CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Pearson

SCIENTIFIC PROGRAM COMMITTEE

Carla Rampichini (chair, University of Florence - Italy)

Claudio Agostinelli (University of Trento - Italy)
Michela Battauz (University of Udine - Italy)
Antonio Canale (University of Padua - Italy)
Carlo Cavicchia (Erasmus University Rotterdam - Netherlands)
Claudio Conversano (University of Cagliari - Italy)
Eustasio del Barrio (University of Valladolid - Spain)
Roberto Di Mari (University of Catania - Italy)
Stefania Fensore (University of "G. d'Annunzio" - Italy)
Nial Friel (University College Dublin - Ireland)
Maria Giovanna Ranalli (University of Perugia - Italy)
Leonardo Grilli (University of Firenze - Italy)
Luigi Grossi (University of Padua - Italy)
Christian Hennig (University of Bologna - Italy)
Mia Hubert (KU Leuven - Belgium)
Alfonso Iodice D'Enza (University of Naples "Federico II" - Italy)
Julien Jacques (University of Lyon - France)
José Joaquim Dias Curto (ISCTE-Instituto Universitário de Lisboa- Portugal)
Michele La Rocca (University of Salerno - Italy)
Silvia Montagna (University of Turin - Italy)
Barbara Pawelek (University of Cracow - Poland)
Fulvia Pennoni (University of Milano-Bicocca - Italy)
Mario Rosario Guarracino (University of Cassino - Italy)
Katrijn Van Deun (University of Tilburg - Netherlands)
Simone Vantini (Politecnico di Milano - Italy)
Donatella Vicari (Sapienza University of Rome - Italy)
Helga Wagner (Johannes Kepler University Linz - Austria)
Hiroshi Yadohisa (Doshisha University - Japan)

LOCAL PROGRAM COMMITTEE

Michele La Rocca (chair, University of Salerno - Italy)

Pietro Coretto (University of Salerno - Italy)
Giuseppe Giordano (University of Salerno - Italy)
Paolo Rocca Comite Mascambruno (University of Salerno - Italy)
Marcella Niglio (University of Salerno - Italy)
Maria Lucia Parrella (University of Salerno - Italy)
Marialuisa Restaino (University of Salerno - Italy)
Domenico Vistocco (University of Naples "Federico II" - Italy)
Maria Prosperina Vitale (University of Salerno - Italy)

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:
14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13, 2023
edited by Carla Rampichini, Michele La Rocca, Pietro Coretto, Giuseppe Giordano, Maria Lucia Parrella

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,
© Tartila / Shutterstock

Published by Pearson Education Resources, Italia
www.pearson.it
ISBN: 9788891935632

INDEX

Preface

This book collects the abstracts and short papers presented at CLADAG 2023, the 14th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Economics and Statistics of the University of Salerno, under the auspices of the University of Salerno, the SIS and the International Federation of Classification Societies (IFCS).

CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research. Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings. CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became an important meeting point for people interested in classification and data analysis. A selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG 2023 conceived the Keynote Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2023 is particularly rich. All in all, it comprises 5 Keynote Lectures, 31 Invited Sessions promoted by the members of the Scientific Program Committee, and 27 Contributed Sessions. We thank all the session organizers for inviting renowned speakers, coming from many different countries. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts and short papers collected in this book. Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2023. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Salerno, September 2023

EMPIRICAL ANALYSIS OF THE QUADRATIC SCORING FOR SELECTING CLUSTERING SOLUTIONS

Luca Coraggio [1], Pietro Coretto[2]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: luca.coraggio@unina.it)

[2] Department of Economics and Statistics, University of Salerno, (e-mail: pcoretto@unisa.it)

ABSTRACT: Selecting an optimal clustering solutions is a difficult problem, and there exist many data-driven validation strategies to perform this task. In this paper, we focus on a recent proposal, the BQH and BQS criteria, based on quadratic discriminant scores and bootstrap resampling. We provide more insight on these criteria, comparing them with a likelihood-based alternative and using different resampling schemes.

KEYWORDS: cluster validation, mixture models, model-based clustering, resampling methods

1 Quadratic scoring, likelihood-based scoring, and resampling

Selecting an optimal clustering solution is not an easy task (von Luxburg *et al.*, 2012). Recently, in Coraggio & Coretto, 2023, we proposed a novel validation index aimed at selecting clustering solutions in cases where clusters can be expected to have elliptic-symmetric shapes, or to be separable by quadratic boundaries.

Let $\mathbb{X}_n$ indicate sample data, and $\mathcal{G}^{(m)} = \left\{ G_k^{(m)},\ k = 1,\ldots,K_m \right\}$ be a clustering solution, obtained running clustering method $m \in \mathcal{M}$. We assume that $\mathcal{G}^{(m)}$ can be meaningfully described by K_m triplets $\boldsymbol{\theta}^{(m)} = \left\{ \boldsymbol{\theta}_k^{(m)},\ k = 1,\ldots,K_m \right\}$, each collecting unique elements of *(i)* π_k, the expected fraction of points belonging to the *k-th* group; *(ii)* $\boldsymbol{\mu}_k \in \mathbb{R}^p$, the *k-th* cluster's center; *(iii)* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ a positive definite scatter matrix. For a point $\boldsymbol{x}$ and a triplet $\boldsymbol{\theta}_k$, we define the quadratic score (inspired to Quadratic Discriminant Analysis; e.g., see Hastie *et al.*, 2009) of point $\boldsymbol{x}$ for the *k-th* cluster as

$$\mathrm{qs}(x, \boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k); \qquad (1)$$

it can be seen as a measure of how well point $\boldsymbol{x}$ is accommodated into cluster k. The hard (QH) and (QS) smooth scores are based on (1), and are essentially

Algorithm 1 Bootstrap likelihood-based scoring

input: observed sample $\mathbb{X}_n$ (with ecdf $\mathbb{F}_n$), $\alpha \in (0,1)$; clustering method $m \in \mathcal{M}$; integers $B > 0$

output: bootstrap likelihood-based scoring for method m: $\widetilde{L}_n^{(m)}$.

(to ease notation, dependence on m is dropped and reintroduced in step 3)

for $b \in \{1,\dots,B\}$ do

(step 1.1) $\mathbb{X}_n^{(b)} \leftarrow$ non-parametric bootstrap resample from $\mathbb{X}_n$ (sample of size n from $\mathbb{F}_n$)

(step 1.2) $\hat{\boldsymbol{\theta}}_n^{(b)} \leftarrow$ triplets of parameters from clustering solution m fitted on $\mathbb{X}_n^{(b)}$

(step 1.3) $S_n^{(b)} \leftarrow l(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n)$ (score solution on $\mathbb{X}_n$)

end for

(step 2) $\widetilde{W}_n \leftarrow \frac{1}{B}\sum_{b=1}^{B} S_n^{(b)} \qquad R_n^{(b)} \leftarrow \sqrt{n}\left(S_n^{(b)} - \widetilde{W}_n\right)$

(step 3) Compute $(\alpha/2)$-level and $(1-\alpha/2)$-level empirical quantiles:

$$\widetilde{L}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge \frac{\alpha}{2} \right\}; \qquad \widetilde{U}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge 1-\frac{\alpha}{2} \right\}$$

weighted averages of the quadratic score (see Coraggio & Coretto, 2023 for details). The quadratic score (1) is strongly connected to likelihood theory, and it is easy to show that it is proportional to the Gaussian density function. Thus, as a natural alternative to the scoring criteria we use the following likelihood function

$$l(\boldsymbol{\theta}^{(m)};\, \mathbb{X}_n) = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathbb{X}_n} \log\left(\sum_{k=1}^{K^{(m)}} \pi_k^{(m)} \phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})\right), \tag{2}$$

where $\phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})$ is the density function of a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.

Choosing the solution that maximizes (2) may give poor results: since the sample data $\mathbb{X}_n$ is used both to estimate $\boldsymbol{\theta}^{(m)}$ and for scoring, overly-complex solutions may be selected due to overoptimism in the evaluation process. Thus, we use the same resampling scheme used for the BQH and BQS scores, proposed in Coraggio & Coretto, 2023, that is to estimate clustering solutions on non-parametric bootstrap resamples (Efron, 1979) from $\mathbb{X}_n$, while using the full data to evaluate the score. The procedure is reviewed in Algorithm 1 for the likelihood-based scoring criterion.

2 Empirical analysis

The experimental analysis is a scaled-down version of that in Coraggio & Coretto, 2023, using the Pentagon5, T510D and Uniform simulated data sets.

Table 1: Selected solution by selection criteria (left-most column). Each subtable shows results from a data set: the first column shows the selected solution, and the second column reports its ARI, computed against true classes.

	(b) Pentagon5		(c) T510D		(d) Uniform	
Criterion	Selected m	ARI	Selected m	ARI	Selected m	ARI
QH	M, K=3, VVV	**0.86**	O, K=10, γ=10^4	0.51	O, K=10, γ=10^3	0
QS	M, K=3, VVV	**0.86**	O, K=10, γ=10^4	0.51	M, K=8, VVV	0
LK	O, K=10, γ=10^3	0.44	O, K=10, γ=10^4	0.51	O, K=10, γ=10^3	0
CVQH	M, K=3, EVE	**0.86**	O, K=6, γ=1	0.73	O, K=5, γ=10^2	0
CVQS	M, K=3, EVE	**0.86**	O, K=5, γ=1	0.97	M, K=1, EEI	**1**
CVLK	M, K=5, EVI	0.86	O, K=8, γ=1	0.60	M, K=7, VEE	0
BQH	M, K=3, EVE	**0.86**	O, K=8, γ=5	0.57	O, K=9, γ=10^4	0
BQS	M, K=3, EVE	**0.86**	O, K=5, γ=5	**0.98**	M, K=1, EEI	**1**
BLK	O, K=5, γ=1	0.85	O, K=8, γ=5	0.57	M, K=10, VVI	0

Since likelihood-based scoring is only justified for model-based clustering, $\mathcal{M}$ includes: *(i)* 140 Gaussian mixture models with covariance matrices restrictions (Banfield & Raftery, 1993), implemented with the Mclust (M) software (Scrucca *et al.*, 2016; setting $K = 1, \ldots, 10$, and 14 covariance models); *(ii)* 180 Gaussian mixture models with eigen-ratio contraints (ERC; Ingrassia, 2004), implemented with Otrimle (O) software (Coretto & Hennig, 2017, Coretto & Hennig, 2021; setting $K \in \{1, \ldots, 10\}$, ERC $\gamma \in \{1, 5, 10, 10^2, 10^3, 10^4\}$, and 3 initialization methods). The criteria compared to select optimal solutions are as follows. QH, QS, and LK: clustering solutions are estimated and scored using the full data, $\mathbb{X}_n$; CVQH, CVQS, CVLK: clustering solutions are estimated on a "train set" and scored on a non-overlapping "test set", using a 10-fold cross-validation scheme, as in Smyth, 2000. BQH, BQS, BLK: clustering solution are estimated and scored according to Algorithm 1, selecting the method m maximizing $\widetilde{L}_n^{(m)}$. For each criterion, the selected solutions are evaluated against the true class labels, reporting the achieved Adjusted Rand Index (ARI, Hubert & Arabie, 1985).

Results are presented in Table 1. The comparison gives a better understanding on the mechanism that lies behind the effectiveness of the BQH and BQS criteria. First, notice that all criteria where solutions are estimated and scored on the full data (QH, QS, LK) always select overly-complex solutions. The extra penalization of the smooth score on overlapping clusters is key to select better solutions in more complicated settings (T510D and Uniform). Finally, the bootstrap scheme improves on the cross-validation. Overall, both the quadratic scores, QH and QS, and the resampling scheme in Algorithm 1 seem

equally important to consistently achieve good results.

3 Conclusion

In this paper, we run an empirical comparison of the BQH and BQS procedures from Coraggio & Coretto, 2023 with a likelihood-based alternative, using different resampling schemes. Our experiments provide new insights on the criteria, showing that both the bootstrap resampling scheme and the quadratic scores contribute equally to the procedure: *(*i) the penalization for clusters' overlap from the quadratic scores allows achieving better results in cases where clusters are not well separated; *(ii)* the bootstrap resampling scheme allows to effectively take into account clustering methods' variability, better than cross-validation would (likely better suited for prediction settings).

References

BANFIELD, JEFFREY D., & RAFTERY, ADRIAN E. 1993. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**(3), 803.

CORAGGIO, LUCA, & CORETTO, PIETRO. 2023. Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. *Journal of Multivariate Analysis*, **196**(July), 105181.

CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2017. Consistency, Breakdown Robustness, and Algorithms for Robust Improper Maximum Likelihood Clustering. *Journal of Machine Learning Research*, **18**(142), 1–39.

CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2021. *otrimle: Robust Model-Based Clustering*. R package version 2.0.

EFRON, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**(1).

HASTIE, TREVOR, TIBSHIRANI, ROBERT, & FRIEDMAN, JEROME. 2009. *The Elements of Statistical Learning*. 2 edn. Springer Series in Statistics (SSS). Springer New York.

HUBERT, LAWRENCE, & ARABIE, PHIPPS. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

INGRASSIA, SALVATORE. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, **13**(2).

SCRUCCA, LUCA, FOP, MICHAEL, MURPHY, T. BRENDAN, & RAFTERY, ADRIAN E. 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, **8**(1), 289–317.

SMYTH, PADHRAIC. 2000. *Statistics and Computing*, **10**(1), 63–72.

VON LUXBURG, ULRIKE, WILLIAMSON, ROBERT C., & GUYON, ISABELLE. 2012. Clustering: Science or Art? Proceedings of Machine Learning Research, vol. 27. Bellevue, Washington, USA: PMLR.