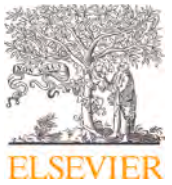
International Journal of Approximate Reasoning 195 (2026) 109698



Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Robust fuzzy clustering with cellwise outliers

Giorgia Zaccaria [a,*], Lorenzo Benzakour [b], Luis A. García-Escudero [c], Francesca Greselin [b], Agustín Mayo-Íscar [c]

[a] Department of Economics, Management and Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan, 20100, Italy
[b] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan, 20100, Italy
[c] Department of Statistics and Operational Research, University of Valladolid, Paseo de Belén 7, Valladolid, 47011, Spain

ARTICLE INFO

Keywords:
Cellwise contamination
Fuzzy assignments
Constrained estimation
Robust clustering
Outlier detection
High contrast property

ABSTRACT

In a data matrix, we may distinguish between cases, each represented by a row vector for a statistical unit, and cells, which correspond to single entries of the data matrix. Recent developments in Robust Statistics have introduced the cellwise contamination paradigm, which assumes contamination on cells rather than on entire cases. This approach becomes particularly relevant as the number of variables increases. Indeed, discarding or downweighting entire cases because of a few anomalous cells in them, as done by traditional (casewise) robust methods, can result in substantial information loss, since the non-contaminated (or reliable) cells can still be highly informative. This philosophy can also be considered in fuzzy clustering, by assuming that reliable cells within a case may still provide useful information for determining fuzzy memberships. A robust fuzzy clustering proposal is thus introduced in this work, combining the advantages of dealing with outlying cells and simultaneously controlling the degree of fuzziness of unit assignments. The cluster-specific relationships among variables, detected by the fuzzy clustering approach, are also key to better identifying outlying cells and correct them. The strengths of the proposed methodology are illustrated through a simulation study and two real-world applications. The effects of the model's tuning parameters are explored, and some guidance for users on how to set them suitably is provided.

1. Introduction

The goal of the unsupervised clustering is to discover subpopulations, called clusters, within the data which share common characteristics, and to estimate their statistical features (e.g., centers, scatter/covariance matrices, etc.). In many cases, clusters are not perfectly separated, and the information provided by partitioning units may offer a narrow perspective on the underlying structure of the data. To overcome this limitation of "hard" clustering algorithms such as, for instance, k-means [1,2], fuzzy clustering approaches have been introduced [3], in which units are not fully assigned to a single cluster but can have positive membership degrees to more than one cluster. Many of these fuzzy clustering approaches are reviewed in Höppner et al. [4], De Oliveira and Pedrycz [5], Giordani et al. [6]. One of the main advantages of fuzzy clustering methods is that they allow for varying degrees of fuzzification through a tuning parameter m, say the fuzzifier parameter, which can be set depending on the purpose of the analysis. This feature cannot be achieved by other clustering methodologies, such as finite mixture models [7], which, although they quantify

* Corresponding author.
E-mail addresses: giorgia.zaccaria@unimib.it (G. Zaccaria), l.benzakour@campus.unimib.it (L. Benzakour), lagarcia@uva.es (L.A. García-Escudero), francesca.greselin@unimib.it (F. Greselin), agustin.mayo.iscar@uva.es (A. Mayo-Íscar).

https://doi.org/10.1016/j.ijar.2026.109698
Received 25 November 2025; Received in revised form 27 March 2026; Accepted 17 April 2026
Available online 21 April 2026

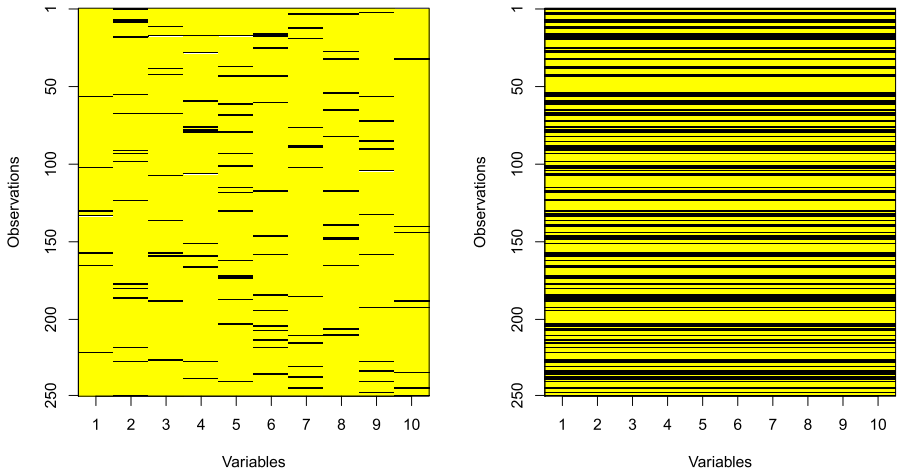


Fig. 1. Data matrix with 250 units, 10 variables, and 5% of contaminated cells shown in black (left); corresponding rows in black that would be trimmed by a casewise robust method due to containing at least one outlying cell (right).

membership uncertainty and provide "soft" clustering through the estimation of posterior probabilities, do not provide control over the desired degree of fuzziness.

That flexibility of the fuzzy clustering methods is especially important in domains such as, for instance, medical diagnosis and treatment planning for subgroups of patients. A concrete example comes from clinical studies on certain diseases, where some patients may present overlapping symptoms or ambiguous laboratory test results, making it difficult to assign them uniquely to a diagnostic category. In such cases, fuzzy clustering enables, on the one hand, the identification of a subset of patients for whom further testing may be particularly recommended, and on the other hand, the possibility of adjusting the size of this subset based on the purpose of the analysis or specific clinical priorities. However, as this example highlights, the interest in fuzzification typically concerns only a subset of units rather than all of them: for many units, the membership to a specific cluster may be sufficiently strong that a hard assignment is appropriate. This motivates the use of fuzzy clustering methods with the so-called "high contrast" property [8], where hard and soft assignments coexist within the same framework.

In real-world applications, data often contain measurement errors, anomalies, or, more generally, outliers, some of which may actually correspond to behaviors of genuine interest in the data. In Robust Statistics [9,10], outliers have been traditionally defined as cases (or units) that do not follow the pattern of the majority of the data, and therefore referred to as casewise outliers. Specifically, given an $(n \times p)$ data matrix $\mathbf{X}$, where p variables are measured on n units, we may distinguish between *cases*, represented by the rows $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, of $\mathbf{X}$, and *cells*, which correspond to the single measurements x_{ij} of the j-th variable ($j = 1, \ldots, p$) for the i-th unit ($i = 1, \ldots, n$). With this notation, the common assumption in Robust Statistics is that an entire case $\mathbf{x}_i$ may be considered atypical.

The presence of contaminating cases $\mathbf{x}_i$ – even in very small numbers – may cause methods traditionally applied to hard and fuzzy clustering to fail dramatically, resulting in the identification of clusters of very limited practical interest. This well-known phenomenon has motivated the development of numerous robust clustering techniques, from both hard and fuzzy perspectives (see Dave and Krishnapuram [11], García-Escudero et al. [12], Banerjee and Davé [13], García-Escudero et al. [14] and references therein), essentially from a casewise point of view. Until the first decade of the 2000s, this was indeed the dominant approach for dealing with outliers. However, the increasing dimensionality of the data makes it reasonable to assume that units may exhibit only a few outlying cells x_{ij}, while the remaining (non-outlying) cells may still contain reliable information for data analysis. This has led to the development of robust approaches aimed at addressing cellwise contamination [15,16], in which only single entries x_{ij} of a data matrix, rather than entire rows $\mathbf{x}_i$, are assumed to be contaminated by arbitrary values. Under this type of contamination, discarding, assigning to a noise component or downweighting entire cases, as done by traditional casewise robust methods for fuzzy clustering, can result in substantial information loss, even when the overall proportion of atypical cells is very small. Additionally, note that, as the number of variables p increases, cellwise contamination is likely to affect nearly all cases, often through a single outlying value in many of them, causing (even robust) fuzzy clustering to fail (see Fig. 1). Moreover, in the cellwise paradigm, it is possible to take advantage of the reliable information within a case to obtain a predicted value for each cell and compare it with the observed one. If the former deviates from the latter, the cells are flagged as contaminated, and corrected through imputation for the parameter estimation, rather than discarding or downweighing them. Motivated by the cellwise robust approach, Raymaekers and Rousseeuw introduced cellMCD [17] for estimating the location and scatter parameters in single-population problems under cellwise contamination, building on the casewise-robust Minimum Covariance Determinant (MCD) estimator [18,19]. The cellMCD approach has been recently extended to mixture modeling in cellGMM [20] (see also Puchhammer et al. [21]). However, none of these proposals is designed to addresses cellwise contamination in the fuzzy clustering framework.

The main contribution of this paper is a novel approach, called cellFCLUST, which combines three key features: (i) cellwise outlier detection and robust parameter estimation; (ii) fuzzy clustering; and (iii) the flexibility of tuning hard and soft assignments for units

(the previously commented "high contrast" property [8]). CellFCLUST arises as a cellwise extension of the F-TCLUST casewise robust fuzzy clustering method introduced in Fritz et al. [22]. F-TCLUST aims to produce a fuzzy partition of the units and to estimate the location and scatter parameters of the clusters using a trimmed classification maximum likelihood approach. The latter relies on the individual contribution to the objective function to identify a subset of cases that are retained as reliable for parameter estimation, while detecting and trimming (removing) the most anomalous cases. Previous related methods can be found in Kim et al. [23], where fuzzy c-means [24] is robustified via trimming, and in Dave [25], in which outliers are not trimmed but instead modeled as belonging to a "noise" cluster. Further approaches for achieving robustness, such as possibilistic clustering [26] or models with heavy-tailed components [27], are, to the best of our knowledge, just focused on the casewise robustness paradigm.

The novel proposal introduced in this work assumes a Gaussian distribution for the clusters and allows cluster covariance matrices to have distinct elliptical shapes, similar to F-TCLUST. The dependence relationships among variables may therefore vary across clusters and can be used for detecting and correcting outlying cells. Specifically, by setting a proportion of cells to flag for the variables, cellFCLUST identifies and imputes them using the information contained in the remaining reliable cells of each case, making the use of an algorithm inspired by the Expectation-Maximization (EM, Dempster et al. [28]) one suitable for performing this correction. As a result of the imputation, *all* units are assigned to clusters with a certain degree of fuzziness, unlike trimming approaches, which entirely discard cases, or noise clustering methods, where outliers are not grouped with regular data. Following this rationale, cellFCLUST can naturally handle missing information, which often occurs in real data sets. Missing entries usually affect cells x_{ij} of a data matrix, but their positions (i, j) are known in advance, unlike cellwise outlying values; both types of cells are referred to as *unreliable* in this paper. Several fuzzy clustering methods have been extended to cope with incomplete data sets, such as fuzzy c-means with incomplete information [29] (see Jyoti et al. [30] for an overview); however, none of them can deal with cellwise contamination, in which the position of the outlying cells is completely unknown. Moreover, in cellFCLUST, hard and soft assignments coexist within the same framework, and the proportions of each type of membership can be tuned by adjusting relevant parameters. This represents a novel feature compared to cellGMM, where this property cannot be achieved.

The paper is organized as follows. A brief review of F-TCLUST and its key features is provided in Section 2 to help the reader follow its extension from casewise to cellwise robust fuzzy clustering. Section 3 formally introduces cellFCLUST, including its parameter estimation and algorithm. A simulation study was conducted to evaluate the performance of cellFCLUST in comparison with alternative methods for fuzzy clustering with and without outlier detection; the results are reported in Section 4. In Section 5, we provide guidance on the tuning parameter selection through examples that illustrate their impact on cluster recovery and outlier detection. Two real data applications from different fields are presented in Section 6. Finally, Section 7 concludes the paper with a discussion on the proposed methodology, as well as potential directions for future developments in the cellwise clustering literature. The notation used throughout the paper is summarized in Appendix A, facilitating the reading of the methodological sections.

2. Background: F-TCLUST for casewise robust fuzzy clustering

Fuzziness in clustering was first introduced through the fuzzy k-means (or fuzzy c-means [24], here denoted as FKM), which assigns units to clusters based on their squared Euclidean distance from the cluster mean vectors, often called centroids. Specifically, let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]'$ be an $(n \times p)$ data matrix. The FKM objective function to be minimized is

$$J_{\text{FKM}}(\mathbf{U}, \{\boldsymbol{\mu}_k\}_{k=1}^K) = \sum_{i=1}^n \sum_{k=1}^K u_{ik}^m \left\|\mathbf{x}_i - \boldsymbol{\mu}_k\right\|^2, \tag{1}$$

where $m \geq 1$ is the fuzzifier tuning parameter (a larger m increases the degree of fuzziness, allowing more overlap between clusters and making cluster memberships less crisp), $\mathbf{U}$ is an $(n \times K)$ membership matrix with $u_{ik} \in [0, 1]$ denoting the membership degree of the i-th unit to the k-th cluster and such that $\sum_{k=1}^K u_{ik} = 1$, for $i = 1, \dots, n$, and $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ are cluster centroids in $\mathbb{R}^p$.

One of the main disadvantages of FKM, as in classical hard k-means, is its preference for spherical clusters with the same covariance structure, which makes it poorly suited for detecting more general types of clusters. Several procedures have been proposed to relax the isotropic assumption (see Gustafson and Kessel [31], Trauwaert et al. [32], Rousseeuw et al. [33], Gath and Geva [34], among others). These methods can be robustified through casewise trimming. In particular, in this section we focus on F-TCLUST [22], which integrates fuzzification into a classification maximum likelihood approach, accommodates different cluster shapes, and removes outlying cases from parameter estimation via impartial trimming. The term "impartial" means that the method itself detects the cases to trim, avoiding any user intervention in their identification; the user only needs to set the fraction α of casewise outliers. The objective function of F-TCLUST to be maximized is

$$J_{\text{F-TCLUST}}(\mathbf{U}, \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) = \sum_{i=1}^n \sum_{k=1}^K u_{ik}^m \log(\pi_k \phi_p(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)), \tag{2}$$

where $\phi_p(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right)$ is the probability density function evaluated at $\mathbf{x}_i \in \mathbb{R}^p$ of a p-variate normal distribution with p-dimensional mean vectors $\boldsymbol{\mu}_k$ and $(p \times p)$ positive definite covariance matrices $\boldsymbol{\Sigma}_k$ holding the dependence relationships among variables within clusters. The considered weights π_k are positive numbers such that $\sum_{k=1}^K \pi_k = 1$. The membership degrees $u_{ik} \in [0, 1]$ satisfy the following constraints

$$\sum_{k=1}^K u_{ik} = 1 \text{ when } i \in \mathcal{I} \text{ and } \sum_{k=1}^K u_{ik} = 0 \text{ when } i \notin \mathcal{I},$$

where $\mathcal{I}$ is the subset of not trimmed cases, with $\#\mathcal{I} = \lceil (1-\alpha)n \rceil$, i.e., α corresponds to the trimming level.

Since we deal with a maximum likelihood-type problem in the clustering framework, unboundedness may potentially arise when maximizing (2). To overcome that issue, F-TCLUST is also subject to the following eigenvalue-ratio constraint

$$\frac{\max_{k=1,\dots,K} \max_{j=1,\dots,p} \lambda_j(\boldsymbol{\Sigma}_k)}{\min_{k=1,\dots,K} \min_{j=1,\dots,p} \lambda_j(\boldsymbol{\Sigma}_k)} \leq c \tag{3}$$

where $c \geq 1$ is a fixed constant, and $\lambda_1(\boldsymbol{\Sigma}_k), \dots, \lambda_p(\boldsymbol{\Sigma}_k)$ are the p eigenvalues of $\boldsymbol{\Sigma}_k$. Constraint (3) ensures the maximization of $J_{\text{F-TCLUST}}$ in (2) as a well-posed problem, preventing its unboundedness (which may occur, for example, by taking $u_{i1} = 1$, $\pi_1 = 1$, $\boldsymbol{\mu}_1 = \mathbf{x}_i$, and letting $|\boldsymbol{\Sigma}_1| \downarrow 0$). As the constant c, which constrains the ratio between the largest and smallest eigenvalues of the covariance matrices within and across clusters, decreases, the cluster configurations become increasingly similar to each other and closer to sphericity, ultimately reaching spherical and equally dispersed clusters when $c = 1$. The eigenvalue-ratio constraint in (3) is also an important tool to avoid detecting so-called "spurious" clusters with little statistical relevance, typically driven by a few nearly collinear cases with $|\boldsymbol{\Sigma}_k| \simeq 0$ (see García-Escudero et al. [35] for details). The detection of spurious solutions constitutes an additional source of lack of robustness in (fuzzy) clustering.

F-TCLUST extends the hard clustering approach known as TCLUST [36] to the robust fuzzy clustering framework. Since TCLUST is based on the classification trimmed likelihood, F-TCLUST reduces to it when $m = 1$, so that the associated u_{ik} obtained by maximizing (2) take values in $\{0, 1\}$ – not in the interval – and sum to 1 for the non-trimmed cases and to 0 for the trimmed ones. A version of TCLUST for robust mixture modeling was introduced in García-Escudero et al. [37], where entire cases can be again trimmed. Note that both approaches – TCLUST and its version for robust mixture fitting – can be used to compute posterior probabilities according to the robustly estimated parameters, thus providing "soft assignments" for the non-trimmed cases. However, crucial differences from the membership degrees of F-TCLUST exist. First, F-TCLUST explicitly embeds fuzzy memberships into the objective function (2), treating them as parameters to be estimated jointly with the cluster weights, mean vectors and covariance matrices. Second, as noted in Section 1, the fuzzifier parameter m allows the user to control the degree of fuzziness.

3. CellFCLUST for cellwise robust fuzzy clustering

In this section, we introduce a novel fuzzy clustering method that handles cellwise outliers. The proposal, called *cellwise Fuzzy Clustering* (cellFCLUST), is formulated within a maximum likelihood framework and can be viewed as the F-TCLUST "counterpart" for dealing with outlying cells rather than entire outlying cases. The methodology is illustrated in Section 3.1, and a detailed description of the algorithm for its parameter estimation is provided in Section 3.2.

3.1. Methodology

The key feature of cellFCLUST is detecting unreliable cells and imputing them, rather than discarding them from the parameter estimation process. Identification is performed variable-by-variable by comparing the contribution of each unit to the objective function when its value is considered contaminated or not, while keeping the reliable values for the other variables. The unreliable cells, both contaminated or missing, are imputed by leveraging the reliable information per unit and cluster-specific relationships among variables. We denote by $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_n]'$ the $(n \times p)$ cellwise indicator matrix, where $w_{ij} = 1$ if the cell is reliable, and $w_{ij} = 0$ otherwise. Differently from the outlying cells, as mentioned, the (i, j)-positions of the missing values in a data matrix are known, and therefore the corresponding zeros into $\mathbf{W}$ can be set a priori. According to $\mathbf{w}_i$, we partition $\mathbf{x}_i$ into $\mathbf{x}_{i[\mathbf{w}_i]}$ and $\mathbf{x}_{i[\mathbf{w}_i^c]}$, where $\mathbf{w}_i^c = \mathbf{1}_p - \mathbf{w}_i$ with $\mathbf{1}_p$ being the unitary vector of dimension p. These two sub-vectors represent the reliable and unreliable cells for the i-th unit, respectively.

The cellFCLUST objective function to maximize is

$$J_{\text{cellFCLUST}}(\mathbf{W}, \mathbf{U}, \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) = \sum_{i=1}^n \sum_{k=1}^K u_{ik}^m \log(\pi_k \phi_{p[\mathbf{w}_i]}(\mathbf{x}_{i[\mathbf{w}_i]}; \boldsymbol{\mu}_{k[\mathbf{w}_i]}, \boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]})), \tag{4}$$

where both $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K$ are restricted in (4) to the sub-vectors and the sub-matrices, respectively, corresponding to the reliable cells for the i-th unit, and $p[\mathbf{w}_i]$ is their number. As in F-TCLUST, $\mathbf{U}$ represents the $(n \times K)$ membership matrix, the weights π_k are positive and such that $\sum_{k=1}^K \pi_k = 1$, and the fuzzifier parameter m must be greater than 1 to obtain strictly fuzzy memberships, since $m = 1$ results in crisp $0 - 1$ assignments for every unit, even though $u_{ik} \in [0, 1]$, as shown in Rousseeuw et al. [8]. The maximization of (4) is subject to both the eigenvalue-ratio constraint in (3) and the following one

$$\sum_{i=1}^n w_{ij} = h, \text{ for } j = 1, \dots, p, \tag{5}$$

which imposes that $h = \lceil (1-\alpha)n \rceil$ cells per variable are reliable. Here, α represents the proportion of flagged cells and it should be at most 0.25 to guarantee that pairs of variables overlap for some units, allowing their covariances to be safely computed. Note that F-TCLUST, instead of considering constraint (5), assumed $w_{ij} = 0$ for all $j = 1, \dots, p$ for the trimmed cases $\mathbf{x}_i$, without any imputation, and $w_{ij} = 1$ for all $j = 1, \dots, p$ for the untrimmed ones. We could have considered the maximization of a version of (4) in which the weights π_k are removed; the effect of this removal will be briefly discussed in Section 5.

Moreover, it is worth noting that casewise methodologies, such as F-TCLUST, systematically require trimming a large fraction of units located in the tails of the distribution in order to eliminate even a small fraction of cellwise contamination. Consequently, this trimming causes covariance matrix estimators to be biased downward due to the removal of part of the clusters' variability. For this reason, covariance matrix estimators resulting from casewise trimming typically require some form of adjustment. This is, for instance, the case of MCD-based covariance matrix estimators, which need to be multiplied by a consistency factor greater than one [38]. In contrast, cellFCLUST limits the amount of information discarded in the tails of the cluster components, making such a correction of the estimated covariance matrices much less critical.

3.2. An EM-inspired algorithm for cellFCLUST

We aim to solve the maximization problem of the objective function in (4) via an EM-inspired algorithm that extends the one for handling missing data in the model-based clustering framework [39], with an additional step for cellwise outlier detection. Specifically, the algorithm is composed of four alternating steps. After the initialization in *Step 0*, its key feature lies in *Step 1*, where $n-h$ cells per variable considered as potentially contaminated are flagged, and the corresponding positions in $\mathbf{W}$ are set to zero. Conditional on the previous update of $\mathbf{W}$, the membership values in $\mathbf{U}$ are updated in *Step 2* so as to decrease the objective function and directly satisfy the "hard contrast" property. Following the EM rationale and treating unreliable cells as missing entries, the parameters of the distribution for the missing part of the data, necessary for overall parameter estimation, are obtained in *Step 3*, which also leads to the imputation of potentially contaminated cells rather than their removal. Finally, *Step 4* uses that information to update the weights π_k, the mean vectors $\boldsymbol{\mu}_k$, and the covariance matrices $\boldsymbol{\Sigma}_k$, $k=1,\dots,K$ (the eigenvalue-ratio constraint in (3) must also be enforced by truncating their eigenvalues). A detailed description of these steps is provided below, along with the pseudocode for the entire procedure given in Algorithm 1.

Step 0. Initial solutions for the parameters are obtained using several applications of TCLUST method in García-Escudero et al. [36]. Specifically, TCLUST is applied to each variable and pairs of variables to initialize $\mathbf{W}$, and then on random subsets of variables to achieve feasible initializations for $\pi_k, \boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k, k=1,\dots,K$ (see Zaccaria et al. [20] for details). Accordingly, an initial solution for $\mathbf{U}$ is computed as described in *Step 2*.

Step 1. Given the current parameters, the membership matrix $\mathbf{U}$, and the actual configuration of $\mathbf{W}$, we update the latter column-by-column. Let consider the objective function in Eq. (4) as the sum of individual contributions, i.e. $J_{\text{cellFCLUST}}(\mathbf{W},\mathbf{U},\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K)=\sum_{i=1}^n J^{(i)}_{\text{cellFCLUST}}(\mathbf{w}_i,\mathbf{u}_i,\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K)$. It can be seen that

$$J^{(i)}_{\text{cellFCLUST}}(\mathbf{w}_i,\mathbf{u}_i,\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K)=\sum_{k=1}^K u_{ik}^m\log(\pi_k)-\frac{1}{2}\sum_{k=1}^K u_{ik}^m\Big[p[\mathbf{w}_i]\log(2\pi)+\log|\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}| + (\mathbf{x}_{i[\mathbf{w}_i]}-\boldsymbol{\mu}_{k[\mathbf{w}_i]})'\boldsymbol{\Sigma}^{-1}_{k[\mathbf{w}_i,\mathbf{w}_i]}(\mathbf{x}_{i[\mathbf{w}_i]}-\boldsymbol{\mu}_{k[\mathbf{w}_i]})\Big], \tag{6}$$

where the first addend does not depend on the elements of $\mathbf{W}$ and can thus be ignored. For each column j, we compare the individual contribution in (6) when we consider its i-th element reliable ($w_{ij}=1$) or contaminated ($w_{ij}=0$) as follows

$$\begin{aligned}\Delta_{ij}&=J^{(i)}_{\text{cellFCLUST}}(\mathbf{w}_i,\mathbf{u}_i,\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K|w_{ij}=1)-J^{(i)}_{\text{cellFCLUST}}(\mathbf{w}_i,\mathbf{u}_i,\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K|w_{ij}=0)\\&=-\frac{1}{2}\sum_{k=1}^K u_{ik}^m\left[\log(2\pi)+\log(C_{ij(k)})+\frac{(x_{ij}-\hat{x}_{ij(k)})^2}{C_{ij(k)}}\right],\end{aligned} \tag{7}$$

where $\hat{x}_{ij(k)}=\boldsymbol{\mu}_{k[j]}+\boldsymbol{\Sigma}_{k[j,\mathbf{w}_i]}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]})^{-1}(\mathbf{x}_{i[\mathbf{w}_i]}-\boldsymbol{\mu}_{k[\mathbf{w}_i]})$ and $C_{ij(k)}=\boldsymbol{\Sigma}_{k[j,j]}-\boldsymbol{\Sigma}_{k[j,\mathbf{w}_i]}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]})^{-1}\boldsymbol{\Sigma}_{k[\mathbf{w}_i,j]}$ are two scalars denoting the expectation and variance, respectively, of the cell associated with the j-th variable in the i-th unit, conditional on the other reliable cells for the i-th unit and given the parameters of the k-th cluster. The proof for the final expression of (7) is provided in the Supplementary Material.

By sorting the Δ_{ij} values in an increasing order, i.e., $\Delta_{(1)j}\le\Delta_{(2)j}\le\cdots\le\Delta_{(n)j}$, we flag as unreliable the cells with indexes $\{i:\Delta_{ij}<\Delta_{(n-h)j}\}$, where $h=\lceil(1-\alpha)n\rceil$, and set the corresponding elements of the j-th column of $\mathbf{W}$ to zero. Equivalently, the cells with $\{i:\Delta_{ij}\ge\Delta_{(n-h)j}\}$ are considered reliable, which is denoted by ones into the j-th column of $\mathbf{W}$. Therefore, the supposedly contaminated cells are those whose Δ_{ij} is negative or small.

If missing values occur for the j-th variable, the corresponding cells of $\mathbf{W}$ are set to zero. Consequently, the Δ_{ij} values are computed only on the observed cells. This approach may result in varying proportions of zeros across variables, depending on the extent of missingness they contain. However, using α to account for both the contaminated and missing cells – and therefore grounding h on n, independently of the number of missing values – could potentially lead to an incorrect identification of the truly contaminated cells. In general, we recommend excluding variables with a high proportion of missing data, particularly when moderate contamination is expected; specifically, the proportion of missing *and* outlying values must not exceed 0.25, as also mentioned in Section 3.1.

Algorithm 1: CellFCLUST.

Input: $\mathbf{X}$, K, α, c, m, ϵ, *maxiter*

$t \leftarrow 0$; $J_0 \leftarrow -\infty$;

Initialization: $\mathbf{W}^{(0)}$, $\boldsymbol{\theta}^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ *and* $\mathbf{U}^{(0)}$ ▷ *See Step 0*

Repeat

  Update $\mathbf{W}$ (conditional on $\mathbf{U}$ and $\boldsymbol{\theta}$): ▷ See *Step 1*

  for $j = 1, \dots, p$ **do**

    • Compute

$$\Delta_{ij} \leftarrow -\frac{1}{2}\sum_{k=1}^{K}\left(u_{ik}^{(t)}\right)^m\left[\log(2\pi) + \log(C_{ij(k)}) + \frac{(x_{ij} - \hat{x}_{ij(k)})^2}{C_{ij(k)}}\right], \quad \text{(Eq. 7)}$$

    where $\hat{x}_{ij(k)} = \boldsymbol{\mu}_{k[j]}^{(t)} + \boldsymbol{\Sigma}_{k[j,\mathbf{w}_i]}^{(t)}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}^{(t)})^{-1}(\mathbf{x}_{i[\mathbf{w}_i]} - \boldsymbol{\mu}_{k[\mathbf{w}_i]}^{(t)})$ and $C_{ij(k)} = \boldsymbol{\Sigma}_{k[j,j]}^{(t)} - \boldsymbol{\Sigma}_{k[j,\mathbf{w}_i]}^{(t)}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}^{(t)})^{-1}\boldsymbol{\Sigma}_{k[\mathbf{w}_i,j]}^{(t)}$ for $\mathbf{w}_i = \mathbf{w}_i^{(t)}$.

    • Set

$$w_{ij}^{(t+1)} \leftarrow \begin{cases} 1 & \text{if } \Delta_{ij} \geq \Delta_{(n-h)j} \\ 0 & \text{if } \Delta_{ij} < \Delta_{(n-h)j} \end{cases}, \text{ for } i = 1, \dots, n,$$

    where $\Delta_{(1)j} \leq \Delta_{(2)j} \leq \dots \Delta_{(n)j}$ are the non-decreasing Δ_{ij} values.

  end

  $\mathbf{w}_i \leftarrow \mathbf{w}_i^{(t+1)}$ for $i = 1, \dots, n$

  Update $\mathbf{U}$ (conditional on $\mathbf{W}$ and $\boldsymbol{\theta}$): ▷ See *Step 2*

$$u_{ik}^{(t+1)} \leftarrow \begin{cases} I\left\{f_{ik} = \max\limits_{k'=1,\dots,K} f_{ik'}\right\} & \text{if } \max\limits_{k'=1,\dots,K} f_{ik'} \geq 1 \\ \left(\sum_{k'=1}^{K}\left(\dfrac{\log(f_{ik})}{\log(f_{ik'})}\right)^{\frac{1}{m-1}}\right)^{-1} & \text{if } \max\limits_{k'=1,\dots,K} f_{ik'} < 1 \end{cases}, \quad \text{(Eq. 8)}$$

  where $f_{ik} = \pi_k^{(t)}\phi_{p[\mathbf{w}_i]}\left(\mathbf{x}_{i[\mathbf{w}_i]}; \boldsymbol{\mu}_{k[\mathbf{w}_i]}^{(t)}, \boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}^{(t)}\right)$.

  $u_{ik} \leftarrow u_{ik}^{(t+1)}$ for $i = 1, \dots, n$, $k = 1, \dots, K$

  Update $\boldsymbol{\theta}$ (conditional on $\mathbf{U}$ and $\mathbf{W}$): ▷ See *Steps 3–4*

  for $k = 1, \dots, K$ **do**

    $\hat{\mathbf{x}}_{i[\mathbf{w}_i^c](k)} \leftarrow \boldsymbol{\mu}_{k[\mathbf{w}_i^c]}^{(t)} + \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i]}^{(t)}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}^{(t)})^{-1}(\mathbf{x}_{i[\mathbf{w}_i]} - \boldsymbol{\mu}_{k[\mathbf{w}_i]}^{(t)})$ for $i = 1, \dots, n$

    $\tilde{\mathbf{x}}_{i(k)} \leftarrow (\mathbf{x}_{i[\mathbf{w}_i]}, \hat{\mathbf{x}}_{i[\mathbf{w}_i^c](k)})$ for $i = 1, \dots, n$

    • Compute

$$\pi_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^n u_{ik}^m}{\sum_{i=1}^n\sum_{k=1}^K u_{ik}^m}, \quad \text{(Eq. 12)}$$

$$\boldsymbol{\mu}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^n u_{ik}^m \tilde{\mathbf{x}}_{i(k)}}{\sum_{i=1}^n u_{ik}^m}, \quad \text{(Eq. 13)}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} \leftarrow \frac{\sum_{i=1}^n u_{ik}^m\left[(\tilde{\mathbf{x}}_{i(k)} - \boldsymbol{\mu}_k^{(t+1)})(\tilde{\mathbf{x}}_{i(k)} - \boldsymbol{\mu}_k^{(t+1)})' + \mathbf{C}_{i[\mathbf{w}_i^c,\mathbf{w}_i^c](k)}\right]}{\sum_{i=1}^n u_{ik}^m}, \quad \text{(Eq. 14)}$$

    where $\mathbf{C}_{i[\mathbf{w}_i^c,\mathbf{w}_i^c](k)} = \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i^c]}^{(t)} - \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i]}^{(t)}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}^{(t)})^{-1}\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i^c]}^{(t)}$ (truncation of the eigenvalues of $\boldsymbol{\Sigma}_k^{(t+1)}$ depending on c may be required).

  end

  $t \leftarrow t + 1$

  Compute objective function: $J_t \leftarrow J_{\text{cellFCLUST}}(\mathbf{W}^{(t)}, \mathbf{U}^{(t)}, \boldsymbol{\theta}^{(t)})$

until $J_t - J_{t-1} < \epsilon$ *or* $t =$ *maxiter*;

Step 2. Given the cellwise indicator matrix and the current parameters, the membership values are updated as

$$u_{ik} = \begin{cases} I\left\{ f_{ik} = \max\limits_{k'=1,\dots,K} f_{ik'} \right\} & \text{if } \max\limits_{k'=1,\dots,K} f_{ik'} \geq 1 \\ \left(\sum_{k'=1}^{K} \left(\dfrac{\log(f_{ik})}{\log(f_{ik'})} \right)^{\frac{1}{m-1}} \right)^{-1} & \text{if } \max\limits_{k'=1,\dots,K} f_{ik'} < 1 \end{cases}, \tag{8}$$

where

$$f_{ik} = \pi_k \phi_{p[\mathbf{w}_i]}(\mathbf{x}_{i[\mathbf{w}_i]}; \boldsymbol{\mu}_{k[\mathbf{w}_i]}, \boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}),$$

and $I\{\cdot\}$ represents the indicator function. In the first case, the i-th unit is fully assigned to the k-th cluster, while in the second case, the assignment is fuzzy. This is a desirable property of a fuzzy clustering approach since not all units necessarily require a soft assignment. Some of them, especially those in the "core" of the clusters, can be unequivocally assigned and, thus, this update of u_{ik} automatically results in the "high contrast" property in Rousseeuw et al. [8].

In order to justify the update of u_{ik} in (8), note that the maximization of the objective function (4) is equivalent to the minimization of $\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m D_{ik}$, with $D_{ik} = -\log(f_{ik})$. For fixed $\mathbf{W}$ and $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$, the function to be minimized is separable with respect to the index i, and hence it suffices to determine how to minimize

$$\sum_{k=1}^{K} u_{ik}^m D_{ik} \tag{9}$$

with respect to $u_{i1}, \dots, u_{iK}$, for a fixed i. If $\max_{k'=1,\dots,K} f_{ik'} < 1$, that is, whenever $D_{ik} > 0$ for $k = 1, \dots, K$, then (9) is a strictly convex function and a unique global minimizer exists. Standard Lagrange multiplier theory shows that the values u_{ik} must be proportional to $D_{ik}^{-1/(m-1)}$, as traditionally occurs in FKM. When combined with the constraint $\sum_{k=1}^{K} u_{ik} = 1$, updating u_{ik} results in fuzzy memberships, that is the second case in (8). On the other hand, if $\max_{k'=1,\dots,K} f_{ik'} \geq 1$ and we define $k^* = \arg\max_{k'=1,\dots,K} f_{ik'}$, it trivially follows that

$$\sum_{k=1}^{K} u_{ik}^m \log(f_{ik}) \leq \sum_{k=1}^{K} u_{ik}^m \log(f_{ik^*}) \leq \log(f_{ik^*}) \sum_{k=1}^{K} u_{ik} = \log(f_{ik^*}),$$

and, consequently, the largest decrease in (9) occurs when $u_{ik^*} = 1$ and $u_{ik} = 0$ for all $k \neq k^*$. Therefore, the complete update for the membership values in (8) is demonstrated.

Step 3. Given the cellwise indicator matrix and the current parameters, we compute the parameters of the distribution for the unreliable data $\mathbf{x}_{i[\mathbf{w}_i^c]}$ conditional on the observed $\mathbf{x}_{i[\mathbf{w}_i]}$ as

$$\boldsymbol{\mu}_{k[\mathbf{w}_i^c|\mathbf{w}_i]} = \boldsymbol{\mu}_{k[\mathbf{w}_i^c]} + \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i]}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]})^{-1}(\mathbf{x}_{i[\mathbf{w}_i]} - \boldsymbol{\mu}_{k[\mathbf{w}_i]}) := \hat{\mathbf{x}}_{i[\mathbf{w}_i^c](k)}, \tag{10}$$

$$\boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i^c|\mathbf{w}_i]} = \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i^c]} - \boldsymbol{\Sigma}_{k[\mathbf{w}_i^c,\mathbf{w}_i]}(\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]})^{-1}\boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i^c]} := \mathbf{C}_{i[\mathbf{w}_i^c,\mathbf{w}_i^c](k)}. \tag{11}$$

Further details on the derivation of these parameters are provided in the Supplementary Material.

Step 4. Given the membership matrix and considering the completed data $\tilde{\mathbf{x}}_{i(k)} = (\mathbf{x}_{i[\mathbf{w}_i]}, \hat{\mathbf{x}}_{i[\mathbf{w}_i^c](k)})$, for $i = 1, \dots, n$ and $k = 1, \dots, K$, where both contaminated and missing values are imputed through *Step 3*, we update the parameters as follows

$$\pi_k = \frac{\sum_{i=1}^{n} u_{ik}^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^m}, \tag{12}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} u_{ik}^m \tilde{\mathbf{x}}_{i(k)}}{\sum_{i=1}^{n} u_{ik}^m}, \tag{13}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{n} u_{ik}^m \left[(\tilde{\mathbf{x}}_{i(k)} - \boldsymbol{\mu}_k)(\tilde{\mathbf{x}}_{i(k)} - \boldsymbol{\mu}_k)' + \mathbf{C}_{i[\mathbf{w}_i^c,\mathbf{w}_i^c](k)} \right]}{\sum_{i=1}^{n} u_{ik}^m}. \tag{14}$$

In (14), the term $\mathbf{C}_{i[\mathbf{w}_i^c,\mathbf{w}_i^c](k)}$ only affects the covariance of the imputed values. Further details on the derivation of (14), which corresponds to the traditional covariance matrix estimation in presence of missing values, are provided in the Supplementary Material. If the update of the covariance matrices in (14) does not satisfy the eigenvalue-ratio constraint in (3), we apply the efficient eigenvalue-truncation procedure proposed by Fritz et al. [40]. This ensures that the constraint in (3) is met.

At the end of *Step 4*, the objective function in (4) is computed. If the focus is on clusters with $\sum_{i=1}^{n} u_{i1}^m = \dots = \sum_{i=1}^{n} u_{iK}^m$, the objective function in (4) should be modified by removing the π_k weights, which approximately corresponds to set $\pi_k = 1/K$ in *Step 4* throughout all iterations (this is an exact correspondence for $m = 1$). *Steps 1–4* are repeated until the increase (and, more precisely, non-decrease) in the objective function falls below a small tolerance value ϵ (e.g., 10^{-6} in our experiments), or until a maximum number of iterations (500 in the examples shown in this work) is reached. The cellFCLUST objective function is non-decreasing at every iteration (see the Supplementary Material for details on the algorithm's monotonicity). To increase the chances of finding the global constrained maximum of the objective function, the algorithm should be run several times with different initializations, retaining the best solution.

4. Simulation study

We conduct a simulation study to evaluate the performance of cellFCLUST relative to several alternative methodologies across three main goals: cluster recovery, parameter estimation, and outlier detection. The competitors we consider are fuzzy clustering methods, both robust and non-robust. It is worth noting that the robust approaches included in the simulation study are designed to handle casewise outliers. Indeed, to the best of our knowledge, prior to cellFCLUST, no fuzzy clustering methodology addressing cellwise contamination has been proposed in the literature, as highlighted in Section 1.

The first competitor is F-TCLUST, which represents the casewise counterpart of our proposal and it is run via the R package **tclust** (Fritz et al. [41], R version 3.3). Two other classes of methods encompass those which are not specifically tailored for outliers: fuzzy k-means (FKM, R function `FKM`) and Gustafson, Kessel and Babuska-like fuzzy k-means (FKMGKB, R function `FKM.gkb`), which relaxes the spherical assumption for the clusters [42]. Their robust versions based on the identification of a noise cluster are implemented via the R functions `FKM.noise` and `FKM.gkb.noise`, respectively. All these methods are included in the R package **fclust** [43]. Additionally, we also consider the Unsupervised Fuzzy Trimmed C Prototypes (UFTCP, Kim et al. [23]), where fuzzy c-means is extended via a trimming approach.

Finally, we compare the proposal with cellGMM [20], a mixture model designed to deal with cellwise contamination. As discussed in Section 1, although mixture models provide soft assignments, they do not offer the same flexibility as fuzzy clustering methods in capturing cluster overlap. Results for cellGMM are compared with those for cellFCLUST in the Supplementary Material.

4.1. Design of the simulation study

Two scenarios with varying numbers of units and clusters are considered, while the number of variables is fixed to $p = 10$. In *Scenario 1*, $n = 250$, $G = 2$: 30% of units are generated from the first cluster, with $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\Sigma}_1 = [\sigma_{jl} = (0.6^{|j-l|})/16, j, l = 1, \dots, p]$; and 70% from the second cluster, with $\boldsymbol{\mu}_2 = [\mu_j = (-1)^j \times 0.5, j = 1, \dots, p]$ and $\boldsymbol{\Sigma}_2 = [\sigma_{jl} = ((-0.6)^{|j-l|})/16, j, l = 1, \dots, p]$. In *Scenario 2*, $n = 500$, $G = 4$, and the sizes of the clusters are more balanced: the first two clusters, each containing 20% of the units, have the same mean vectors and covariance matrices as in Scenario 1; the other two clusters, each composed of 30% of the units, have mean vectors $\boldsymbol{\mu}_3 = [\mu_j = (j \mod 3) - 1, j = 1, \dots, p]$ and $\boldsymbol{\mu}_4 = [\mu_j = \big((j+1 \mod 4) - 1\big)/2, j = 1, \dots, p]$, and covariance matrices $\boldsymbol{\Sigma}_3 = [\sigma_{jl} = (0.7^{|j-l|})/16, j, l = 1, \dots, p]$ and $\boldsymbol{\Sigma}_4 = [\sigma_{jl} = \big((-0.7)^{|j-l|}\big)/16, j, l = 1, \dots, p]$. The maximum overlap between pairs of clusters – where the overlap is defined in Maitra and Melnykov [44] as the sum of the two misclassification probabilities – ranges from 0.01 to 0.05 in both scenarios, indicating moderate cluster separation, which represents a suitable configuration for fuzzy clustering methodologies. For each scenario, 100 random samples are obtained and contaminated with 0% (baseline), 1%, 5%, and 10% of outlying cells per variable (i.e., in each column of the data matrix) randomly drawn from a uniform distribution on the interval $[-30, 30]$. The contamination is also performed in such a way that no outlying unit lies within the 99-th percentile ellipsoid of any cluster, comparing their Mahalanobis distance, computed using the parameters employed for data generation, with the 99-th percentile of a chi-squared distribution with p degrees of freedom.

The fuzzifier parameter m is set to 2 for cellFCLUST and F-TCLUST, to 1.35 for FKM, FKM.noise and UFTCP, and to 1.5 for FKM.gkb and FKM.gkb.noise. To select these values, we compare the proportion of Weak Assignments (WA) – defined as the proportion of units whose highest membership is below 0.90 – resulting from the estimation of each methodology under $\alpha = 0$ (baseline) with the theoretical ones. The latter are computed from the membership matrices in (8) using the generating parameters and $\alpha = 0$, yielding WA proportions of 0.05 and 0.07 in the two scenarios, respectively. It should be noted that we focus on u_{ik} for the unit i belonging to cluster k rather than the fuzziness itself, i.e., $u_{ik'}$ with $k' \neq k$, since the latter depends on the number of clusters considered. Additionally, analyzing the highest membership value across clusters provides information on the "weight" a unit has in the cluster it is assigned to: if it is a representative unit of that cluster, its membership will be high – potentially approaching one – regardless of the number of clusters. In our experience, when a unit is weakly assigned to a cluster, the remaining membership is typically shared among only a few clusters, that is, it is not evenly split across all $K - 1$ clusters. For the methods that require eigenvalue-ratio constraints given by c, i.e., cellFCLUST and F-TCLUST, this is set according to the ratio of the eigenvalues computed from the true (i.e., data-generating) covariance structure. For FKM.gkb and FKM.gkb.noise, we specify precise constraints: the volume parameters ρ_g as obtained from the aforementioned covariance matrices, and $\gamma = 0.1$ to ensure numerical stability. These two parameters are therefore changed from their default values, which were set only to prevent numerical singularities. Instead, we apply fine-tuning to choose more appropriate values that improve the performance of these clustering methods. The flagging level α is fixed to the true generating contamination for cellFCLUST, while the trimming one is set to 0.25 for F-TCLUST, as previously employed in Zaccaria et al. [20]. However, due to the nature of cellwise contamination, with $p = 10$ and cellwise outliers representing 1%, 5%, and 10% of the cells, the percentage of cases affected by at least one outlying cell can reach 10%, 40%, and 65% of the total cases, respectively. For noise clustering methodologies, the number of outliers is automatically selected within the algorithm. Similarly, UFTCP chooses the appropriate outlier proportion from a set of possible levels ranging from 0 to 0.5, in increments of 0.05, ultimately retaining the value that maximizes a density criterion [34] as the validity measure.

4.2. Results

The results of the simulation study are evaluated in terms of cluster recovery, parameter estimation, and outlier detection. In the following sections, we analyze the performance of the proposed and alternative methods focusing separately on these objectives.

4.2.1. Cluster recovery

Cluster recovery is assessed through the Misclassification Rate (MR) and the Adjusted Rand Index (ARI, Hubert and Arabie [45]) by comparing the theoretical and estimated assignments of units to clusters, determined from the maximum values of u_{ik} across $k = 1 \dots K$ for each $i = 1, \dots, n$. Lower MR values and higher ARI values indicate better recovery, with perfect agreement between theoretical and estimated partitions when MR is zero and ARI is one.

As shown in Fig. 2, cellFCLUST has substantially lower MR and higher ARI (close to 1) in both scenarios across contamination levels, since it is the only model designed to address cellwise contamination. The difference with the other methods increases as the proportion of outlying cells in the data grows, since even robust (casewise) methodologies can break down when contamination spreads across the cases. Among the competitors, FKM, FKM.gkb, and UFTCP show the highest MR (and lowest ARI) for cellwise contamination of 1%. However, at 5% and 10% contamination, only cellFCLUST and FKM.gkb.noise maintain good clustering performance. This result is mainly due to the large number of cases containing outlying cells as the contamination level increases, and their ability to accommodate non-spherical clusters. Moreover, unlike F-TCLUST, where α is fixed in advance, FKM.gkb.noise automatically selects the proportion of outlying cases, allowing it to stabilize at lower MR (and higher ARI) values than cellFCLUST's natural casewise counterpart. Although this flexibility, cellFCLUST consistently outperforms FKM.gkb.noise due to its higher efficiency in handling cellwise outliers, without discarding or downweighting entire contaminated units during the estimation procedure.

4.2.2. Parameter estimation

Parameter estimation is evaluated via the Root Mean Squared Error (RMSE) of the estimated membership matrix $\mathbf{U}$ and cluster mean vectors $\{\boldsymbol{\mu}_k\}_{i=1}^K$, and the Kullback-Leibler (KL) discrepancy of the estimated cluster covariance matrices $\{\boldsymbol{\Sigma}_k\}_{i=1}^K$, compared to the corresponding theoretical parameters used to generate the data (Section 4.1). To properly perform this comparison, the label switching problem is resolved by finding the cluster label ordering that minimizes RMSE between the theoretical and estimated cluster means over all possible permutations.

The results for the membership matrix are shown in Fig. 2, while those for the cluster mean vectors and covariance matrices are reported in Fig. 3. RMSE of the membership matrices exhibits a pattern similar to MR and ARI, as does RMSE of the cluster mean vectors, which deteriorates at 1% of contamination for the non-robust models. Additionally, the latter significantly degrades at 5% and 10% for F-TCLUST and FKM.noise, whereas the FKM.gkb.noise estimates remain acceptable but suboptimal with respect to cellFCLUST. In the estimation of the cluster covariance matrices, cellFCLUST outperforms the competitors from the first contamination level, as indicated by the KL discrepancy measure. On the other hand, F-TCLUST provides good estimates only when the number of contaminated cases in at least one cell is smaller than the number of trimmed cases, handling at most 1% of cellwise contamination when $p = 10$. Unlike the other indices, the difference between cellFCLUST and FKM.gkb.noise becomes evident in the estimation of the covariance matrices. Indeed, for FKM.gkb-type models – especially the noise version – increasing the smallest eigenvalue of the covariance matrix $\boldsymbol{\Sigma}_k$ to ensure its determinant equals ρ_g does not yield results comparable to those obtained with the efficient procedure proposed in Fritz et al. [40], which provides a closed-form solution for the truncated eigenvalues satisfying constraint (3). This is particularly pronounced in Fig. 3 at 1% contamination, where the two models implementing the aforementioned procedure – namely, cellFCLUST and F-TCLUST – produce better estimates of the covariance matrices. It is worth noting that F-TCLUST is implemented with $\alpha = 0.25$, while the maximum percentage of cases with at least one contaminated cell is 10% when the true $\alpha = 0.01$. At higher contamination levels, the KL discrepancy of FKM.gkb.noise is lower than that of F-TCLUST due to the larger proportion of outliers not flagged by the latter. Regarding cellFCLUST, the combined effect of the eigenvalue-ratio constraint and the cellwise outlier detection mechanism allows this methodology to achieve lower KL values than FKM.gkb.noise.

4.2.3. Outlier detection

We compute the True Positive Rate (TPR), False Positive Rate (FPR), and False Negative Rate (FNR) as the proportions of cells correctly flagged, reliable cells incorrectly flagged as contaminated, and contaminated cells not recognized as such, respectively, to evaluate outlier detection. For the robust (casewise) competitors, all the cells of the outlying cases are considered contaminated. It is worth noting that a high value of FNR is more problematic than a high value of FPR for a robust model, as it implies that contaminated cells are not properly accounted for and their values affect the parameter estimates.

Looking at Table 1, we can notice that TPR shows a near-perfect score for cellFCLUST and FKM.gkb.noise in both scenarios and across contamination levels, whereas F-TCLUST and FKM.noise start to deteriorate at 5%, and UFTCP at 10%. Additionally, cellFCLUST is the only model capable of achieving a zero FPR. It is worth recalling that, for the casewise robust models, all cells of the cases flagged as outlier are considered as contaminated, leading to a higher number of false positives. Consequently, the reliable information discarded by casewise robust models results in higher errors even when outlying values are detected. This effect is noticeable when comparing the overall performance of cellFCLUST and FKM.gkb.noise.

Similarly to TPR, the best performance for FNR is achieved by cellFCLUST and FKM.gkb.noise. As mentioned before, a high FNR is more detrimental than a high FPR, and this effect is also reflected in cluster recovery and parameter estimation. This can be seen, for instance, in F-TCLUST: at 1% contamination, it exhibits a high FPR (0.25) and no false negatives, resulting in good cluster recovery (see Fig. 2). When contamination increases, FPR slightly decreases (0.23), while FNR increases substantially (0.31 at 5% of contamination), leading to a marked deterioration in clustering performance, with a sharp increase in MR and a corresponding decrease in ARI.

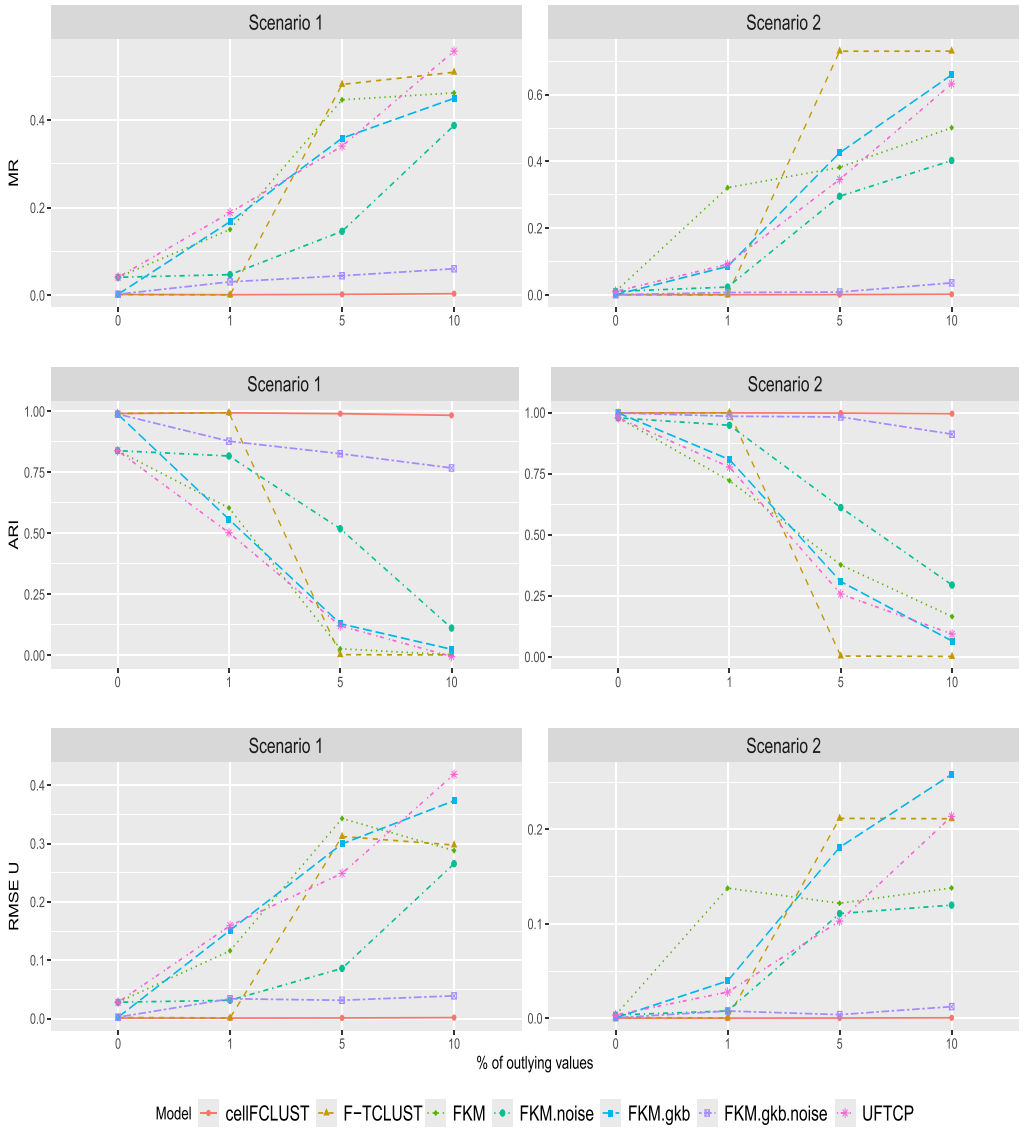


Fig. 2. Results of the simulation study: Misclassification Rate (MR), Adjusted Rand Index (ARI), and Root Mean Squared Error (RMSE) of the membership matrix averaged over 100 samples per scenario, percentage of contamination, and model. For F-TCLUST, FKM.noise, FKM.gkb.noise, and UFTCP the indices are computed only on the units not flagged as outliers.

5. Effects of the cellFCLUST tuning parameters

The proposed methodology depends on several tuning parameters: the number of clusters (K), the flagging level (α), the constant for the eigenvalue-ratio constraint (c), and the fuzzifier (m). Additionally, we can also consider the scale factor (S), which arises when we modify x_{ij} by x_{ij}/S for $i = 1 \dots, n$ and $j = 1, \dots, p$, as a further parameter. The effect of the scale factor was already noted in Fritz et al. [22] and is inherent when considering general covariance matrices and the definition of the membership values in (8). However, S can be viewed as an actionable tuning parameter as well.

We illustrate the role of all tuning parameters using one of the samples generated in Scenario 1 of the simulation study (Section 4) with 5% of contamination. It has to be highlighted that these parameters are interrelated, and a unified approach for their setting is needed. In this section, we provide tools for helping the user in selecting these parameters, which are particularly useful when no prior information is available, as is often the case in real-data applications.

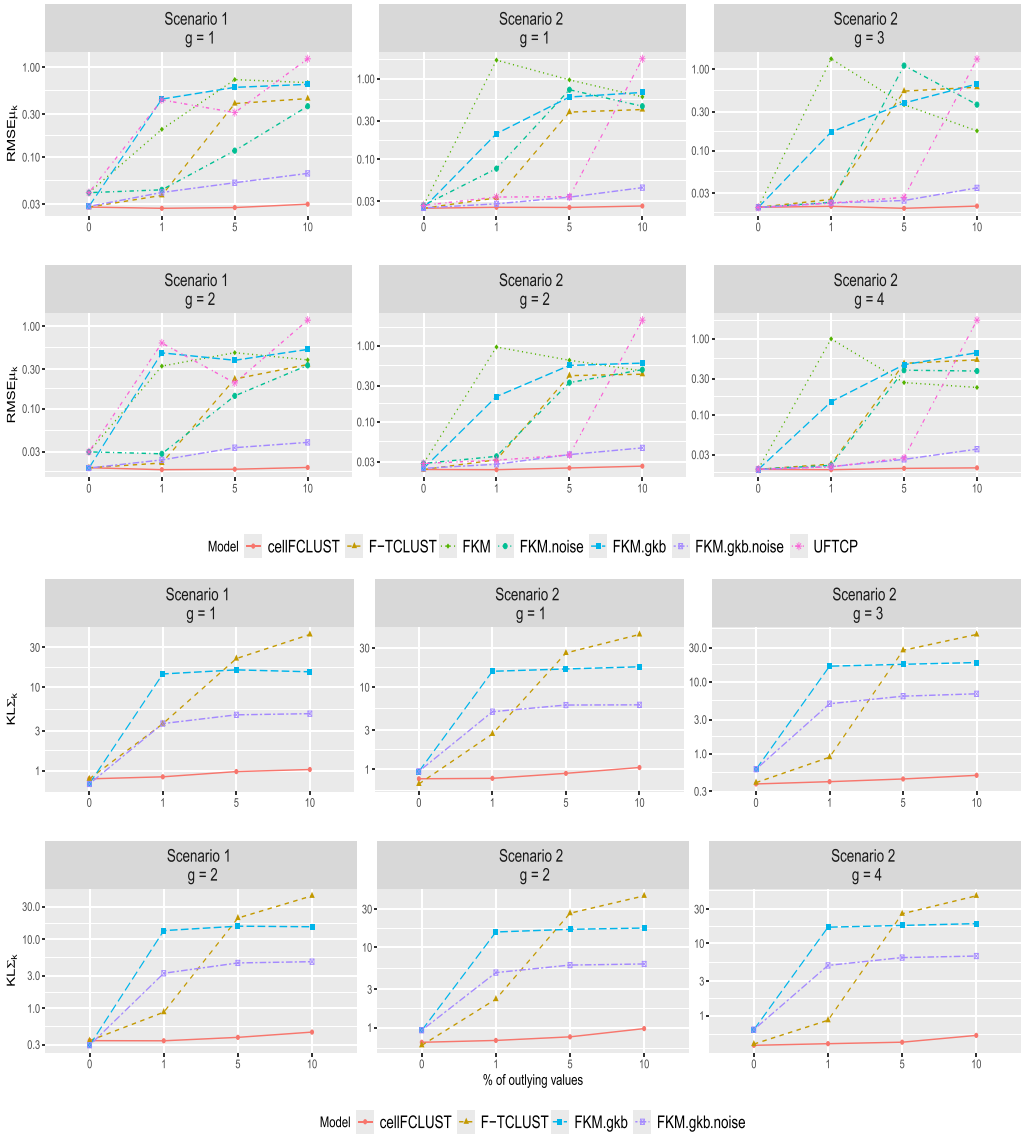


Fig. 3. Results of the simulation study: Root Mean Squared Error (RMSE) of the cluster mean vectors and Kullback-Leibler (KL) discrepancy for the cluster covariance matrices averaged over 100 samples per scenario, percentage of contamination, and model. The values are represented via log-transformation, while the y-axis ticks are labeled using the original scale.

5.1. Number of clusters K and flagging level α

The number of clusters in a data set cannot be chosen independently of the level of cells flagged as contaminated per variable. In this framework, we evaluate the behavior of the curves representing the objective function in (4) at convergence, by running cellFCLUST with different values of K and α, when $c = 14$ and $m = 2$. We vary α over the set $\{0, 0.01, 0.025, 0.05, 0.075, 0.10\}$, considering that the true contamination level is 0.05. As shown in Fig. 4, the objective functions differ significantly when $\alpha = 0, 0.01, 0.025$, since underestimating the contamination level requires additional clusters beyond those originally generated. It is worth noting that when $K = 4$ and $\alpha = 0$ or 0.01, cellFCLUST finds solutions with an empty cluster, causing the algorithm to stop – the corresponding objective function values are not reported in the figure. On the other hand, when $\alpha = 0.05$, which is the so generated contamination level in the data set, the objective functions for $K = 2, 3, 4$ are close to each other, indicating that two is a suitable choice for the number of clusters. Indeed, $K = 2$, which corresponds to the true value, captures the underlying structure of the data well, and increasing this number does not lead to a meaningful improvement in the objective function [46].

Table 1
Memberships and outlier detection: proportion of Weak Assignments (WA), True and False Positive Rate (TPR and FPR), and False Negative Rate (FNR) averaged over 100 samples per scenario, percentage of contamination, and model.

		Scenario 1				Scenario 2			
% out	Method	WA	TPR	FPR	FNR	WA	TPR	FPR	FNR
0	cellFCLUST	0.03	–	–	–	0.05	–	–	–
	F-TCLUST	0.07	–	–	–	0.12	–	–	–
	FKM	0.10	–	–	–	0.04	–	–	–
	FKM.noise	0.10	–	–	–	0.04	–	–	–
	FKM.gkb	0.07	–	–	–	0.05	–	–	–
	FKM.gkb.noise	0.09	–	–	–	0.08	–	–	–
	UFTCP	0.10	–	–	–	0.04	–	–	–
1	cellFCLUST	0.03	1.00	0.00	0.00	0.05	1.00	0.00	0.00
	F-TCLUST	0.00	1.00	0.25	0.00	0.01	1.00	0.24	0.00
	FKM	0.18	–	–	–	0.13	–	–	–
	FKM.noise	0.11	0.92	0.06	0.08	0.06	0.79	0.07	0.21
	FKM.gkb	0.20	–	–	–	0.22	–	–	–
	FKM.gkb.noise	0.57	1.00	0.08	0.00	0.47	1.00	0.09	0.00
	UFTCP	0.17	0.93	0.09	0.07	0.14	1.00	0.09	0.00
5	cellFCLUST	0.04	0.99	0.00	0.01	0.06	0.99	0.00	0.01
	F-TCLUST	1.00	0.69	0.23	0.31	1.00	0.67	0.23	0.33
	FKM	0.39	–	–	–	0.43	–	–	–
	FKM.noise	0.28	0.69	0.23	0.31	0.26	0.60	0.20	0.40
	FKM.gkb	0.35	–	–	–	0.47	–	–	–
	FKM.gkb.noise	0.22	1.00	0.36	0.00	0.17	1.00	0.37	0.00
	UFTCP	0.45	0.99	0.37	0.01	0.45	1.00	0.40	0.00
10	cellFCLUST	0.05	0.99	0.00	0.01	0.07	0.99	0.00	0.01
	F-TCLUST	1.00	0.47	0.23	0.53	1.00	0.46	0.23	0.54
	FKM	0.66	–	–	–	0.70	–	–	–
	FKM.noise	0.53	0.60	0.31	0.40	0.56	0.50	0.24	0.50
	FKM.gkb	0.33	–	–	–	0.54	–	–	–
	FKM.gkb.noise	0.22	0.99	0.60	0.01	0.20	0.98	0.60	0.02
	UFTCP	0.60	0.81	0.45	0.19	0.65	0.80	0.45	0.20

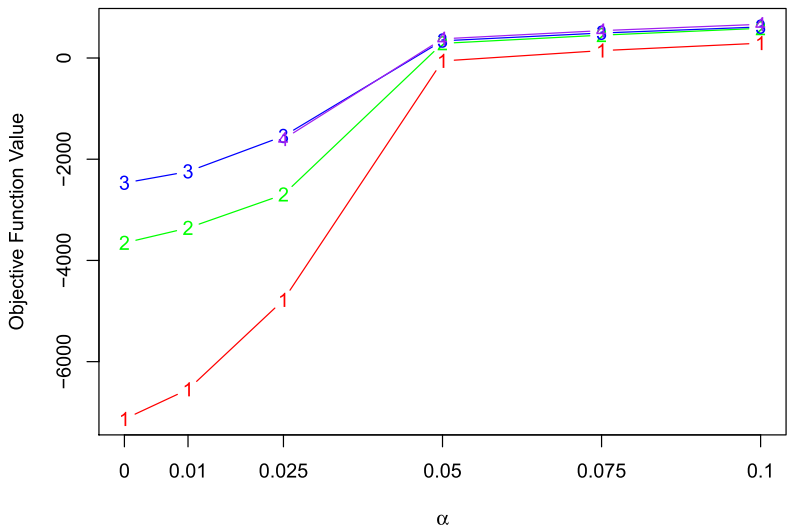


Fig. 4. Synthetic data: objective function curves.

We further investigate how the cellFCLUST results change with the flagging level α to support the previous finding by analyzing the behavior of $\{\Delta_{ij}\}_{i=1}^{n}$ for $j = 1, \ldots, p$. These are computed as defined in (7), using the parameters estimated at convergence for the selected K. For each variable, the ordered values $\Delta_{(i)j}$, where $\Delta_{(1)j} \leq \Delta_{(2)j} \leq \ldots \leq \Delta_{(n)j}$, can be plotted. In each plot – which we discuss in detail later – we identify the knee point, corresponding to the maximum distance between each $\Delta_{(i)j}$ and the straight line connecting the endpoints $\Delta_{(1)j}$ and $\Delta_{(n)j}$, as α varies. Therefore, we have p knee points (one per variable) for each α. Fig. 5 shows the median difference between the $p = 10$ knee points and the corresponding α, for different values of α. The difference approaches zero

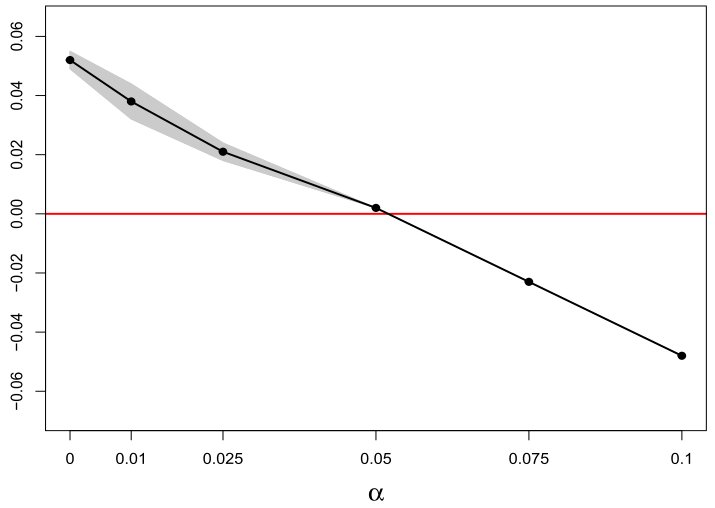


Fig. 5. Synthetic data: difference between the knee point of $\{\Delta_{ij}\}_{i=1}^{n}$ and the α value. The curve shows the median differences across variables as α varies, with $K = 2$. The shaded gray area represents the variability.

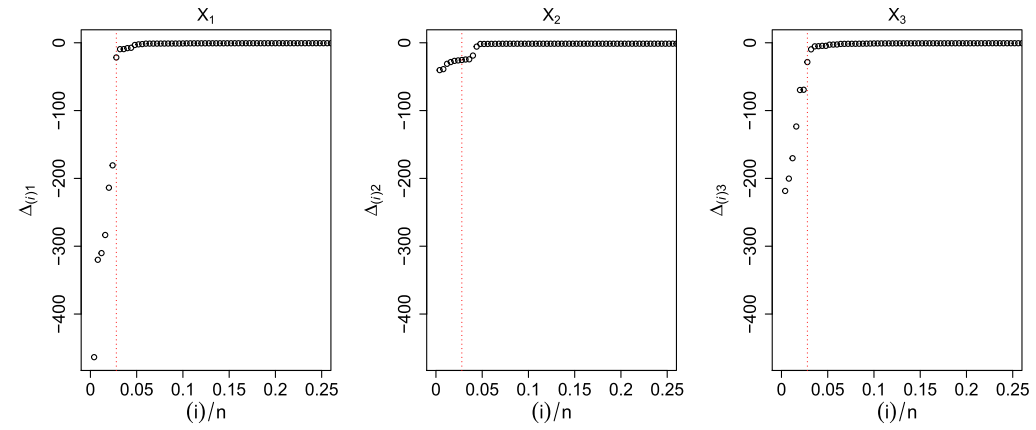
when $\alpha = 0.05$, with variability (gray area, calculated using the Median Absolute Deviation) close to zero, confirming the selection of this value as the appropriate flagging level for these data. Furthermore, we can verify the behavior of Δ_{ij} for each variable, recalling that they correspond to the difference between the individual contribution to the objective function when a cell is considered reliable or unreliable. In Fig. 6, we plot points $\{\big((i)/n, \Delta_{(i)j}\big)\}_{i=1}^{n}$ for the first three variables as examples, where units are ordered according to their Δ values (see the Supplementary Material for the plots of the remaining seven variables). When $\alpha = 0.025$, some units with a small value of Δ are not considered as contaminated, resulting in a lower flagging level than the theoretical contamination one (Fig. 6a). The same occurs for smaller values of α. On the other hand, too many cells than needed are flagged as unreliable when $\alpha = 0.075$, as shown in Fig. 6c, where the values of Δ level off before the considered α – the same happens for $\alpha = 0.10$. The appropriate choice turns out to be 0.05 (Fig. 6b), which corresponds to the true proportion of contaminated cells in the simulated sample. Indeed, at this level of flagged cells, 5% of the Δ values per variable are significantly lower than the majority of the others. In general, we recommend initially adopting a conservative choice of α, then gradually decreasing it while monitoring the behavior of the objective function curves. The additional tools provided for supporting the choice of α, such as those in Figs. 5 and 6, are particularly useful when the objective function curves are smoother.

5.2. Constant c for the eigenvalue-ratio constraint

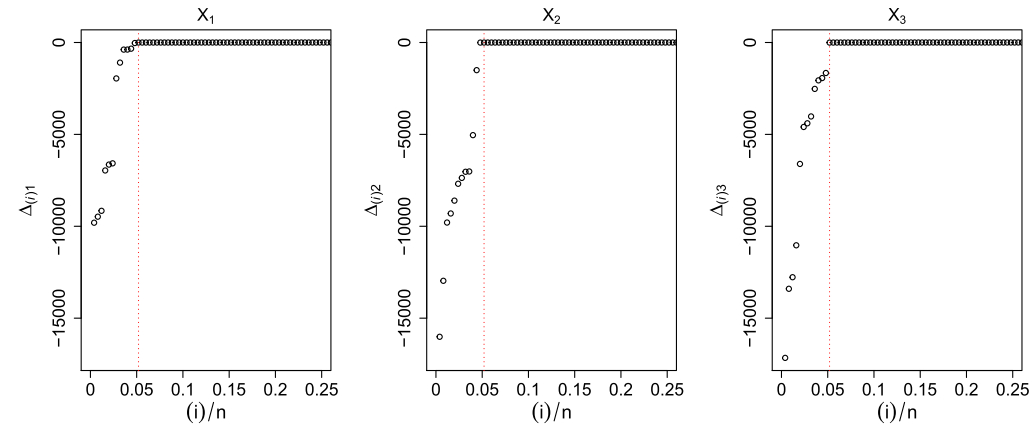
The constant for the eigenvalue-ratio constraint in (3) allows for different cluster shapes. When $c = 1$, the clusters become spherical, and cellFCLUST with $\alpha = 0$ and $\pi_1 = \ldots = \pi_k$ produces similar results to those of FKM. Indeed, a small value of c reduces the cellFCLUST ability to detect elongated and/or differently dispersed clusters, as illustrated in Fritz et al. [22] and García-Escudero and Mayo-Iscar [47]. However, in some applications, the user may be interested in more spherical types of clusters (see, for instance, Hennig and Liao [48]).

5.3. Fuzzifier parameter m, scale factor S and their interaction

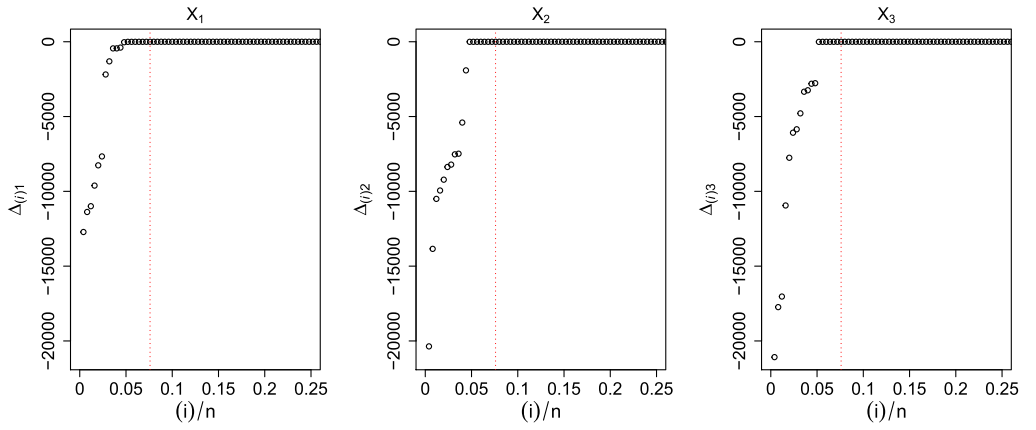
CellFCLUST shows different behavior depending on the fuzzifier parameter and the scaling of the data. The former affects the degree of fuzziness by letting the memberships become more similar as m increases. Points that are more fuzzily assigned to the clusters are usually units located farther from the core of the clusters in the p-dimensional space. On the other hand, the scale factor S influences the proportion of Hard Assignments (HA) obtained by cellFCLUST, which usually correspond to units within the core of the clusters carrying more weight in the estimation of their parameters. This effect was previously observed in Gath and Geva [34] and Rousseeuw et al. [33]. Given the property of high contrast, we thus treat the scale factor S as a key tuning parameter that allows us to manage the mentioned level of hard assignment depending on the purpose of the data analysis. Furthermore, the scale factor and the fuzzifier interplay, resulting in a change in the clustering structure estimated by cellFCLUST when both S and m vary. Specifically, as the scale factor decreases and the fuzzifier increases, the assignments tend to become less crisp and more fuzzy. However, a complete fuzzification is not desirable from an interpretation point of view. It is worth noting that when $m = 1$, the results are scale-independent since cellFCLUST returns a completely hard assignment. Since the effect of the tuning parameters m and S in cellFCLUST is similar to its casewise counterpart F-TCLUST, the reader can refer to Fritz et al. [22] for further insights and numerical examples.



(a) $\alpha = 0.025$



(b) $\alpha = 0.05$



(c) $\alpha = 0.075$

Fig. 6. Synthetic data: Δ plots where units are sorted according to their Δ values. Vertical, dashed, red line corresponds to the value of α used for the model implementation when $K = 2$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
List of variables of the body fat data set.

ID	Name	Measur. unit	ID	Name	Measur. unit
1	Adiposity index (BMI)	kg/m^2	7	Knee circumference (knee)	cm
2	Neck circumference (neck)	cm	8	Ankle circumference (ankle)	cm
3	Chest circumference (chest)	cm	9	Extended biceps circumference (bicep)	cm
4	Abdomen circumference (abdomen)	cm	10	Forearm circumference (forearm)	cm
5	Hip circumference (hip)	cm	11	Wrist circumference (wrist)	cm
6	Thigh circumference (thigh)	cm			

5.4. Restrictions on the weights π_k

One of the advantages of cellFCLUST is its ability to accommodate different cluster sizes via the weights π_k in the objective function $J_{\text{cellFCLUST}}$ (4), in a similar manner to F-TCLUST. It is worth noting that removing the weights π_k in (4) would favor clusters with comparable values of $\sum_{i=1}^{n} u_{ik}^m$, which are the quantities corresponding exactly to cluster sizes in a hard clustering approach ($m = 1$). We do not reproduce here examples illustrating the effect of constraining π_k, $k = 1, \dots, K$, to be equal, which the reader can find in Fritz et al. [22]. Moreover, Fritz et al. [22] highlight that when K is misspecified by the user with a value larger than necessary, some cluster weights π_k can become very close to 0, resulting in almost "empty" clusters, i.e., clusters with small u_{ik} for all units. This occurrence can reveal the potential misspecification of K to the user.

5.5. Guidelines for the parameter setting

The material presented above exemplify the role and effect of the cellFCLUST tuning parameters. All these parameters are closely related to each other, and their choice strongly depends on the application under analysis. If prior knowledge about the data set is available – for instance, regarding the expected clustering structure or the desired degree of fuzzification – this information can guide the choice of some (or, in rare cases, all) tuning parameters. When this does not occur, we suggest selecting appropriate values for the tuning parameters by taking into account the purpose of the analysis. Specifically, as a first step, a simple proposal is to choose the constant c by considering the order of magnitude of the ratio between the maximum and the minimum eigenvalues computed on the entire data set. Secondly, the scale factor can be set by analyzing the proportion of HA as S varies, selecting a reasonable subset of values for S consistent with the data under study. Indeed, although S and m are interconnected, the former mostly affects HA, while the latter affects the degree of fuzzification. In this regard, m can be chosen by studying the proportion of WA. The user may prefer to select a value of m that results in a proportion of WA within a specific range, depending on the goal of the analysis and the cluster configuration. Finally, the objective function curves presented in this section are a useful tool for selecting K and α, as previously described. The flagging level α can be deepened using the Δ plots. It is worth highlighting that the choice of S and m is connected to the selection of both the number of clusters and the level of flagged cells.

6. Real data analyses

In this section, we illustrate the potential of cellFCLUST through two real data applications. The first one focuses on identifying clusters of individuals with different levels of body fat-related risk (Section 6.1). In the second application, regions of the OECD countries are grouped based on eleven variables related to well-being (Section 6.2).

6.1. Body fat data

The data on body fat, available in the R package `UsingR`, contain physiological measurements of 250 men – units 172 and 182 have been discarded due to erroneous values. Among the 18 variables, we consider only those directly observed. Additionally, we exclude the variable *Age*, as it masks the clustering structure due to its cross effect on body phenotypes; the variables *Weight* and *Height*, as they define *BMI* (i.e., Body Mass Index); and the variable *Fat Free Weight*, since it is derived from estimated quantities not directly observed. Table 2 lists the eleven variables used for the analysis. Due to their different scales, the variables are standardized using a robust procedure that replaces the mean with the median and the standard deviation with the median absolute deviation. The presence of outlying values can be observed both by examining the boxplot and the pair plot, which are reported in the Supplementary Material. CellFCLUST is particularly suitable for analyzing this data set for two main reasons. First, since the variables are highly correlated, there is more information available to impute the potentially outlying values. Second, although the elongated shape of the data suggests the presence of a single cluster, this is not reasonable in this context as it would merge individuals with very different physical characteristics. This motivates considering a small value of c, similarly to the reasoning in Hennig and Liao [48] for the social stratification analysis. Assuming $c = 2$, we search for more spherical clusters that can sometimes be very close to each other, up to overlapping, making the fuzzy clustering approach extremely useful.

As a preliminary analysis, we consider the proportions of HA and WA obtained by varying the scale factor. Fixing $m = 1.7$, which proves to be a suitable value for the degree of fuzzification in this data set (see the Supplementary Material), we choose $S = 2$, as it

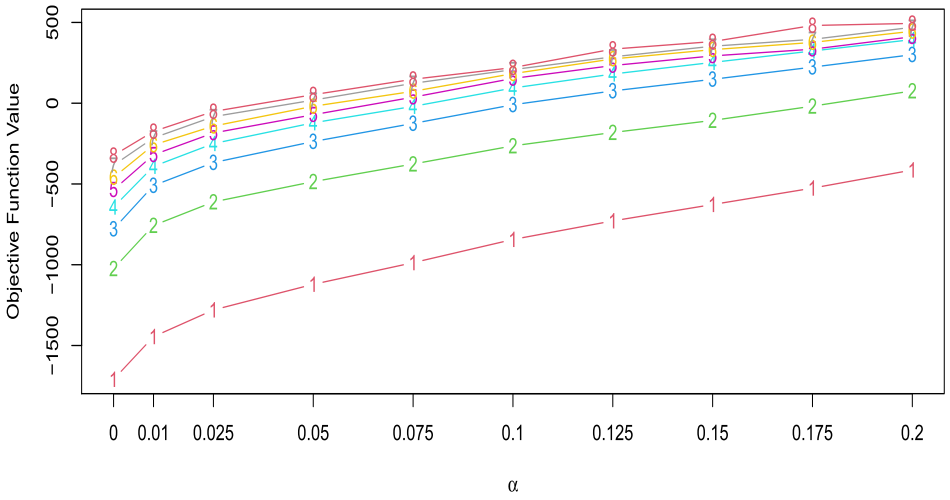


Fig. 7. Body fat data: objective function curves.

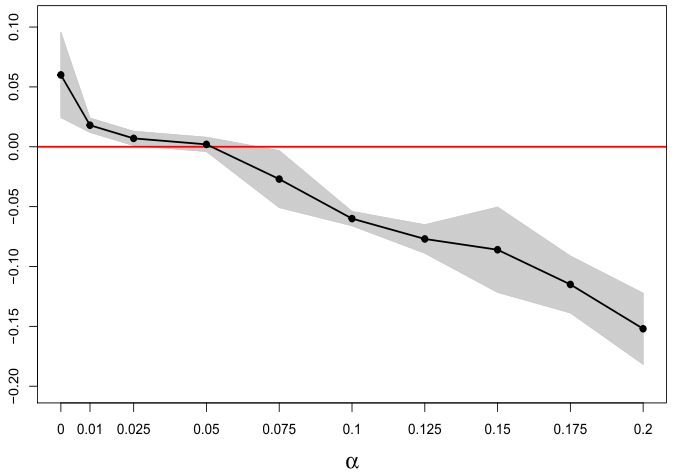


Fig. 8. Body fat data: difference between the knee point of Δ_{ij} and the α value. The plot shows the median differences across variables as α varies, with $K = 4$. The shaded gray area represents the variability.

yields a proportion of HA and WA of 0.43 and 0.32, respectively. This results to be the best choice since a small value of S, i.e. $S = 1$, excessively fuzzifies all units, while higher values of S correspond to an overly high proportion of hard assignments given the data configuration – for $S = 3$ and $S = 4$, the proportions of HA are 0.89 and 0.98, while those of WA are 0.04 and 0.01, respectively. After selecting the fuzzifier parameter, the scale factor, and the constant for the eigenvalue ratio, we determine the values of K and α based on the objective function curves reported in Fig. 7. Specifically, we vary $K \in \{1, \dots, 8\}$ and α from the set that increases by 0.025 increments from 0.025 to 0.20, including also 0 and 0.01. Looking at Fig. 7, we can choose $K = 4$, since it is the value from which the increase in the objective function does not justify an increase in the number of clusters. However, there is no clear indication of the optimal flagging level, which can be further investigated via the plots of the knee points. Following the same reasoning illustrated for the artificial data, we select $\alpha = 0.05$, as this is the level at which the median difference between the knee points of the Δ values and the corresponding flagging level is close to zero, and the variability is minimal (Fig. 8). This choice can be corroborated via the inspection of the Δ plots provided in the Supplementary Material.

The clustering results from cellFCLUST, using the tuning parameters set as previously detailed, are shown in Fig. 9. In the tetrahedral representation, points closer to the vertices have higher membership to the corresponding cluster, while those on the edges, or generally among vertices, are more fuzzified. The four clusters can be interpreted based on their configuration across variables (a graphical representation is provided in the Supplementary Material). *Cluster 1* (black) consists of individuals with the lowest measurements for all variables, including *BMI*, whose values, on the original scale, fall within the normal weight range, i.e. $BMI < 25$. *Cluster 2* (blue) includes men with generally higher measurements than Cluster 1 across the eleven variables, with *BMI* values still mostly within the normal range. For *Cluster 3* (red), some overlap with Cluster 2 is observed on specific variables (e.g., wrist), while *BMI* predominantly falls within the overweight range, i.e. $BMI \in [25, 30)$. Finally, *Cluster 4* (green) groups individuals in the severe

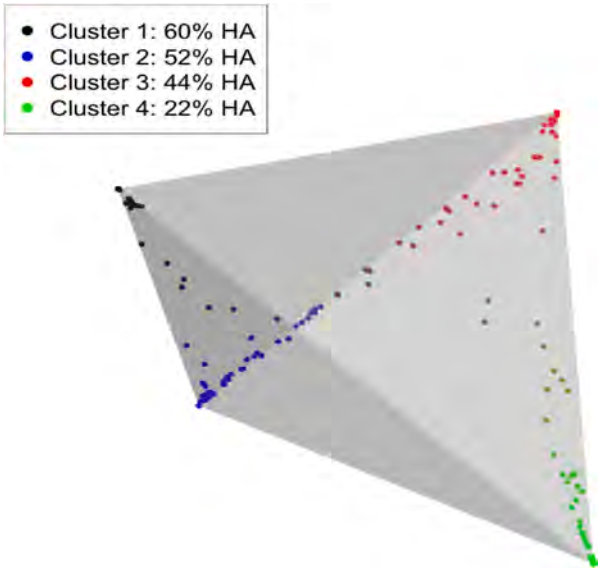


Fig. 9. Body fat data: tetrahedron plot showing cluster assignments (Cluster 1: black, Cluster 2: blue, Cluster 3: red, Cluster 4: green). Color intensity and point position reflect the degree of fuzziness (i.e., hard and soft memberships). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Proportion of outlying values detected by cellFCLUST in the body fat data set per variable and cluster. Total per row results in a fixed flagging level, which is 0.048 in this case since $n - h = 250 - \lceil 0.95 \times 250 \rceil = 250 - 238 = 12$ units.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
BMI	0.000	0.012	0.020	0.016
neck	0.012	0.020	0.008	0.008
chest	0.004	0.012	0.020	0.012
abdomen	0.000	0.016	0.016	0.016
hip	0.004	0.012	0.012	0.020
thigh	0.004	0.004	0.012	0.028
knee	0.004	0.012	0.016	0.016
ankle	0.012	0.020	0.004	0.012
bicep	0.024	0.012	0.000	0.012
forearm	0.012	0.012	0.000	0.024
wrist	0.020	0.012	0.004	0.012

overweight to obese range – the latter corresponds to $BMI \geq 30$ – with the highest values for all variables. As expected, Cluster 1 shows the highest proportion of HA (0.60 of the units in the cluster), and only 4 units with WA. The fuzziness of these units is directed toward Cluster 2, with membership degrees to the latter ranging from 0.19 to 0.37. Cluster 2, which has a proportion of HA of 0.52, contains 31 units with maximum membership below 0.9; among them, 24 have their second-highest membership in Cluster 3, and 7 in Cluster 1. Cluster 3 includes 23 weak assignments, with 13 represented by men whose second-highest membership lies in Cluster 2. The remaining 10 units would be secondarily assigned to Cluster 4, which, in turn, contains 17 weakly assigned units whose second-highest membership is in Cluster 3. These individuals may represent a subgroup of men for whom further analyses could be conducted – for instance, to assess whether they should follow specific dietary plans designed for individuals with obesity.

Finally, we analyze the distribution of flagged cells per variable across the four clusters (Table 3). Specifically, it is worth noting that *BMI* shows no unreliable cells in the first cluster, which corresponds to normal weight, while the cluster with the highest proportion of flagged cells – considering a fixed total per variable – is the third one. Similarly, Cluster 1 includes all reliable units for the variable *abdomen*, while the other clusters share the same proportion of unreliable cells. Higher flagging levels are usually observed in Cluster 4, i.e., the group containing men with obesity. This cluster can also encompasses extremely obese individuals, whom the model is able to discover and assign to the correct cluster. It is important to highlight that the detection of outlying values by cellFCLUST does not occur marginally, but takes into account all reliable cells for each unit.

Table 4
Number of regions by OECD country.

Country	Regions	Country	Regions	Country	Regions
Australia	8	Greece	13	New Zealand	14
Austria	9	Hungary	8	Norway	6
Belgium	3	Iceland	2	Poland	17
Canada	13	Ireland	3	Portugal	7
Chile	16	Israel	6	Slovak Republic	4
Colombia	33	Italy	21	Slovenia	2
Costa Rica	6	Japan	10	Spain	17
Czech Republic	8	Korea	7	Sweden	8
Denmark	5	Latvia	6	Switzerland	7
Estonia	5	Lithuania	10	Türkiye	26
Finland	5	Luxembourg	1	United Kingdom	12
France	18	Mexico	32	United States	51
Germany	16	Netherlands	12		

6.2. Well-being: OECD regional data

The second empirical analysis focuses on well-being indicators published by the Organization for Economic Co-operation and Development (OECD) on a regional basis (https://www.oecdregionalwellbeing.org/). The dataset comprises 447 statistical regions from 38 OECD countries (Table 4) and 11 indicators, each ranging in $[0, 10]$. These are *Education, Jobs, Income, Safety, Health, Environment, Civic Engagement, Accessibility to services, Housing, Community, Life satisfaction*. Although the scores have already been winsorized to mitigate the effect of marginal outlying values, the presence of non-marginal outliers justifies the need for a cellwise robust approach. Missing information is present in 1.6% of the cells. For this data set a natural cluster structure can be expected, as dissimilarities between regions of different countries, or even continents, could be significant. Finally, the fuzzy approach can uncover hidden links between apparently distant regions.

Unlike the previous application, we choose a higher value of c, allowing for elongated clusters. This choice is motivated by the reasonable assumption of correlation between variables (e.g., *Education* and *Income*) within the same cluster, as well as different degrees of variability between clusters. Specifically, we impose $c = 50$, which corresponds to the order of magnitude of the ratio between the eigenvalues of the covariance matrix computed on the entire data set. Since fuzzy models are scale-dependent and a constraint on the eigenvalues is imposed, careful consideration should be given to the preprocessing of the variables. Given the nature of the indicators, we consider a selected number of scaling factors $S = \{1, 2, 5, 10\}$, which allow to adjust the proportion of hard assignments while maintaining high comprehensibility of the scaled scores. We select $S = 5$ as the proportion of HA approaches 0.74 – it ranges from 0.14 to 0.91 depending on the choices of K and α, as shown in Fig. 10c. The fuzzifier parameter m is set to 1.8 as it provides desirable levels of WA, with 29 weakly assigned regions out of 447 (18 for $m = 1.6$ and 55 for $m = 2$, as shown in the Supplementary Material). Once c, S, and m are defined, the objective function curves can be computed. As reported in the Supplementary Material, these curves suggest $K = 5$ as a sensible choice. To support this choice, we include the plot of the knee points in Fig. 11.

As we can see in Fig. 12, clusters have a clear spatial characterization, which can be summarized as follows: Northern Europe and Oceania (Cluster 1), Eastern Europe (Cluster 2), United States of America (Cluster 3), Latin America and the Aegean Sea (Cluster 4), and Southern Europe, Asia and Chile (Cluster 5). Regions within the same country are often clustered together, with limited exceptions. Examples are Canadian regions, which are divided between United States of America and Northern Europe and Oceania, mainly driven by differences in *Income* and *Health*. Korean regions are also split between Eastern Europe and Southern Europe, Asia and Chile, as differences in *Civic Engagement* and *Life Satisfaction* discriminate the assignment. Another interesting case is that of France, where metropolitan regions are assigned to Northern Europe and Oceania, except for Pays de la Loire and Corsica, which are assigned to Southern Europe, Asia and Chile. However, overseas French regions such as Guadeloupe and French Guiana – which are assigned to Latin America and the Aegean Sea – do not fall into the same cluster as the mainland France. This may be due to their generally lower values of the indicators compared to those of Northern Europe and Oceania. Moreover, relevant fuzzy assignments can be found in many Greek regions, such as North Aegean, Peloponnese, South Aegean, Western Macedonia, Epirus, Ionian Islands, which belong to both Southern Europe, Asia and Chile, and Latin America and the Aegean Sea, with varying degrees. This is due to similarities within the cluster of Southern Europe, Asia and Chile offset by lower scores in *Jobs*, *Civic Engagement*, *Community* and *Life Satisfaction*, which are closer to those of Latin America and the Aegean Sea. Other fuzzy regions to highlight are the Colombian San Andrés, Amazonas, Guainía, and Guaviare which, although assigned to Eastern Europe, have a high membership value for Latin America and the Aegean Sea, and the Costa Rica Central, Central Pacific, Brunca and Huetar Caribbean, partly assigned to Southern Europe, Asia and Chile, due to similarities with the latter. A complete overview of the weakly assigned regions and their fuzzification is provided in the Supplementary Material.

In terms of unreliable cells, we see in Fig. 13 that California is considered outlying with respect to *Income*, as the difference between salaries in this region and those in Northern Europe and Oceania – the cluster which California belongs – is significant. Another example is Lombardy, which has a much lower score for *Environment*, being a region with higher levels of industrialization,

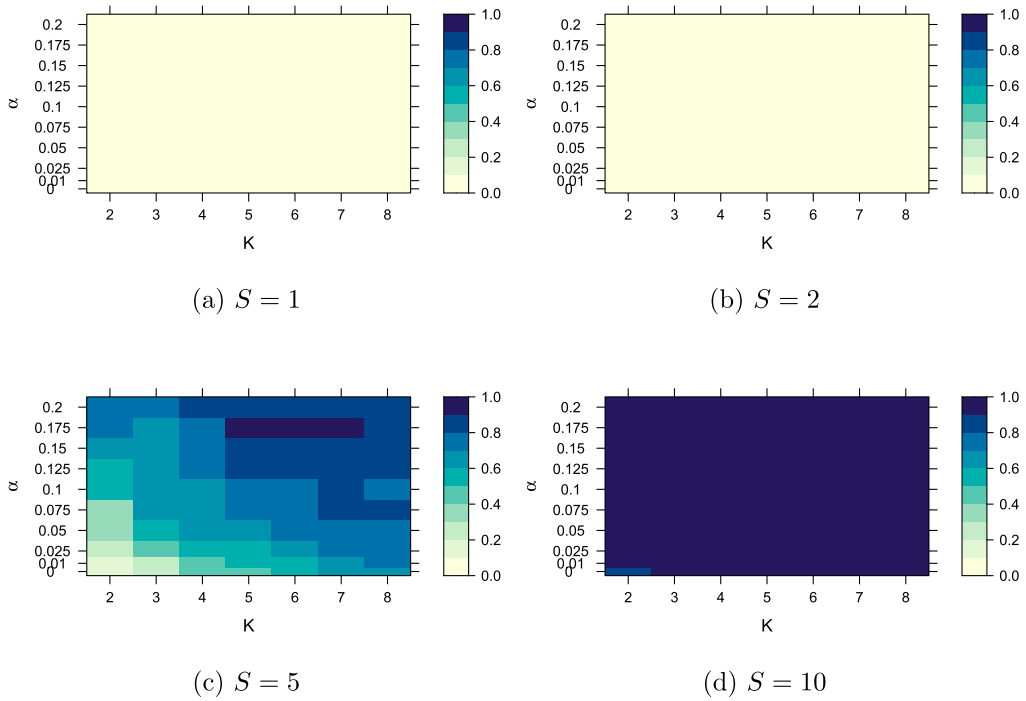


Fig. 10. OECD data: proportion of hard assignments depending on K and α for four different levels of S.

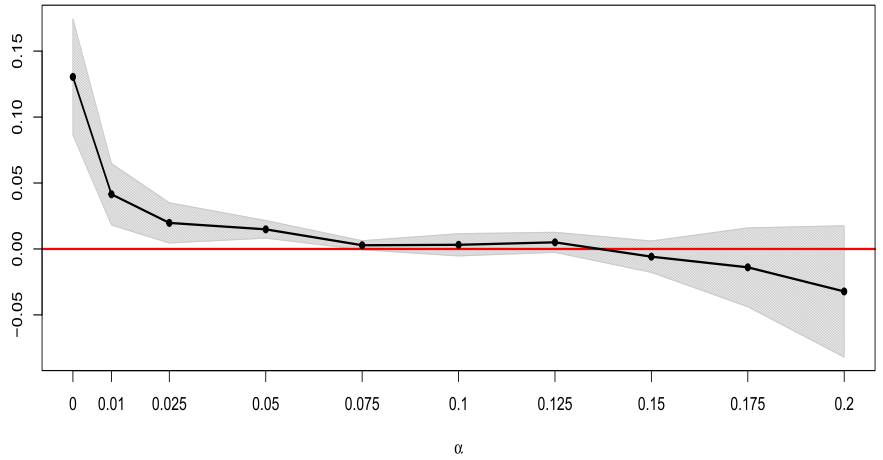


Fig. 11. OECD data: difference between the knee point of Δ_{ij} and the α value. The plot shows the median differences across variables as α varies, with $K = 5$. The shaded gray area represents the variability.

compared to other Italian regions or other regions of Southern European countries, causing significant pollution. Silesia represents another outlier in *Environment*, due to the presence of a substantial coal mining industry in the region. With respect to *Housing*, the Stockholm region presents an outlying value, as it hosts one of the most populated cities in all of Northern Europe and Oceania. New York is outlying both in *Income*, due to notably high salaries, and *Civic Engagement*, representing an example of a non-marginal outlier with lower and higher values than expected, respectively, given the other indicators. Jalisco, as other regions in its cluster, has a low score of *Safety*, caused by increased exposure to criminal activities compared to other regions. It is also possible to observe countries where all the cells for a specific variable have been flagged as outlying. Particularly, Swiss regions are all considered outlying in *Income*, which is unusually high compared to their peers, and *Civic Engagement*, as voter turnout in the country is significantly lower than in most European nations. The same occurs for South Korea in *Environment*, as the levels of pollution in the peninsula are consistently high. This result from cellFCLUST is interesting, as it helps uncover differences within clusters, where subgroups (e.g., all, or almost all, regions of a country) may differ from other units only in a small subset of variables. Depending on the purpose of the

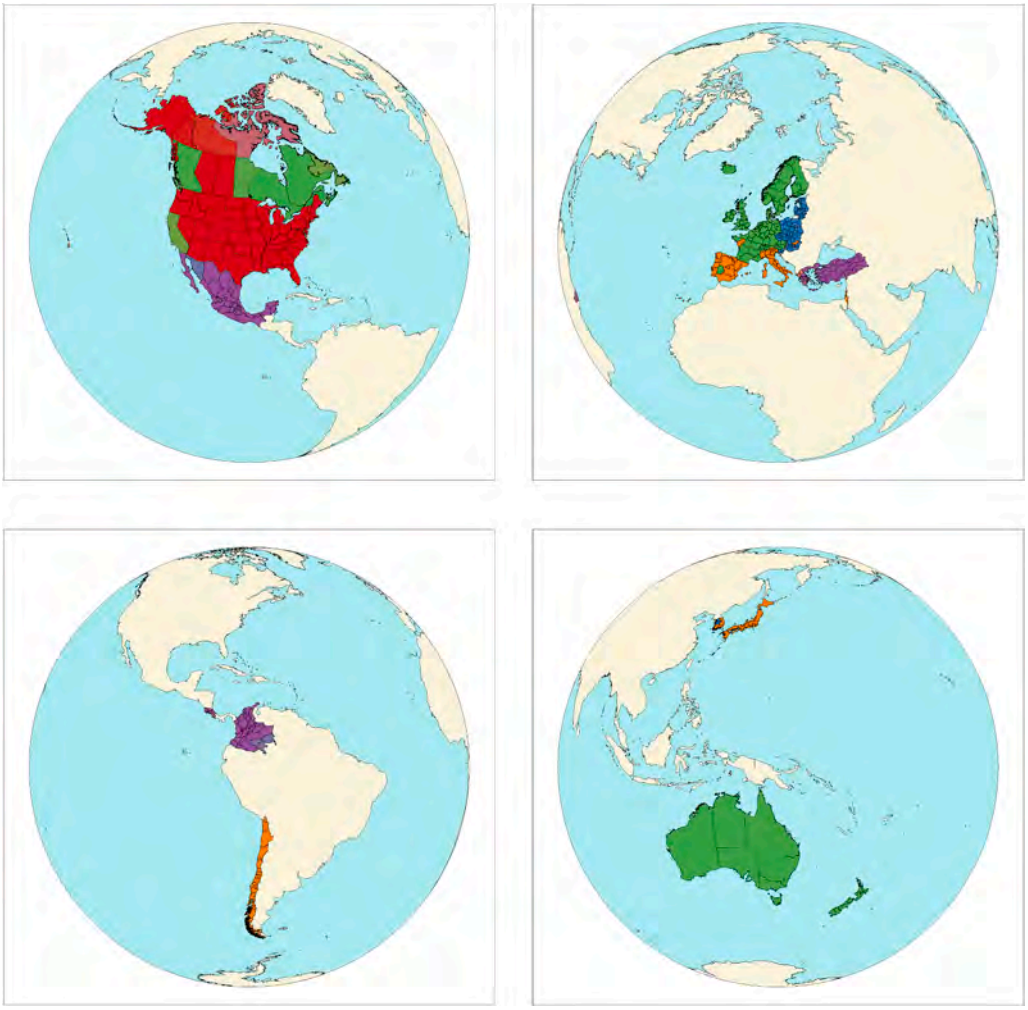
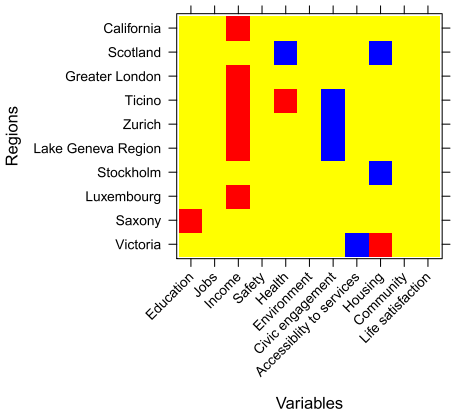
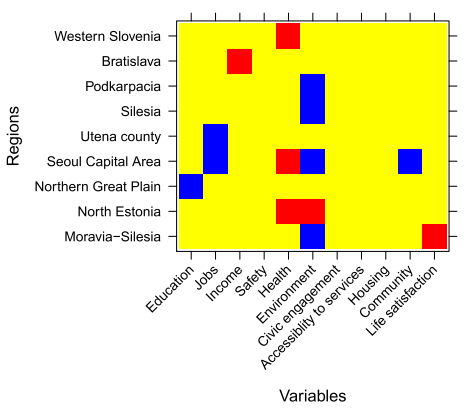


Fig. 12. OECD data: clustering results. The color gradient indicates fuzzy assignments. Cluster 1: – Cluster 2: – Cluster 3: – Cluster 4: – Cluster 5: .

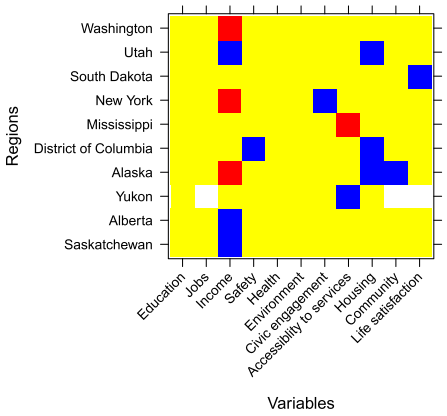
analysis, the user may increase the number of clusters and reduce the flagging level by merging these regions into additional clusters and avoiding the flagging of their values, if these are of interest. However, this comes at the cost of reduced model parsimony, which may not always be desirable. A complete overview of the flagged cells in each cluster is available in the Supplementary Material.



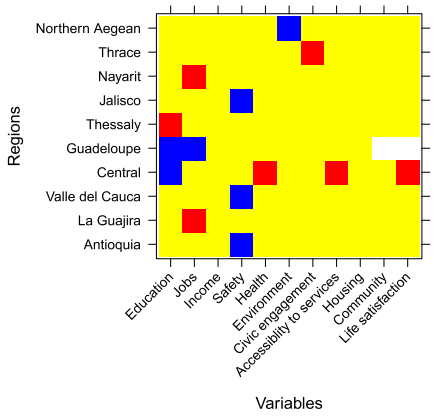
(a) Northern Europe and Oceania



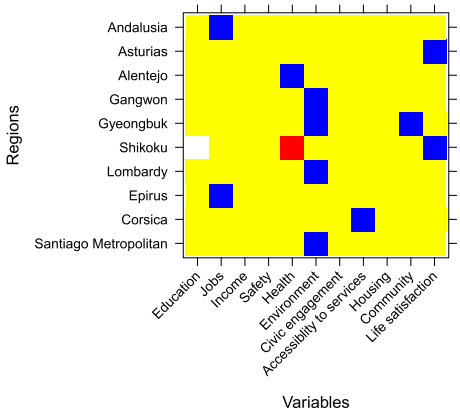
(b) Eastern Europe



(c) United States of America



(d) Latin America and the Aegean Sea



(a) Southern Europe, Asia and Chile

Fig. 13. OECD data: outlying cells of selected regions (yellow: reliable cells; blue: flagged cells imputed with a higher value than the original one; red: flagged cells imputed with a lower value than the original one; white: missing values). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7. Conclusions

A new methodology, called cellFCLUST, has been introduced for fuzzy clustering and cellwise outlier detection. Following the recent paradigm of cellwise contamination, which considers single cells of a data matrix to be potentially contaminated, we have developed an algorithm intended for flagging these cellwise outliers, which are then treated as missing values, and imputed rather than discarded, before estimating the model parameters. If not properly identified, outlying values and fuzzy assignments can interplay in ways that change the clustering structure and bias the parameter estimates. Notably, cellFCLUST encompasses both hard and soft robust clustering methodologies: crisp assignment decisions are made for cases located in the central regions or cores of the clusters, while fuzzy assignments are retained for ambiguous units.

Through a simulation study and two real-world applications, we have shown the effectiveness and usefulness of the proposed model. The former illustrate the performance of cellFCLUST in cluster recovery, parameter estimation and outlier detection compared to other fuzzy clustering methodologies. The results demonstrate the advantages of the proposal when cellwise contamination occurs. Regarding the real data analyses, the first one focuses on analyzing risk levels among individuals in the overweight to obese range, leveraging fuzzification to identify units whose measurements warrant closer inspection; the second application concerns indicators used to study well-being across regions of the OECD countries. In this framework, cellFCLUST enables the identification of common patterns among countries in terms of well-being, serving two main purposes: on one hand, to reveal different behaviors among regions within the same country – for example, California is grouped with Northern European countries based on its measurement; and on the other hand, to highlight indicators on which some countries (and their regions) display anomalous values relative to the cluster they belong to. This is the case, for instance, of Switzerland, whose income is higher even than that of Northern European countries.

A crucial role in the implementation of cellFCLUST is played by its tuning parameters, as the clustering and outlier detection results can be sensitive to their choice. We have illustrated their effects on artificial data, providing guidance to users for their selection. It is worth noting that all these parameters – namely, the number of clusters K, the flagging level α, the constant c for the eigenvalue-ratio constraint, the fuzzifier parameter m, and the scale factor S – are interconnected and require interrelated analyses for their appropriate setting. The dependence of cellFCLUST on these parameters offers users a flexible methodology that accommodates on the purpose of the analysis and prevents issues in the parameter estimation. Simplifying these connected tasks remains an open issue, leaving room for improvement and future developments aimed at identifying a procedure for the simultaneous selection of all tuning parameters in cellFCLUST. Future work could also explore extending the robust fuzzy clustering approach with a factorial structure for the cluster covariance matrices [49] to the cellwise contamination framework.

Code availability

The R code for the implementation of cellFCLUST and the simulation study in Section 4 is available at https://github.com/giorgiazaccaria/cellFCLUST.

CRediT authorship contribution statement

Giorgia Zaccaria: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization; **Lorenzo Benzakour:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation; **Luis A. García-Escudero:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization; **Francesca Greselin:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization; **Agustín Mayo-Íscar:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization.

Data availability

The real data used in Section 6 are publicly available, as detailed in the corresponding subsections.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The research of Giorgia Zaccaria and Francesca Greselin was supported by Milano-Bicocca University Fund for Scientific Research, 2023-ATE-0448. Francesca Greselin's research was also supported by PRIN2022 - 2022LANNKC. The research of Luis A. García-Escudero and Agustín Mayo-Íscar was partially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, grant PID2024-162240NB-I00, and Junta Castilla y León grant VA064G24 .

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.ijar.2026.109698

Appendix A. Notation

For the convenience of the reader, the notation used in this paper is listed in Table A.1.

Table A.1
Notation of the paper.

Notation	Description
n, p, K	Number of units, variables, clusters, respectively (scalars).
m	Fuzzifier tuning parameter (scalar).
α	Proportion of cells per variable flagged as contaminated in a data matrix, referred to as *flagging level* (scalar).
$\mathbf{X}$	$(n \times p)$ data matrix, where $\mathbf{x}_i$ denotes its i-th row (unit) containing p variable measurements, and x_{ij} represents the cell corresponding to the j-th measurement for the i-th unit.
$\mathbf{W}$	$(n \times p)$ cellwise indicator matrix, where zeros denote contaminated or missing cells.
$\mathbf{U}$	$(n \times K)$ membership matrix of units to cluster.
π_k	Weight of cluster k, $k = 1, \dots, K$ (scalar).
$\boldsymbol{\mu}_k$	p-dimensional vector of variable means for cluster k, $k = 1, \dots, K$.
$\boldsymbol{\mu}_{k[j]}, \boldsymbol{\mu}_{k[\mathbf{w}_i]}$	Sub-vectors of $\boldsymbol{\mu}_k$ corresponding to the j-th variable and the variables for which $w_{ij} = 1, j \in \{1, \dots, p\}$, respectively.
$\boldsymbol{\Sigma}_k$	$(p \times p)$ covariance matrix for cluster k, $k = 1, \dots, K$.
$\boldsymbol{\Sigma}_{k[j,j]}$	(j, j)-element of $\boldsymbol{\Sigma}_k$ (scalar), corresponding to the variance of the j-th variable.
$\boldsymbol{\Sigma}_{k[j,\mathbf{w}_i]}, \boldsymbol{\Sigma}_{k[\mathbf{w}_i,\mathbf{w}_i]}$	Row vector and sub-matrix of $\boldsymbol{\Sigma}_k$, respectively, corresponding to the covariances between the j-th variable and the variables for which $w_{ij} = 1$, and to the covariances between the latter.
c	Constant for the eigenvalue-ratio constraint on covariance matrices (scalar).

References

[1] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: Statistics, University of California Press, Berkeley, Calif., 1967, pp. 281–297.
[2] G.H. Ball, D.J. Hall, A clustering technique for summarizing multivariate data, Syst. Res. 12 (1967) 153–155.
[3] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[4] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, John Wiley & Sons, 1999.
[5] J.V. De Oliveira, W. Pedrycz, Advances in Fuzzy Clustering and its Applications, John Wiley & Sons, 2007.
[6] P. Giordani, M. Ferraro, F. Martella, Introduction to clustering, in: An Introduction to Clustering with R, Springer, 2020, pp. 3–5.
[7] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[8] P.J. Rousseeuw, E. Trauwaert, L. Kaufman, Fuzzy clustering with high contrast, J. Comput. Appl. Math. 64 (1) (1995) 81–90.
[9] P. Huber, Robust Statistics, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, New York, 1981.
[10] R. Maronna, R. Martin, V. Yohai, M. Salibián-Barrera, Robust Statistics: Theory and Methods (with R), John Wiley & Sons, 2019.
[11] R. Dave, R. Krishnapuram, Robust clustering methods: a unified view, IEEE Trans. Fuzzy Syst. 5 (2) (2002) 270–293.
[12] L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Íscar, A review of robust clustering methods, Adv. Data Anal. Classif. 4 (2) (2010) 89–109.
[13] A. Banerjee, R. Davé, Robust clustering, Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov. 2 (1) (2012) 29–59.
[14] L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Íscar, C. Hennig, Robustness and outliers, in: C. Hennig, M. Meilà, F. Murtagh, R. Rocci (Eds.), Handbook of Cluster Analysis, Chapman & Hall/CRC, Boca Raton, FL, 2015, pp. 653–678.
[15] F. Alqallaf, S. Van Aelst, V.J. Yohai, R.H. Zamar, Propagation of outliers in multivariate data, Ann. Stat. 37 (1) (2009) 311–331.
[16] J. Raymaekers, P.J. Rousseeuw, Challenges of cellwise outliers, Economet. Stat. 38 (2026) 6-25.
[17] J. Raymaekers, P.J. Rousseeuw, The cellwise minimum covariance determinant estimator, J. Am. Stat. Assoc. 119 (548) (2023) 2610-2621.
[18] P.J. Rousseeuw, Least median of squares regression, J. Am. Stat. Assoc. 79 (388) (1984) 871–880.
[19] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, Dordrecht, Reidel, 1985, pp. 283–297.
[20] G. Zaccaria, L. García-Escudero, F. Greselin, A. Mayo-Íscar, Cellwise outlier detection in heterogeneous populations, Technometrics 67 (4) (2025) 643-654.
[21] P. Puchhammer, I. Wilms, P. Filzmoser, A smooth multi-group gaussian mixture model for cellwise robust covariance estimation, arXiv:2504.02547 (2025).
[22] H. Fritz, L.A. García-Escudero, A. Mayo-Íscar, Robust constrained fuzzy clustering, Inf. Sci. 245 (2013) 38–52.
[23] J. Kim, R. Krishnapuram, R. Davé, Application of the least trimmed squares technique to prototype-based clustering, Pattern Recognit. Lett. 17 (6) (1996) 633–641.
[24] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybernet. 3 (3) (1973) 32-57.
[25] R. Dave, Characterization and detection of noise in clustering, Pattern Recognit. 12 (11) (1991) 657–664.
[26] R. Krishnapuram, J. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (2) (2002) 98–110.
[27] S. Chatzis, T. Varvarigou, Robust fuzzy clustering using mixtures of student's-t distributions, Pattern Recognit. Lett. 29 (13) (2008) 1901–1905.
[28] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc., Ser. B (Statist. Methodol.) 39 (1) (1977) 1–38.
[29] R.J. Hathaway, J.C. Bezdek, Fuzzy c-means clustering of incomplete data, IEEE Transact. Syst. Man Cybernet. Part B (Cybernet.) 31 (5) (2001) 735–744.
[30] Jyoti, J. Singh, A. Gosain, Handling missing values using fuzzy clustering: a review, in: A. Bhattacharya, S. Dutta, P. Dutta, V. Piuri (Eds.), Innovations in Data Analytics. ICIDA 2022. Advances in Intelligent Systems and Computing, 1442, Springer, Singapore, 2023, p. 341-353.

[31] D.E. Gustafson, W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: 1978 IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes, San Diego, CA, USA, 1979, p. 761-766.
[32] E. Trauwaert, L. Kaufman, P. Rousseeuw, Fuzzy clustering algorithms based on the maximum likelihood priciple, Fuzzy Sets Syst. 42 (2) (1991) 213–227.
[33] P.J. Rousseeuw, E. Trauwaert, L. Kaufman, Fuzzy clustering using scatter matrices, Comput. Stat. Data Anal. 23 (1) (1996) 135–151.
[34] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 773–780.
[35] L.A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, A. Mayo-Íscar, Eigenvalues and constraints in mixture modeling: geometric and computational issues, Adv. Data Anal. Classif. 12 (2018) 203-233.
[36] L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Íscar, A general trimming approach to robust cluster analysis, Ann. Stat. 36 (3) (2008) 1324–1345.
[37] L.A. García-Escudero, A. Gordaliza, A. Mayo-Iscar, A constrained robust proposal for mixture modeling avoiding spurious solutions, Adv. Data Anal. Classif. 8 (1) (2014) 27–43.
[38] C. Croux, G. Haesbroeck, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, J. Multivar. Anal. 71 (2) (1999) 161–190.
[39] Z. Ghahramani, M. Jordan, Learning from incomplete data, Technical Report AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, 1995.
[40] H. Fritz, L.A. García-Escudero, A. Mayo-Íscar, A fast algorithm for robust constrained clustering, Comput. Stat. Data Anal. 61 (2013) 124–136.
[41] H. Fritz, L.A. García-Escudero, A. Mayo-Íscar, tclust: an R package for a trimming approach to cluster analysis, J. Stat. Softw. 47 (12) (2012) 1–26.
[42] R. Babuska, P.J. van der Veen, U. Kaymak, Improved covariance estimation for Gustafson-Kessel clustering, in: 2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No.02CH37291), Honolulu, HI, USA, 2, 2002, pp. 1081–1085.
[43] M.B. Ferraro, P. Giordani, A. Serafini, fclust: an R package for fuzzy clustering, R J. 11 (1) (2019) 198–210.
[44] R. Maitra, V. Melnykov, Simulating data to study performance of finite mixture modeling and clustering algorithms, J. Comput. Graph. Stat. 19 (2) (2010) 354-376.
[45] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[46] L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Íscar, Exploring the number of groups in robust model-based clustering, Stat. Comput. 21 (4) (2011) 585-599.
[47] L.A. García-Escudero, A. Mayo-Iscar, Robust clustering based on trimming, Wiley Interdiscip. Rev. Comput. Stat. 16 (4) (2024) e1658.
[48] C. Hennig, T. Liao, How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, J. R. Stat. Soc., C: Appl. Stat. 62 (3) (2013) 309–369.
[49] L.A. García-Escudero, F. Greselin, A. Mayo-Iscar, Robust fuzzy and parsimonious clustering based on mixtures of factor analyzers, Int. J. Approx. Reason. 94 (2018) 60–75.