stats

MDPI

*Article*

# Parameter Choice, Stability and Validity for Robust Cluster Weighted Modeling

**Andrea Cappozzo [1,†], Luis Angel García Escudero [2], Francesca Greselin [3,*,†] and Agustín Mayo-Iscar [2]**

[1] MOX-Department of Mathematics, Politecnico di Milano, 20133 Milan, Italy; andrea.cappozzo@polimi.it
[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, 47002 Villadolid, Spain; lagarcia@uva.es (L.A.G.E.); agustin.mayo.iscar@uva.es (A.M.-I.)
[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, 20126 Milan, Italy
* Correspondence: francesca.greselin@unimib.it
† These authors contributed equally to this work.

**Abstract:** Statistical inference based on the cluster weighted model often requires some subjective judgment from the modeler. Many features influence the final solution, such as the number of mixture components, the shape of the clusters in the explanatory variables, and the degree of heteroscedasticity of the errors around the regression lines. Moreover, to deal with outliers and contamination that may appear in the data, hyper-parameter values ensuring robust estimation are also needed. In principle, this freedom gives rise to a variety of "legitimate" solutions, each derived by a specific set of choices and their implications in modeling. Here we introduce a method for identifying a "set of good models" to cluster a dataset, considering the whole panorama of choices. In this way, we enable the practitioner, or the scientist who needs to cluster the data, to make an educated choice. They will be able to identify the most appropriate solutions for the purposes of their own analysis, in light of their stability and validity.

**Keywords:** cluster-weighted modeling; outliers; trimmed BIC; eigenvalue constraint; monitoring; constrained estimation; model-based clustering; robust estimation



**Citation:** Cappozzo, A.; García Escudero, L.A.; Greselin, F.; Mayo-Iscar, A. Parameter Choice, Stability and Validity for Robust Cluster Weighted Modeling. *Stats* **2021**, *4*, 602–615. https://doi.org/10.3390/stats4030036

Academic Editor: Wei Zhu

Received: 30 April 2021
Accepted: 30 June 2021
Published: 6 July 2021



## 1. Introduction

One of the most fundamental problems tackled in data mining is clustering. A plethora of algorithms, procedures, and theoretical investigations have been developed in the literature to identify groups in data. Several monographs have been published on the topic, to cite a few excellent ones we suggest [1–3], among many others. Applications can be found in virtually every possible area, spanning from bioinformatics, marketing, image analysis to text and web mining.

Clustering is the "art" of decomposing a given data set into subgroups, where observations are as similar as possible within clusters, while being the most heterogeneous between them. Apart from this informal description, however, there is no universally appropriate unique formalism, algorithm, and/or evaluation measure for clustering. The very same definition of cluster and, as a consequence, the most appropriate clustering procedure, heavily depends on the application at hand and on the (subjective) rationale defining similarity between units. These considerations can be subsumed by saying that clustering per se is an ill-posed problem, where the number of clusters, their shape, and their parameters depend, in general, on a multiplicity of choices made by the modeler. We refer the interested reader to the thought-provoking work in [4] for a deeper discussion on the topic. A complementary point of view on this regard stems as well from the machine learning community, where the stability of clustering solutions has been treated in a principled way in [5,6]. All in all, only in a few cases there is no ambiguity on a partitioning solution.