53
Detecting and interpreting the consensus ranking based on the weighted Kemeny distance <i>Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio</i>	57
Predictive principal components analysis <i>Simona Balzano, Maja Bozic, Laura Marcis, Renato Salvatore</i>	61
Flexible model-based trees for count data <i>Federico Banchelli</i>	63
Euclidean distance as a measure of conformity to Benford's law in digital analysis for fraud detection <i>Mateusz Baryła, Józef Pociecha</i>	67
The evolution of the purchase behavior of sparkling wines in the Italian market <i>Francesca Bassi, Fulvia Pennoni, Luca Rossetto</i>	71
Modern likelihood-frequentist inference at work <i>Ruggero Bellio, Donald A. Pierce</i>	75
Ontology-based classification of multilingual corpuses of documents <i>Sergey Belov, Salvatore Ingrassia, Zoran Kalinić, Paweł Lula</i>	79
Modeling heterogeneity in clustered data using recursive partitioning <i>Moritz Berger, Gerhard Tutz</i>	83

Mixtures of experts with flexible concomitant covariate effects: a bayesian solution <i>Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy, Saverio Ranciati</i>	87
Sampling properties of an ordinal measure of interrater absolute agreement <i>Giuseppe Bove, Pier Luigi Conti, Daniela Marella</i>	91
Tensor analysis can give better insight <i>Rasmus Bro</i>	95
A boxplot for spherical data <i>Davide Buttarazzi, Giuseppe Pandolfo, Giovanni C. Porzio, Christophe Ley</i>	97
Machine learning models for forecasting stock trends <i>Giacomo Camba, Claudio Conversano</i>	99
Tree modeling ordinal responses: CUBREMOT and its applications <i>Carmela Cappelli, Rosaria Simone, Francesca Di Iorio</i>	103
Supervised learning in presence of outliers, label noise and unobserved classes <i>Andrea Cappelozzo, Francesca Greselin, Thomas Brendan Murphy</i>	104
Asymptotics for bandwidth selection in nonparametric clustering <i>Alessandro Casa, José E. Chacón, Giovanna Menardi</i>	108
Foreign immigration and pull factors in Italy: a spatial approach <i>Oliviero Casacchia, Luisa Natale, Francesco Giovanni Truglia</i>	112
Dimensionality reduction via hierarchical factorial structure <i>Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria</i>	116
Likelihood-type methods for comparing clustering solutions <i>Luca Coraggio, Pietro Coretto</i>	120
Labour market analysis through transformations and robust multilevel models <i>Aldo Corbellini, Marco Magnani, Gianluca Morelli</i>	124
Modelling consumers' qualitative perceptions of inflation <i>Marcella Corduas, Rosaria Simone, Domenico Piccolo</i>	128
Noise resistant clustering of high-dimensional gene expression data <i>Pietro Coretto, Angela Serra, Roberto Tagliaferri</i>	132
Classify X-ray images using convolutional neural networks <i>Federica Crobu, Agostino Di Ciaccio</i>	136

A compositional analysis approach assessing the spatial distribution of trees in Guadalajara, Mexico <i>Marco Antonio Cruz, Maribel Ortego, Elisabet Roca</i>	140
Joining factorial methods and blockmodeling for the analysis of affiliation networks <i>Daniela D'Ambrosio, Marco Serino, Giancarlo Ragozini</i>	142
A latent space model for clustering in multiplex data <i>Silvia D'Angelo, Michael Fop</i>	146
Post processing of two dimensional road profiles: variogram scheme application and sectioning procedure <i>Mauro D'Apuzzo, Rose-Line Spacagna, Azzurra Evangelisti, Daniela Santilli, Vittorio Nicolosi</i>	150
A new approach to preference mapping through quantile regression <i>Cristina Davino, Tormod Naes, Rosaria Romano, Domenico Vistocco</i>	154
On the robustness of the cosine distribution depth classifier <i>Houyem Demni, Amor Messaoud, Giovanni C. Porzio</i>	158
Network effect on individual scientific performance: a longitudinal study on an Italian scientific community <i>Domenico De Stefano, Giuseppe Giordano, Susanna Zaccarin</i>	162
Penalized vs constrained maximum likelihood approaches for clusterwise linear regression modelling <i>Roberto Di Mari, Stefano Antonio Gattone, Roberto Rocci</i>	166
Local fitting of angular variables observed with error <i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	170
Quantile composite-based path modeling to estimate the conditional quantiles of health indicators <i>Pasquale Dolce, Cristina Davino, Stefania Taralli, Domenico Vistocco</i>	174
AUC-based gradient boosting for imbalanced classification <i>Martina Dossi, Giovanna Menardi</i>	178
How to measure material deprivation? A latent Markov model based approach <i>Francesco Dotto</i>	182
Decomposition of the interval based composite indicators by means of biclustering <i>Carlo Drago</i>	186
Consensus clustering via pivotal methods <i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	190

Robust model-based clustering with mild and gross outliers <i>Alessio Farcomeni, Antonio Punzo</i>	194
Gaussian processes for curve prediction and classification <i>Sara Fontanella, Lara Fontanella, Rosalba Ignaccolo, Luigi Ippoliti, Pasquale Valentini</i>	198
A new proposal for building immigrant integration composite indicator <i>Mario Fordellone, Venera Tomaselli, Maurizio Vichi</i>	199
Biodiversity spatial clustering <i>Francesca Fortuna, Fabrizio Maturo, Tonio Di Battista</i>	203
Skewed distributions or transformations? Incorporating skewness in a cluster analysis <i>Michael Gallagher, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu</i>	207
Robust parsimonious clustering models <i>Luis Angel Garcia-Escudero, Agustin Mayo-Isacar, Marco Riani</i>	208
Projection-based uniformity tests for directional data <i>Eduardo García-Portugués, Paula Navarro-Esteban, Juan Antonio Cuesta-Albertos</i>	212
Graph-based clustering of visitors' trajectories at exhibitions <i>Martina Gentilin, Pietro Lovato, Gloria Menegaz, Marco Cristani, Marco Minozzo</i>	214
Symmetry in graph clustering <i>Andreas Geyer-Schulz, Fabian Ball</i>	218
Bayesian networks for the analysis of entrepreneurial microcredit: evidence from Italy <i>Lorenzo Giammei, Paola Vicard</i>	222
The PARAFAC model in the maximum likelihood approach <i>Paolo Giordani, Roberto Rocci, Giuseppe Bove</i>	226
Structure discovering in nonparametric regression by the GRID procedure <i>Francesco Giordano, Soumendra Nath Lahiri, Maria Lucia Parrella</i>	230
A microblog auxiliary part-of-speech tagger based on bayesian networks <i>Silvia Golia, Paola Zola</i>	234
Recent advances in model-based clustering of high dimensional data <i>Isobel Claire Gormley</i>	238
Tree embedded linear mixed models <i>Anna Gottard, Leonardo Grilli, Carla Rampichini, Giulia Vannucci</i>	239

Weighted likelihood estimation of mixtures <i>Luca Greco, Claudio Agostinelli</i>	243
A canonical representation for multiblock methods <i>Mohamed Hanafi</i>	247
An adequacy approach to estimating the number of clusters <i>Christian Hennig</i>	251
Classification with weighted compositions <i>Karel Hron, Julie Rendlova, Peter Filzmoser</i>	255
MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers <i>Mia Hubert, Peter J. Rousseeuw, Wannes Van den Bossche</i>	256
Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present <i>Maria Iannario, Claudia Tarantola</i>	258
A multi-criteria approach in a financial portfolio selection framework <i>Carmela Iorio, Giuseppe Pandolfo, Roberta Siciliano</i>	262
Clustering of trajectories using adaptive distances and warping <i>Antonio Irpino, Antonio Balzanella</i>	266
Sampling and learning Mallows and generalized Mallows models under the Cayley distance: short paper <i>Ekhine Irurozki, Borja Calvo, Jose A. Lozano</i>	270
The gender parity index for the academic students progress <i>Aglaia Kalamatianou, Adele H. Marshall, Mariangela Zenga</i>	274
Some asymptotic properties of model selection criteria in the latent block model <i>Christine Keribin</i>	278
Invariant concept classes for transcriptome classification <i>Hans Kestler, Robin Szekely, Attila Klimmek, Ludwig Lausser</i>	282
Clustering of ties defined as symbolic data <i>Luka Kronegger</i>	283
Application of data mining in the housing affordability analysis <i>Viera Labudová, Eubica Sipková</i>	284
Cylindrical hidden Markov fields <i>Francesco Lagona</i>	288

Comparing tree kernels performances in argumentative evidence classification <i>Davide Liga</i>	292
Recent advancement in neural network analysis of biomedical big data <i>Pietro Liò, Giovanna Maria Dimitri, Chiara Sopegno</i>	296
Bias reduction for estimating functions and pseudolikelihoods <i>Nicola Lunardon</i>	297
Large scale social and multilayer networks <i>Matteo Magnani</i>	301
Uncertainty in statistical matching by BNs <i>Daniela Marella, Paola Vicard, Vincenzina Vitale</i>	305
Evaluating the recruiters' gender bias in graduate competencies <i>Paolo Mariani, Andrea Marletta</i>	309
Dynamic clustering of network data: a hybrid maximum likelihood approach <i>Maria Francesca Marino, Silvia Pandolfi</i>	313
Stability of joint dimension reduction and clustering <i>Angelos Markos, Michel Van de Velden, Alfonso Iodice D'Enza</i>	317
Hidden Markov models for clustering functional data <i>Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni</i>	321
Composite likelihood inference for simultaneous clustering and dimensionality reduction of mixed-type longitudinal data <i>Antonello Maruotti, Monia Ranalli, Roberto Rocci</i>	325
Bivariate semi-parametric mixed-effects models for classifying the effects of Italian classes on multiple student achievements <i>Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni</i>	329
Multivariate change-point analysis for climate time series <i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Carlo Blasi</i>	333
A dynamic stochastic block model for longitudinal networks <i>Catherine Matias, Tabea Rebafka, Fanny Villers</i>	337
Unsupervised fuzzy classification for detecting similar functional objects <i>Fabrizio Mauro, Francesca Fortuna, Tonio Di Battista</i>	339
Mixture modelling with skew-symmetric component distributions <i>Geoffrey McLachlan</i>	343

New developments in applications of pairwise overlap <i>Volodymyr Melnykov, Yana Melnykov, Domenico Perrotta, Marco Riani, Francesca Torti, Yang Wang</i>	344
Modelling unobserved heterogeneity of ranking data with the bayesian mixture of extended Plackett-Luce models <i>Cristina Mollica, Luca Tardella</i>	346
Issues in nonlinear time series modeling of European import volumes <i>Gianluca Morelli, Francesca Torti</i>	350
Gaussian parsimonious clustering models with covariates and a noise component <i>Keefe Murphy, Thomas Brendan Murphy</i>	352
Illumination in depth analysis <i>Stanislav Nagy, Jiří Dvořák</i>	353
Copula-based non-metric unfolding on augmented data matrix <i>Marta Nai Ruscone, Antonio D'Ambrosio</i>	357
A statistical model for software releases complexity prediction <i>Marco Ortu, Giuseppe Destefanis, Roberto Tonelli</i>	361
Comparison of serious diseases mortality in regions of V4 <i>Viera Pacáková, Lucie Kopecká</i>	365
Price and product design strategies for manufacturers of electric vehicle batteries: inferences from latent class analysis <i>Friederike Paetz</i>	369
A Mahalanobis-like distance for cylindrical data <i>Lucio Palazzo, Giovanni C. Porzio, Giuseppe Pandolfo</i>	373
Archetypes, prototypes and other types <i>Francesco Palumbo, Giancarlo Ragozini, Domenico Vistocco</i>	377
Generalizing the skew-t model using copulas <i>Antonio Parisi, Brunero Liseo</i>	381
Contamination and manipulation of trade data: the two faces of customs fraud <i>Domenico Perrotta, Andrea Cerasa, Lucio Barabesi, Mario Menegatti, Andrea Cerioli</i>	385
Bayesian clustering using non-negative matrix factorization <i>Michael Porter, Ketong Wang</i>	389

Exploring gender gap in international mobility flows through a network analysis approach <i>Ilaria Primerano, Marialuisa Restaino</i>	393
Clustering two-mode binary network data with overlapping mixture model and covariates information <i>Saverio Ranciati, Veronica Vinciotti, Ernst C. Wit, Giuliano Galimberti</i>	395
A stochastic blockmodel for network interaction lengths over continuous time <i>Riccardo Rastelli, Michael Fop</i>	399
Computationally efficient inference for latent position network models <i>Riccardo Rastelli, Florian Maire, Nial Friel</i>	403
Clustering of complex data stream based on barycentric coordinates <i>Parisa Rastin, Basarab Matei, Guénaél Cabanes</i>	407
An INDSCAL based mixture model to cluster mixed-type of data <i>Roberto Rocci, Monia Ranalli</i>	411
Topological stochastic neighbor embedding <i>Nicoleta Rogovschi, Nistor Grozavu, Basarab Matei, Younès Bennani, Seiichi Ozawa</i>	415
Functional data analysis for spatial aggregated point patterns in seismic science <i>Elvira Romano, Jonatan González Monsalve, Francisco Javier Rodríguez Cortés, Jorge Mateu</i>	419
ROC curves with binary multivariate data <i>Lidia Sacchetto, Mauro Gasparini</i>	420
Silhouette-based method for portfolio selection <i>Marco Scaglione, Carmela Iorio, Antonio D'Ambrosio</i>	424
Item weighted Kemeny distance for preference data <i>Mariangela Sciandra, Simona Buscemi, Antonella Plaia</i>	428
A fast and efficient modal EM algorithm for Gaussian mixtures <i>Luca Scrucca</i>	432
Probabilistic archetypal analysis <i>Sohan Seth</i>	436
Multilinear tests of association between networks <i>Daniel K. Sewell</i>	438

Use of multi-state models to maximise information in pressure ulcer prevention trials <i>Linda Sharples, Isabelle Smith, Jane Nixon</i>	442
Partial least squares for compositional canonical correlation <i>Violetta Simonacci Massimo Guarino, Michele Gallo</i>	445
Dynamic modelling of price expectations <i>Rosaria Simone, Domenico Piccolo, Marcella Corduas</i>	449
Towards axioms for hierarchical clustering of measures <i>Philipp Thomann, Ingo Steinwart, Nico Schmid</i>	453
Influence of outliers on cluster correspondence analysis <i>Michel Van de Velden, Alfonso Iodice D'Enza, Lisa Schut</i>	454
Earthquake clustering and centrality measures <i>Elisa Varini, Antonella Peresan, Jiancang Zhuang</i>	458
Co-clustering high dimensional temporal sequences summarized by histograms <i>Rosanna Verde, Antonio Irpino, Antonio Balzanella</i>	462
Statistical analysis of item pre-knowledge in educational tests: latent variable modelling and optimal statistical decision <i>Chen Yunxiao, Lu Yan, Iriini Moustaki</i>	466
Evaluation of the web usability of the University of Cagliari portal: an eye tracking study <i>Gianpaolo Zammarchi, Francesco Mola</i>	468
Application of survival analysis to critical illness insurance data <i>David Zapletal, Lucie Kopecka</i>	472

CONSENSUS CLUSTERING VIA PIVOTAL METHODS

Leonardo Egidi¹, Roberta Pappadà¹, Francesco Pauli¹ and Nicola Torelli¹

¹ Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche ‘Bruno de Finetti’, Università degli Studi di Trieste, (e-mail: legidi@units.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it)

ABSTRACT: We propose an approach to the cluster ensemble problem based on pivotal units extracted from a co-association matrix. It can be seen as a modified version of K -means method, which utilizes pivots for careful seeding. Different criteria for identifying the pivots are discussed, as well as preliminary results concerning the comparison with alternative ensemble methods.

KEYWORDS: cluster ensemble, pivot, K -means.

1 Introduction

Ensembles methods have recently emerged as a valid alternative to conventional clustering techniques and have shown to effectively improve the quality of clustering results and achieve robustness (see, e.g., Strehl & Ghosh, 2002, Jain, 2010). Such methods require a strategy to generate multiple clusterings of the same data set (the ensemble) and then combine them into a *consensus* partition (presumably superior), by following the idea of evidence accumulation, i.e., by viewing each clustering result as an independent evidence of data structure. A common way to do this is to obtain a new pairwise similarity matrix, or co-association matrix, by taking the co-occurrences of pairs of points in the same group across all partitions (Fred & Jain, 2005). Then, a similarity-based clustering algorithm can be applied to this matrix to yield the final partition.

We propose to use the co-association matrix to find some specific units (hereafter, pivots) which are representative of the group they belong to (because they never or very rarely co-occur with members of other groups). Various criteria for detecting the pivots are proposed in Section 2. Section 3 illustrates the use of pivotal methods for data clustering, and compare the proposed approach with classical K -means and other common ensemble methods.

Pivotal methods and related clustering procedures are implemented via the R package `pivmet`, available from the Comprehensive R Archive Network at

<http://CRAN.R-project.org/package=pivmet>.

2 Pivotal methods based on co-association

Let $\mathbf{Y} = (y_1, \dots, y_n)$ be a set of n observations, where $y_i \in \mathbb{R}^d$. Consider a set $\mathcal{P} = \{P^1, P^2, \dots, P^H\}$ of H partitions of the data points into K disjoint clusters, derived from an arbitrary clustering algorithm. Note that the number of groups is pre-specified and equal for all P^h . \mathcal{P} can be summarized via the $n \times n$ co-association matrix C with generic element

$$c_{i,j} = \frac{1}{H} \sum_{h=1}^H |P^h(y_i) = P^h(y_j)|, \quad (1)$$

where $|\cdot|$ denotes the indicator function, and $P^h(y_i), P^h(y_j)$, represent the clusters of the objects y_i and y_j in P^h , respectively. Clearly, units which are very dissimilar from each other are likely to have zero co-occurrences; as a consequence, C is expected to contain a non-negligible number of zeros. Given a large and sparse 0-1 matrix, the Maxima Units Search (MUS) algorithm seeks those elements, among a pre-specified number of candidate pivots, whose corresponding rows contain more zeros compared to all other units (Egidi *et al.*, 2018c). Define a reference partition, G_1, \dots, G_K of y_1, \dots, y_n obtained by applying, for instance, an agglomerative hierarchical algorithm into K groups. The MUS procedure takes C as input and outputs a set of K units—one for each group of the reference partition—that exhibit the highest degree of separation (Egidi *et al.*, 2018b). As an alternative approach, the pivot y_{i_k} for group G_k can be chosen so that it is as far as possible from units that might belong to other groups and/or as close as possible to units that belong to the same group, according to one of the following objective functions

$$(a) \max_{i_k} \sum_{j \in G_k} c_{i_k,j} \quad (b) \min_{i_k} \sum_{j \notin G_k} c_{i_k,j} \quad (c) \max_{i_k} \sum_{j \in G_k} c_{i_k,j} - \sum_{j \notin G_k} c_{i_k,j}, \quad (2)$$

where $c_{i,j}$ is defined as in (1). Ideally, the $K \times K$ submatrix of C with only the rows and columns corresponding to i_1, \dots, i_K will be the identity matrix. In practice, it may contain few nonzero elements off the diagonal.

3 A simulation experiment

In order to illustrate the proposed algorithm, we simulate bivariate data from a mixture of three Gaussian distributions with mean vectors $\boldsymbol{\mu}_1 = (1, 5)$, $\boldsymbol{\mu}_2 =$

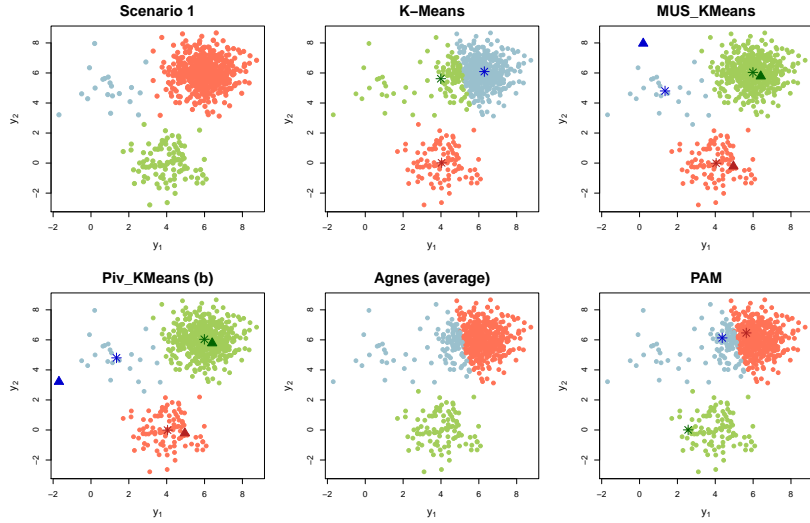


Figure 1. Mixture of three Gaussian distributions (sample size $n=620$). Cluster centers and/or pivots for each method are marked via asterisks and triangles, respectively.

$(4,0)$, $\mu_3 = (6,6)$, and the 2×2 identity matrix as covariance matrix. The components have sample size 20, 100 and 500, respectively (see Figure 1, top-left panel). The K -means algorithm with random seeds is used to generate a cluster ensemble of $H = 1000$ partitions, and obtain the co-association matrix C . For each simulated dataset, we proceed as follows:

1. For a given number of clusters K , obtain a partition of the data G_1, \dots, G_K (reference partition);
2. Apply the MUS algorithm or one alternative criterion in (2) to the matrix C to find K (distinct) pivots y_{i_1}, \dots, y_{i_K} ;
3. Run the K -means algorithm using the pivots as initial cluster centers.

The proposed modification of the standard K -means technique introduces a pivot-based initialization step with the aim of reducing the effect of random seeding (see also Egidi *et al.*, 2018a). An alternative approach to careful seeding can be found in Arthur & Vassilvitskii, 2007. Figure 1 shows the solution from K -means, using $K = 3$, and by pivotal methods MUS and criterion (b) in Eq. (2), where Average-Linkage (AL) agglomerative clustering is used to obtain the reference partition. The results of consensus clustering using PAM (Partitioning Around Medoids) method and AL-agglomerative hierar-

chical clustering (agnes) are also shown (Single Linkage (SL) and Complete Linkage (CL) give similar results). Table 1 reports the comparison between the different methods in terms of Adjusted Rand Index (ARI), used to quantify the agreement between two partitions. The mean value is considered for 1000 simulations. Preliminary results suggest that the pivot-based approach outperforms the competing similarity-based ensemble methods and the standard K -means, which gives a mean ARI of 0.659.

Table 1. 2D Gaussian data: mean ARI (1000 simulations) between the final clustering and the true data partition. Ensemble methods use dissimilarities $1 - c_{i,j}$.

Pivotal methods	MUS	(a)	(b)	(c)
	0.857	0.865	0.883	0.779
Ensemble methods	agnes (AL)	agnes (SL)	agnes (CL)	PAM
	0.512	0.535	0.514	0.506

References

- ARTHUR, D., & VASSILVITSKII, S. 2007. k-means++: The advantages of careful seeding. *Pages 1027–1035 of: Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete algorithms.*
- EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018a. K-means seeding via MUS algorithm. *Pages 256–262 of: Book of Short Papers SIS 2018.*
- EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018b. Maxima Units Search (MUS) algorithm: methodology and applications. *Pages 71–81 of: Studies in Theoretical and Applied Statistics.*
- EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018c. Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, **28**, 957–969.
- FRED, A. L. N., & JAIN, A. K. 2005. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 835–850.
- JAIN, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651 – 666.
- STREHL, A., & GHOSH, J. 2002. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, **3**, 583–617.