Graphical and computational tools to guide parameter choice for the cluster weighted robust model

Andrea Cappozzo
MOX, Department of Mathematics
Politecnico di Milano
Luis Angel García-Escudero
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid
Francesca Greselin
Department of Statistics and Quantitative Methods
University of Milano-Bicocca
and Agustín Mayo-Iscar
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid

November 21, 2022

Abstract

The Cluster Weighted Robust Model (CWRM) is a recently introduced methodology to robustly estimate mixtures of regressions with random covariates. The CWRM allows users to flexibly perform regression clustering, safeguarding it against data contamination and spurious solutions. Nonetheless, the resulting solution depends on the chosen number of components in the mixture, the percentage of impartial trimming, the degree of heteroscedasticity of the errors around the regression lines and of the clusters in the explanatory variables. Therefore an appropriate model selection is crucially required. Such a complex modeling task may generate several "legitimate" solutions: each one derived from a distinct hyper-parameters specification. The present paper introduces a two step-monitoring procedure to help users effectively explore such a vast model space. The first phase uncovers the most appropriate percentages of trimming, whilst the second phase explores the whole set of solutions, conditioning on the outcome derived from the previous step. The final output singles out a set of "top" solutions, whose optimality, stability and validity is assessed. Novel graphical and computational tools - specifically tailored for the CWRM framework - will help the user make an educated choice among the optimal solutions. Three examples on real datasets showcase our proposal in action. Supplementary files for this article are available online.

Keywords: Cluster-weighted modeling; Outliers; Trimmed BIC; Eigenvalue constraint; Monitoring; Model-based clustering; Robust estimation

1 Introduction

Clustering has been defined as one of the core tasks in data mining (Bezdek, 2013). In the last decades, different philosophical points of view have generated different definitions about what constitutes clustering. In a top-down view, clustering means partitioning a heterogeneous population into a number of more homogeneous subgroups. Conversely, in a bottom-up view, a criterion of similarity can be employed to find groups in a dataset. Henceforth, a plethora of models, algorithms and criteria arose in the literature for the purpose of grouping. In essence, as the notion of cluster cannot be uniquely defined, clustering is an “ill-posed” problem. It has been written that “clustering is in the eye of the beholder” (von Luxburg et al., 2012; Estivill-Castro, 2002), and as such, the most appropriate clustering method depends on the knowledge of the field of application and on the (subjective) aim of the task (Hennig, 2015). Finally, evidence from many examples suggests that a given process can be fruitfully modeled in several ways.

One of the trickiest choices in cluster analysis lies in identifying the number of clusters G. In some cases, G is known in advance, being part of the context-specific information. Nevertheless, more often than not, G has to be inferred, assuming that the data carry information about the process generating them. Among the many contributions to this literature stream, we refer to Milligan and Cooper (1985); Rousseeuw (1987a); Tibshirani et al. (2001); Cerioli et al. (2018) and references therein. In model-based clustering, it is customary to select G based on the optimization of a penalized likelihood function. This is an effective criterion, balancing the trade-off between the goodness of fit of the model and the simplicity of the model itself. Maximum likelihood procedures, with their underlying elegant theory, critically depend on the knowledge of the exact parent distributions and the proper fulfillment of those model-based assumptions, and hence they clearly lack robustness.

When contamination appears in the data, adaptive procedures that use most of the sample information are preferable, as they are not hampered by outlying units at the price of losing some efficiency. In what follows, with contamination we identify any mechanism that obscures the relationship between covariates, response and cluster membership.

Such considerations hold true in the general mixture modeling framework, and particularly for mixtures of regressions with random covariates, also known as Cluster Weighted Models (CWMs), which will be the focus of the present paper. Firstly introduced by Gershenfeld (1997) as a machine learning technique for prediction of non-linear time series, their reformulation in a statistical setting is provided in Ingrassia et al. (2012), where the authors showcase that CWMs represent a very general family of mixture models, including finite mixtures of distributions and finite mixtures of regressions as special cases. In details, cluster weighted modeling defines a mixture approach for flexibly learning the joint probability of a response variable and a set of explanatory variables; a thorough survey on the topic can be found in Dang et al. (2017).

Coming back to the data contamination issue, García-Escudero et al. (2017) introduce a methodology for robustly fitting mixtures of regression with random covariates, named Cluster Weighted Robust Model. To reach robustness, two steps are included within the parameters estimation procedure. Firstly, the less plausible observations under the currently estimated model are discarded, to eliminate their problematic contribution to inference. This is achieved using impartial trimming, for which a fixed proportion of observations, hereafter denoted with α, is left unassigned. Secondly, the estimation of the covariance matrices for the covariates, as well as of the variances of the regression error terms, are performed under two user-defined constraints to sweep aside degeneracies and uninteresting spurious solutions. Therefore, the resulting robust estimation depends on three hyper-parameters: the percentage of trimmed observations α and the two constraints for the covariance matrices and the regression error variances, respectively. Even if there are situations in which the user is able to indicate how to set these hyper-parameters; in general their specification is not straightforward and their automatic selection is still an open issue.

Inspired from previous works tackling this challenge for the case of mixtures of Gaussians (Riani et al., 2019; Cerioli et al., 2018), the present paper introduces a set of graphical and computational tools to guide the final user in making an informed choice for the hyper-parameters for the more complicated case of the Cluster Weighted Robust Model. We also provide a fully automated procedure to yield a small and ranked list of optimal solutions, featured by their properties of stability and validity.

The structure of the paper is as follows. In Section 2 an outline of the CWM is recalled. Afterwards, the robust version of the model is presented and a short discussion on the role of the hyper-parameters is given. In addition, a specific penalized criterion for model selection is proposed. In Section 3, a two-step monitoring procedure is presented to efficiently explore the space of solutions. In the first step, a few sensible values of the trimming level α are suggested, monitoring some crucial metrics for the CWM. In the second step, varying the constraints for the covariates and the regression errors, and the number of groups, an algorithm for finding a short list of optimal solutions is introduced. Stability and validity of each solution are also provided through a new ad-hoc graphical tool. Real applications on three datasets are presented in Section 4, and results are discussed. Finally, Section 5 concludes the manuscript and present some directions for future research.

2 The Cluster Weighted Robust Model

Let $\mathbf{X}$ be a vector of covariates with values in $\mathbb{R}^d$, and let Y be a dependent (or response) variable with values in $\mathbb{R}$. Assume that the regression of Y on $\mathbf{X}$ varies across G levels, say groups or clusters, of a categorical latent variable Z. In other words, potential relationships between the variables in $\mathbf{X}$ convey information on group membership, and this in turn may modify the regression in Y, as represented by the directed graph in Figure 1. The aim is thus to identify groups in the data based simultaneously on the conditional distribution of $Y|\mathbf{X}$ and on the marginal distribution of $\mathbf{X}$. Such situations are common, for instance, in demography, psychology and marketing, where groups of people are identified on the basis of core variables that are costly to obtain (e.g. test-item responses, indicating behavior or propensity). The groups need to be simultaneously profiled with concomitant variables that are cheaper to collect or widely available, such as demographic data. Once the groups are identified, new subjects are classified using mainly demographic data. In details, the

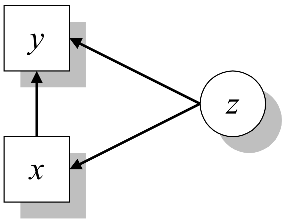

Figure 1: Directed graph showing the relationships among the latent (z) and the manifest variables (x, y) under the Cluster Weighted Model.

CWM models the joint distribution of $(\mathbf{X}, Y)$ as:

$$p(\mathbf{x}, y; \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g f(y|\mathbf{x}; \boldsymbol{\theta}_g) p(\mathbf{x}; \boldsymbol{\xi}_g), \tag{1}$$

where $\pi_g = P(Z = g)$ are the positive mixing weights summing up to 1, $f(y|\mathbf{x}; \boldsymbol{\theta}_g)$ is the conditional density of $Y|\mathbf{X}, Z = g$, depending on the parameter $\boldsymbol{\theta}_g$, $p(\mathbf{x}; \boldsymbol{\xi}_g)$ is the density of $\mathbf{X}|Z = g$, depending on $\boldsymbol{\xi}_g$, and $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G)$ is the resulting parameter space. We will focus on the particular case of the *linear Gaussian CWM*, given by

$$p(\mathbf{x}, y; \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g \phi_1(y; \boldsymbol{b}'_g \mathbf{x} + b^0_g, \sigma^2_g) \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2}$$

where $\phi_d(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes the density of the d-variate Gaussian distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Y is related to $\mathbf{X}$ by a linear model in (2), that is, $Y = \boldsymbol{b}'_g \mathbf{x} + b^0_g + \varepsilon_g$ with $\varepsilon_g \sim N(0, \sigma^2_g)$, $\boldsymbol{b}_g \in \mathbb{R}^d$, $b^0_g \in \mathbb{R}$, $\sigma^2_g \in \mathbb{R}^+$, for every $g = 1, \ldots, G$. Under the given framework, the parameters are denoted with:

$$\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_{G-1}, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_G, b^0_1, \ldots, b^0_G, \sigma^2_1, \ldots, \sigma^2_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}. \tag{3}$$

Based on a set of N i.i.d. samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$ drawn from $(\mathbf{X}, Y)$, maximum likelihood parameter estimation for the linear Gaussian CWM is generally carried out by means of the EM algorithm (Dempster et al., 1977). The observed log-likelihood function

$$\ell(\boldsymbol{\Theta}|\mathbf{X}, Y) = \sum_{i=1}^{N} \log \left[\sum_{g=1}^{G} \pi_g \phi(y_i; \boldsymbol{b}'_g \mathbf{x}_i + b^0_g, \sigma^2_g) \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right], \tag{4}$$

is maximized with respect to (3), and the Bayesian Information Criterion (Schwarz, 1978) is widely used for model selection.

Unfortunately, ML inference on models based on normal assumptions suffers from two major drawbacks. First off, the objective function in (4) is unbounded over $\boldsymbol{\Theta}$ so its optimization results in an ill-posed mathematical problem. Secondly, the resulting inference is strongly affected by outliers: see, e.g., Huber and Ronchetti (2009). To overcome both issues in this specific CWM framework, García-Escudero et al. (2017) introduced the Cluster Weighted Robust Model (CWRM). CWRM is based on the maximization of the trimmed log-likelihood (Neykov et al., 2007):

$$\ell_{\text{trimmed}}(\boldsymbol{\Theta}|\mathbf{X}, Y) = \sum_{i=1}^{N} z(\mathbf{x}_i, y_i) \log \left[\sum_{g=1}^{G} \pi_g \phi(y_i; \boldsymbol{b}_g' \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right], \tag{5}$$

where $z(\cdot, \cdot)$ is a 0-1 trimming indicator function denoting whether observation $(\mathbf{x}_i, y_i)$ is trimmed off ($z(\mathbf{x}_i, y_i) = 0$), or not ($z(\mathbf{x}_i, y_i) = 1$). A fixed fraction α of observations is left unassigned by setting $\sum_{i=1}^{N} z(\mathbf{x}_i, y_i) = N - [N\alpha]$, with α denoting the trimming level. Impartial trimming ensures that the $\alpha 100\%$ of most outlying units, according to the postulated model, is not accounted for in the optimization procedure, ultimately producing an estimator with desirable robustness properties (Hennig, 2004). Algorithmically, trimming is implemented by means of a “concentration” step (Rousseeuw and Driessen, 1999) carried out at each iteration of the EM algorithm. The interested reader is referred to García-Escudero et al. (2017) for a thorough description of the algorithm proposed.

To deal with the unboundedness of (4) a doubly-constrained maximization is considered in the CWRM specification, extending the approach proposed in Hathaway (1985) to the CWM. The first constraint is applied to the set of eigenvalues $\{\lambda_l(\boldsymbol{\Sigma}_g)\}_{l=1,\dots,d}$ of the scatter matrices $\boldsymbol{\Sigma}_g$ by requiring

$$\lambda_{l_1}(\boldsymbol{\Sigma}_{g_1}) \leq c_X \lambda_{l_2}(\boldsymbol{\Sigma}_{g_2}) \qquad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G. \tag{6}$$

The second bound is enforced to the variances σ_g^2 of the regression error terms as follows

$$\sigma_{g_1}^2 \leq c_y \sigma_{g_2}^2 \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G. \tag{7}$$

The constants c_X and c_y in (6) and (7) are finite (not necessarily equal) real numbers, such that $c_X, c_y \geq 1$. They prevent degenerate cases with $|\boldsymbol{\Sigma}_g| \to 0$ and $\sigma_g^2 \to 0$ from appearing, allowing the discarding of uninteresting spurious solutions. Finally, notice that if

$$\hat{\boldsymbol{\Theta}}_G^{c_X, c_y} = \{\hat{\pi}_1, \dots, \hat{\pi}_{G-1}, \hat{\boldsymbol{b}}_1, \dots, \hat{\boldsymbol{b}}_G, \hat{b}_1^0, \dots, \hat{b}_G^0, \hat{\sigma}_1^2, \dots, \hat{\sigma}_G^2, \hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_G, \hat{\boldsymbol{\Sigma}}_1, \dots, \hat{\boldsymbol{\Sigma}}_G\}$$

is the optimal set of parameters found maximizing (5) for fixed G, c_X and c_y, and

$$p(\mathbf{x}_{(1)}, y_{(1)}; \hat{\boldsymbol{\Theta}}_G^{c_X,c_y}) \leq \ldots \leq p(\mathbf{x}_{(N)}, y_{(N)}; \hat{\boldsymbol{\Theta}}_G^{c_X,c_y})$$

are the sorted values when $p(\cdot, \cdot; \hat{\boldsymbol{\Theta}}_G^{c_X,c_y})$ is defined as in (2), then the non-trimmed units with $z(\mathbf{x}_i, y_i) = 1$ are those with

$$p(\mathbf{x}_i, y_i; \hat{\Theta}_G^{c_X,c_y}) \geq p(\mathbf{x}_{([N\alpha]+1)}, y_{([N\alpha]+1)}; \hat{\Theta}_G^{c_X,c_y}).$$

2.1 Penalized likelihood for the CWRM

We now need to specialize the general theory of model selection (refer to Claeskens and Hjort, 2008, for a detailed review) to derive a penalized likelihood criterion for the CWRM. Based on (5), for a specific value of α the number of components G and the hyper-parameters c_X and c_y will be chosen by minimizing the following Trimmed BIC (TBIC) criterion:

$$\text{TBIC}(G, c_X, c_y) = -2\ell_{\text{trimmed}}(\hat{\boldsymbol{\Theta}}_G^{c_X,c_y}|\mathbf{X}, Y) + \nu_G^{c_X,c_y}, \tag{8}$$

where $\ell_{\text{trimmed}}(\hat{\boldsymbol{\Theta}}_G^{c_X,c_y}|\mathbf{X}, Y)$ is the maximized trimmed log-likelihood for a model with G components and constraints c_X and c_y, while the term $\nu_G^{c_X,c_y}$ denotes a penalty factor accounting for model complexity. In particular, the adaptability entailed by relaxing the constrained estimation shall be taken into account in $\nu_G^{c_X,c_y}$, along the lines of Cerioli et al. (2018). Therefore, the following penalty term is proposed:

$$\begin{aligned}\nu_G^{c_X,c_y} =& \{(G-1) + Gd + G(d+1) + \\ & 1 + ((Gd-1) + Gd(d-1)/2)\,(1 - 1/c_X) + \\ & 1 + (G-1)\,(1 - 1/c_y)\} \log\left(N - [N\alpha]\right).\end{aligned} \tag{9}$$

In the first line of (9) we count parameters required for the mixture weights, the cluster means of the covariates, and the regression coefficients. Afterwards, we have the contributions given by modeling the $\boldsymbol{\Sigma}_g$ in $\mathbf{X}$. Based on their eigenvalue decomposition, we have 1 free eigenvalue and $Gd-1$ constrained eigenvalues, plus the $Gd(d-1)/2$ rotation matrices. Except for the first term, the remaining ones are multiplied by $(1 - 1/c_X)$ to account for constrained estimation. Finally, there is the part relative to modeling scatters for the regressions on $Y|\mathbf{X}$, with one free σ_g^2 and $G-1$ constrained by c_y. Again, except for the first

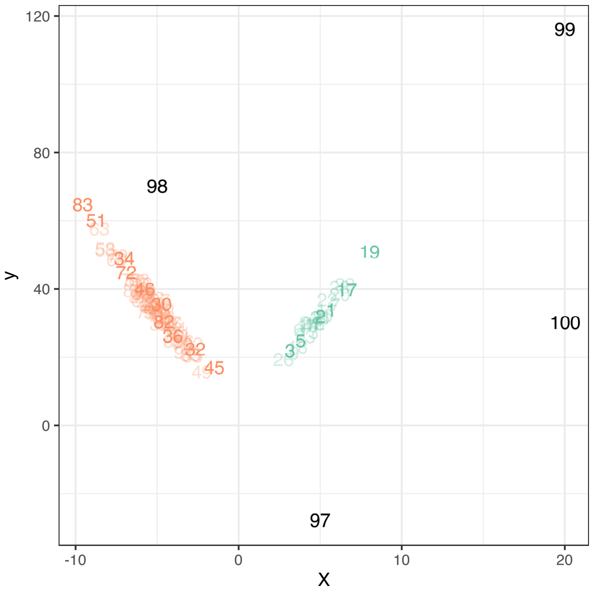

Figure 2: Toy data with the result of a bivariate linear Gaussian CWM with 4 appended outliers ($N = 100$). Row indexes are used as labels in the scatter plot.

term, the other ones should be multiplied by $(1-1/c_y)$ to incorporate the constraint induced by c_y. In expression (9), differently from Cerioli et al. (2018), we opted for multiplying all the variance parameters (rotation and eigenvalues) by the factor $(1 - 1/c_X)$, to account for the fact that rotation loses its meaning for $c_X \to 1$. Numerical experiments (not reported here) have nevertheless demonstrated that the two criteria provide substantial agreement when it comes to models ranking, as performed in our monitoring procedure. Additionally, yet another information criterion based on more general decompositions of the $\boldsymbol{\Sigma}_g$ scatter matrices, recently developed for Gaussian mixtures (García-Escudero et al., 2020, 2022), could be extended to the CWRM framework. Notwithstanding, exploring these new types of constraints is out of the scope of the present manuscript and the criterion defined in (8) will then be hereafter employed. Lastly, observe that the penalized criterion in (8) reduces to the standard Bayesian Information Criterion (Schwarz, 1978) when α goes to zero, and both c_X, c_y go to infinity.

3 Screening the space of solutions for CWRM

The selection of the most appropriate model among the set of potential solutions, as a function of G, c_X, c_y and α, results in a seemingly ungovernable task whenever little or no prior information is available in advance. This is the price to be paid for the great flexibility and robustness achieved by the Cluster Weighted Robust Model. Nonetheless, as we discussed in the introduction, in general a clustering process is not supposed to single out a unique and stand-alone result. To this extent, we propose here a two-step monitoring procedure to fully explore the space of CWRM solutions. In the first step, detailed in Section 3.1, the focus is on gaining insights for later investigation. Specific graphical and exploratory tools for CWRM are employed for determining one or more plausible values for the trimming level α. In the second step, described in Section 3.2, on the base of the values of α selected in the previous step, a further exploration is developed when G, c_X and c_y are free to vary within a grid of values and a list of candidate optimal solutions is generated. The quality of the restricted set of solutions is investigated by means of tailored silhouette plots; in this way, we can also inspect the nature and extent of the identified outliers. Lastly, the validity of the cluster weighted solutions is also assessed by means of the total sum of squares decomposition introduced in Ingrassia and Punzo (2020). A comprehensive account of this process is reported in Section 3.3.

Throughout the next subsections we will make use of a simple toy example to aid the understanding of the proposed methodology, and to motivate and justify the monitoring tools presented hereafter. In details, a total of 96 data points are generated from the linear Gaussian CWM in (2) with $d = 1$ and $G = 2$. In addition, 4 outlying points are appended to the synthetic dataset, with very distinct characteristics:

- a CWM outlier: a point non-outlying in the covariates and with a fitting regression line for one of the $G = 2$ components, but with an outlying pattern according to the joint CWM density (unit 97);
- a vertical outlier: a point non-outlying in the covariates but with a non-fitting regression line (unit 98);
- a group-specific good leverage point: a point outlying in the covariates but with a

fitting regression line for one of the $G = 2$ components (unit 99);

- a bad leverage point: a point outlying in the covariates and with a non-fitting regression line (unit 100).

The scatterplot of the resulting simulated dataset, encompassing $N = 100$ samples, is reported in Figure 2. Clearly, in this trivial situation, the correct $\alpha = 0.04$ and $G = 2$ can be immediately eyeballed by looking at the bivariate plot. Notice that, despite being in principle less relevant for applications, even in this simple situation the true constraints values $c_X = c_y = 2$ cannot be easily inferred. We provide a monitoring procedure to select α, G, c_X, c_y in a semi-automatic way in the upcoming sections.

3.1 Step 1: monitoring tools to validate the trimming level

In robust procedures based on hard trimming, a very crucial role is played by α, which determines the size of the subsets over which the likelihood is maximized.

In this phase, we leverage from previous work in Riani et al. (2019), where a plot of the Adjusted Rand Index (ARI) between consecutive cluster allocations for a grid of α has been proposed to visually assess the contamination rate in a given dataset. The mentioned approach is based on a first bet on the value of G and on the single constraint needed for Gaussian mixtures. The ARI plot shows changes in the clustering structure for different trimming levels, remaining close to its maximum value when solutions are similar one to another. Indeed, this is an effective tool to detect noise in the form of bridges, when the proper underlying partition is uncovered only by adopting the correct level of trimming. However, in case of scattered noise, we argue that the clustering structure could evolve very smoothly from an initial solution, obtained without trimming, to a pretty different final one. Hence, the ARI plot between consecutive allocations may display no jumps, resulting in no meaningful indications about an appropriate choice for α. The same consideration holds true in case of point-wise contamination, when one small group of observation is fitted by one component of the mixture, without biasing model estimation for the main bulk of the data.

To overcome these limitations, we introduce two modifications to the monitoring strategy in Riani et al. (2019). On the one hand, instead of a-priori setting any hyper-parameter, we let the best model be determined by the penalized criterion introduced in Section 2.1, suitably varying G, c_X and c_y, for each α. In details, we consider sequences $G^* = \{1, 2, \ldots, G^{\text{MAX}}\}$, where G^{MAX} is the maximal number of sought clusters, and c_X^*, c_y^* of possible constraint values for c_X and c_y, respectively. Without loss of generality and for easing the notation, in the following we will adopt the same grid of restrictions $c_X^* = c_y^* = c^* = \{c_1, \ldots, c_C\}$, $C \in \mathbb{N}$, for both the covariates and the regression errors. The finite sequence of powers of 2, $c_1 = 2^0, c_2 = 2^1, \ldots, c_C = 2^{C-1}$, for example, is a straightforward way to generate a grid of values that becomes sharper close to 1, to account for possible different solutions that arise when the constraints are tighter, compared to when c_X and c_y are large (García-Escudero et al., 2015). In this way, we aim at fully exploring the model space for a grid of α values, avoiding any subjectivity in the selection. On the other hand, we enrich the set of monitoring aids, extending the ARI plot with graphical tools tailored for the CWRM framework. In details, varying the trimming level α, the following metrics are inspected:

i) *Groups proportion:* estimated proportion of observations in each component are profiled in a stacked barplot;

ii) *Regression slopes* are profiled via a G-lines plot, to monitor increase and/or decrease in parameters magnitude;

iii) *Standard deviations* $\hat{\sigma}_g$, $g = 1, ..., G$ *of the residual error terms* are represented in a G-lines plot, profiling the increase and/or decrease in variability around the regression fits;

iv) *Cluster volumes* are monitored via a G-lines plot, profiling the increase and/or decrease in $\left|\hat{\boldsymbol{\Sigma}}_g\right|^{1/d}$, $g = 1, \ldots, G$;

v) ARI *between consecutive cluster allocations* is tracked via a line plot, as in Riani et al. (2019). Notice that only the resulting data partition of adjacent solutions are compared; while the true label set, that may not even exist, is never considered in building such metric.

vi) *Proportion of doubtful assignments* (e.g., the proportion of units whose cluster allocation is uncertain) is monitored via a line plot. For each observation i we compute the a posteriori probabilities $D_g(\mathbf{x}_i, y_i)$ to belong to cluster g (up to a normalizing constant):

$$D_g(\mathbf{x}_i, y_i) = \hat{\pi}_g \phi_1(y_i; \hat{\boldsymbol{b}}_g' \mathbf{x}_i + \hat{b}_g^0, \hat{\sigma}_g^2) \phi_d(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g), \quad g = 1, \dots, G. \tag{10}$$

Such quantities are sorted, yielding $D_{(1)}(\mathbf{x}_i, y_i) \leq \ldots \leq D_{(G)}(\mathbf{x}_i, y_i)$. Recall that the usual maximum a posteriori (MAP) rule assigns observation i to the component g for which $D_g(\mathbf{x}_i, y_i)$ is highest, i.e., $D_g(\mathbf{x}_i, y_i) = D_{(G)}(\mathbf{x}_i, y_i)$. Then, we compute the discriminant factor DF(i) to measure the strength of the group membership of observation i as a function of the posterior probabilities (Van Aelst et al., 2006; García-Escudero et al., 2011; Fritz et al., 2012). In details, in the framework of CWRM, DF(i)'s are defined as follows:

$$\mathrm{DF}(i) = \begin{cases} \log\left(D_{(G-1)}(\mathbf{x}_i, y_i)/D_{(G)}(\mathbf{x}_i, y_i)\right) & \text{for } i \text{ not trimmed} \\ \log\left(D_{(G)}(\mathbf{x}_i, y_i)/D_{(G)}(\mathbf{x}_{([N\alpha]+1)}, y_{([N\alpha]+1)})\right) & \text{for } i \text{ trimmed} \end{cases} \quad i = 1, \dots, N. \tag{11}$$

For a non trimmed unit i, DF(i) assesses the strength of the assignment of unit i, comparing the largest $D_{(G)}(\mathbf{x}_i, y_i)$ to the second best possible cluster assignment $D_{(G-1)}(\mathbf{x}_i, y_i)$. If i clearly belongs to its assigned group, $D_{(G)}(\mathbf{x}_i, y_i) \gg D_{(G-1)}(\mathbf{x}_i, y_i)$, yielding a large negative value for DF(i). For a trimmed unit i, instead, DF(i) evaluates the strength of the trimming decision. It compares the unnormalized posterior probability $D_{(G)}(\mathbf{x}_i, y_i)$ for i to belong to its most plausible cluster (to which it is not assigned) to the maximum posterior probability of the first not trimmed unit $\left(\mathbf{x}_{([N\alpha]+1)}, y_{([N\alpha]+1)}\right)$. In this way, DF$(i) \leq 0$ for every i, and large DF values (i.e., values close to zero) indicate doubtful assignments or trimming decisions. We will deem unit i to be doubtfully assigned if DF(i) is greater than a given threshold. Along the lines of Fritz et al. (2012), observation i is considered as doubtful if the strength of the assignment to the second best cluster is larger than one tenth of the actually made decision, say DF$(i) \geq \log(1/10)$.

Jointly monitoring the evolution of the selected metrics is an effective approach to uncover the most sensible trimming level/levels to be employed in the subsequent analysis. In

Figure 3 we report the resulting graphical output for the toy dataset introduced in the previous section. As expected, it is immediately noticed that an $\alpha \geq 0.04$ is needed to provide reliable inference, and several patterns pointing this out can be identified. First off, when α is smaller than 0.04 models with extra components ($G = 4$ and $G = 3$) are preferred according to the TBIC defined in (8) because they are needed in order to fit the standalone units that should have not entered the estimation procedure. By looking at the stacked barplot we can nevertheless discern that these are nothing but spurious clusters with very low mixing proportion, which disappear when a higher trimming level is considered. The line plots related to model parameters remain fairly stable throughout the α grid for all components but the spurious ones, further validating the conjecture that the two extra classes are random patterns inappropriately captured by the (not robust enough when $\alpha < 0.04$) model. As anticipated previously, when outliers enter one at a time in the search, clear evidence may not be extracted by looking at the ARI panel only, so much so that other plotting tools are needed even in this very simple example. Lastly, the proportion of doubtful assignments line reaches its minimum at the correct trimming level $\alpha = 0.04$, confirming once more such choice for this dataset.

To carry out the first monitoring step, the well-known label-switching problem of mixture models should be tackled with extreme care. Otherwise, the component-dependent metrics, concerning estimated model parameters, cannot be safely compared across trimming levels. Notice that the only metrics that do not suffer from the label-switching problem are the ARI *between consecutive cluster allocations* and the *proportion of doubtful assignments*. The former is a general measure of similarity between two partitions (with possibly different number of groups), while the latter simply counts the number of units that are doubtfully assigned according to a specific CWRM solution. We construct a relabeling strategy based on the postulated model density. In details, the relabeling procedure proceeds as follows: starting from the solution obtained with the highest amount of trimming, the $(d + 1)$-dimensional quantities $\mathbf{r}_g = \left(\hat{\boldsymbol{\mu}}_g, \hat{b}_g^0 + \hat{\boldsymbol{b}}_g' \hat{\boldsymbol{\mu}}_g\right)$ are stored for each g, $g = 1, \ldots, G$. Notice that $\mathbf{r}_g$ is exactly the estimated marginal d-dimensional cluster mean, to which the conditional estimate according to the regression term is appended. The quantities $\mathbf{r}_g$ are the g cluster "representatives" that are employed in the relabeling

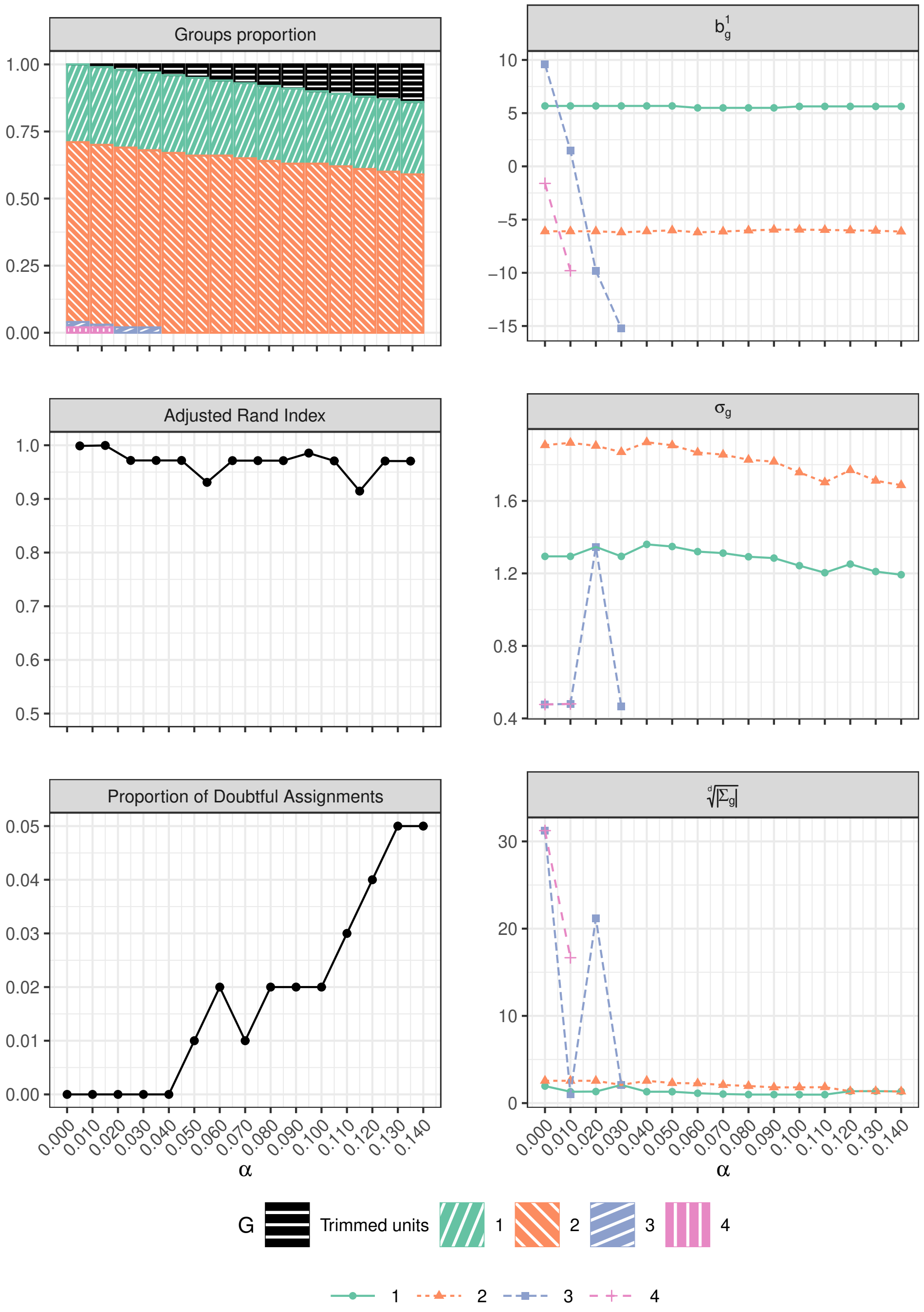

Figure 3: Monitoring tools in Step 1 for the "Toy dataset". From the top left corner, moving counterclockwise the following plots are displayed: groups proportion, ARI between consecutive cluster allocations, proportion of doubtful assignments, cluster volumes, regression standard deviations and regression slopes (see bulletpoints i) to vi) in Section 3.1 for details). Metrics are monitored as a function of the trimming level α.

process of the subsequent solutions, via the MAP rule, when the trimming level decreases. Specifically, for the l-th value α_l in the sequence of considered trimming levels, the label associated to $\mathbf{r}_c = \left(\hat{\boldsymbol{\mu}}_c, \hat{b}_c^0 + \hat{\boldsymbol{b}}_c' \hat{\boldsymbol{\mu}}_c\right)$, $c = 1, \ldots, G$ is computed as follows:

$$\arg\max_g \hat{\pi}_{g,l} \phi_1(\hat{b}_c^0 + \hat{\boldsymbol{b}}_c' \hat{\boldsymbol{\mu}}_c; \hat{\boldsymbol{b}}_{g,l}' \hat{\boldsymbol{\mu}}_c + \hat{b}_{g,l}^0, \hat{\sigma}_{g,l}^2) \phi_d(\hat{\boldsymbol{\mu}}_c; \hat{\boldsymbol{\mu}}_{g,l}, \hat{\boldsymbol{\Sigma}}_{g,l}) \tag{12}$$

where with the subscripts g, l we denote the estimated parameters of the g-th group with trimming level α_l. Whenever a solution possesses a higher number of clusters than the previous one, a new $\mathbf{r}_g$ is computed and stored as the representative of such new component. Conversely, whenever a solution has a lower number of clusters than the previous one, the $\mathbf{r}_g$ quantity for the merged components is identified and the set of representative units is updated accordingly. Clearly, this heuristic may fail in cases where the clustering structure deeply changes when evolving from a solution to its adjacent one. Nonetheless, when the α grid is quite dense, the failure of this procedure clearly indicates that some mechanism must have spoiled the inferential process. Alternatively, a nonparametric approach based on depth measures (Singh et al., 1999) may also be profitably employed in this context.

3.2 Step 2: monitoring tools for choosing G, c_X and c_y

Having identified a/some “reasonable” value/values for α, we propose to screen the space of solutions $\mathcal{E}_0$ generated by varying the number of clusters G, and the pair of hyper-parameters c_X and c_y over a grid, conditioned on a fixed trimming level.

We aim at collecting a reduced list $\mathcal{O}$ of “optimal” solutions, qualified by two features: their stability across hyper-parameter values, and their optimality in terms of the penalized criterion defined in (8). We elaborate on the two algorithms presented in Cerioli et al. (2018) by unifying them in a single searching process, enabled to encompass the more complex framework of Cluster Weighted modeling.

Given a triplet (G, c_X, c_y), let $P(G, c_X, c_y)$ denote the partition into G clusters, obtained by optimizing (5) under the constraints (6) and (7). Let $\text{ARI}(A, B)$ denote the ARI between partitions A and B. We consider that two partitions A and B are “similar” when $\text{ARI}(A, B) \geq \eta$, for a fixed threshold η. Clearly, the higher the value η the greater the number of retained distinct solutions. While η can certainly be application-dependent, its selection does not dramatically affect the procedure described hereafter, and values for η equal to 0.7 or 0.8 are generally considered in the literature (Riani et al., 2019; Cerioli et al., 2018). We will set $\eta = 0.8$ in all applications described in Section 4. Lastly, let us reconsider the sequences G^*, c_X^* and c_y^* previously employed in the first step. In this setting, the proposed procedure for finding $\mathcal{O}$, the set of $T \leq L$ optimal solutions is summarized in Algorithm 1, where L denotes a pre-specified upper bound for the maximum number of optimal solutions to be retained. The resulting strategy simplifies the set of operations originally proposed in Cerioli et al. (2018).

Algorithm 1 Optimal solutions finder

```
Initialize the space to be explored $\mathcal{E}_0 = \{(G, c_X, c_y) \in G^* \times c_X^* \times c_y^*\}$ and the empty list of optimal solutions $\mathcal{O}$
while $\mathcal{E}_t \neq \emptyset$ or $t \leq L$ do
    Obtain $(G^t, c_X^t, c_y^t) = \arg\min_{(G,c_X,c_y)\in\mathcal{E}_{t-1}} \text{TBIC}(G, c_X, c_y)$ and append it to list $\mathcal{O}$
    Obtain from $\mathcal{E}_{t-1}$ the set $\mathcal{I}$ of triplets $(G, c_X, c_y)$ that induce a "similar" partition to $P(G^t, c_X^t, c_y^t)$, that is
        $\mathcal{I} = \left\{(G, c_X, c_y) : \text{ARI}\left(P(G, c_X, c_y), P(G^t, c_X^t, c_y^t)\right) \geq \eta, \quad \text{for} \quad (G, c_X, c_y) \in \mathcal{E}_{t-1}\right\}$
    $\mathcal{E}_t = \mathcal{E}_{t-1} \setminus \mathcal{I}$
end while
return $\mathcal{O} = \{(G^1, c_X^1, c_y^1), \dots, (G^T, c_X^T, c_y^T)\}$
```

Once the optimal set has been identified, we further define two sets of "best" and "stable" intervals for each optimal solution (G^t, c_X^t, c_y^t) in $\mathcal{O}$, respectively defined as follows:

$$\begin{aligned} \mathcal{B}_t = \Big\{(G, c_X, c_y) : \text{TBIC}(G, c_X, c_y) \leq \text{TBIC}(G^{t+1}, c_X^{t+1}, c_y^{t+1}) \\ \text{and} \\ \text{ARI}\left(P(G^t, c_X^t, c_y^t), P(G, c_X, c_y)\right) \geq \eta \quad \text{for} \quad (G, c_X, c_y) \in \mathcal{E}_0\Big\}, \end{aligned} \tag{13}$$

$$\mathcal{S}_t = \left\{(G, c_X, c_y) : \text{ARI}\left(P(G^t, c_X^t, c_y^t), P(G, c_X, c_y)\right) \geq \eta \quad \text{for} \quad (G, c_X, c_y) \in \mathcal{E}_0\right\}. \tag{14}$$

In $\mathcal{B}_t$, we want to identify the set of parameter values for which an optimal solution remains "best". In doing so, we include in $\mathcal{B}_t$ all solutions in $\mathcal{E}_0$ ARI-similar to the optimal, and

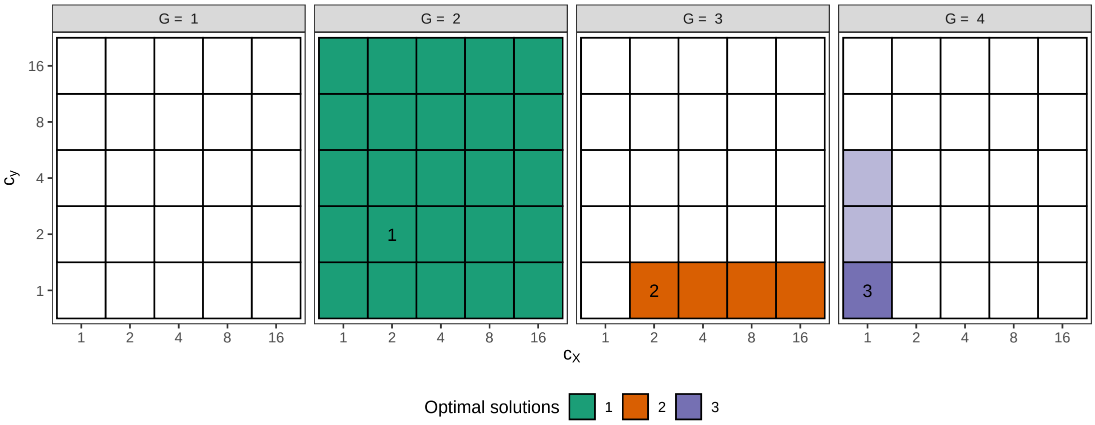

Figure 4: Monitoring tools in Step for the "Toy dataset". The optimal solutions are indicated by the cells with ordinal numbers 1, 2 and 3 ($\alpha = 0.04$). Each solution is featured by one color, showing the range of cases in which it is best (darker opacity cells), and stable (lighter opacity cells), varying G, c_X (horizontal axis) and c_y (vertical axis) in $\mathcal{E}_0$.

not worse than the next optimal solution. In $\mathcal{S}_t$, we want to identify the set of parameter values for which an optimal solution is "stable", including in $\mathcal{S}_t$ all solutions ARI-similar to the optimal. Therefore, we have $\mathcal{B}_t \subseteq \mathcal{S}_t$ and, graphically, solutions in $\mathcal{B}_t$ and $\mathcal{S}_t$ will be represented by darker and lighter opacity cells, with colors depending on the optimal solution they correspond to.

By referring again to the toy dataset of Figure 2, and after having selected $\alpha = 0.04$ in the first step, the resulting output is depicted in Figure 4. The optimal solutions are indicated by ordinal numbers, from 1 to 4. A different color is associated to each optimal solution, with a darker nuance used to depict the range of hyper-parameters for which it remains the best, while a lighter shading indicates the set of associated stable solutions. This novel graphical tool is an extension of the car-bike plot (Cerioli et al., 2018) specifically designed to monitor the space of solutions of CWRMs. Each facet incorporates models with the same number of mixture components, while the x-axis and the y-axis display different values for c_X and c_y, respectively. As expected, it is evident that the first optimal solution is achieved setting $G = 2$, which remains best (i.e., with lower TBIC with respect to the second optimal) for the whole range of c_X and c_y. We notice that the first optimal solution recovers the true values of $c_X = c_y = 2$. The second optimal model possesses three components with homoscedastic error lines ($c_y = 1$), while the third one has also spherical covariance structure ($c_X = 1$). It is apparent that for this toy dataset the only sensible result is the first one, nevertheless it may happen that more than one solution, depending on the field of application, could be of interest. In this regard, a thought-provoking socio-economic analysis is reported in Hennig and Liao (2013).

3.3 Exploring optimal solutions and related outliers

Once a reduced set of optimal solutions has been identified, it remains to determine the one that could be, in principle, better suited to solve the problem at hand and, if of interest, to characterize further the units that have been flagged as outliers. While the thorough treatment of these two issues is application-dependent and thus domain expertise should never be left aside, we hereafter provide yet another graphical tool to assist the decision-making process.

We defined in (11) the Discriminant Factor: a quantity that measures the strength of the assignment/trimming of each unit compared to the second best alternative. Its definition within the CWRM framework entirely agrees with the original formulation, introduced in García-Escudero et al. (2011), developed for robust Gaussian mixtures. Nonetheless, the CWM characterization in (2) clearly breaks down the overall mixture density in the contribution of the g regression lines and the component-wise random covariates. Therefore, the strength of the assignment/trimming may be driven to a greater extent by one of the two terms comprising the CWM formulation. On this wise, and specifically for the CWRM, we define the $Y|X$ *Discriminant Factor* $\mathrm{DF}_{Y|X}(i)$ and the X *Discriminant Factor* $\mathrm{DF}_X(i)$ for observation i. In order to do that, let us consider two additional quantities:

$$D_g^{Y|X}(\mathbf{x}_i, y_i) = \hat{\pi}_g \phi_1(y_i; \hat{\boldsymbol{b}}_g' \mathbf{x}_i + \hat{b}_g^0, \hat{\sigma}_g^2) \quad \text{and} \quad D_g^X(\mathbf{x}_i) = \hat{\pi}_g \phi_d(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g),$$

for $g = 1, \ldots, G$. As done for the terms in (10), we sort the newly defined quantities

$$D_{(1)}^{Y|X}(\mathbf{x}_i, y_i) \leq \ldots \leq D_{(G)}^{Y|X}(\mathbf{x}_i, y_i), \text{ and, analogously, } D_{(1)}^X(\mathbf{x}_i) \leq \ldots \leq D_{(G)}^X(\mathbf{x}_i).$$

Consequently, the $Y|X$*-Discriminant Factor* and the X*-Discriminant Factor* are respectively defined as follows:

$$\mathrm{DF}_{Y|X}(i) = \begin{cases} \log\left(D^{Y|X}_{(G-1)}(\mathbf{x}_i, y_i)/D^{Y|X}_{(G)}(\mathbf{x}_i, y_i)\right) & \text{for } i \text{ not trimmed} \\ \log\left(D^{Y|X}_{(G)}(\mathbf{x}_i, y_i)/D^{Y|X}_{(G)}(\mathbf{x}_{([N\alpha]+1)}, y_{([N\alpha]+1)})\right) & \text{for } i \text{ trimmed} \end{cases} \quad i = 1, \dots, N, \tag{15}$$

and

$$\mathrm{DF}_{X}(i) = \begin{cases} \log\left(D^{X}_{(G-1)}(\mathbf{x}_i)/D^{X}_{(G)}(\mathbf{x}_i)\right) & \text{for } i \text{ not trimmed} \\ \log\left(D^{X}_{(G)}(\mathbf{x}_i)/D^{X}_{(G)}(\mathbf{x}_{([N\alpha]+1)})\right) & \text{for } i \text{ trimmed} \end{cases} \quad i = 1, \dots, N. \tag{16}$$

The rationale behind the definitions of $\mathrm{DF}_{Y|X}$ and DF_X mirrors the one outlined in Section 3.1 for DF, with the difference that with (15) and (16) we are separately assessing the strength of the assignment/trimming for each unit in relation to the regression lines and the covariates, respectively. Contrarily to DF, it thus may happen that $\mathrm{DF}_X(i)$ and $\mathrm{DF}_{Y|X}(i)$ assume a positive value for a trimmed unit i. The reason being that trimming is enforced by looking at the overall contribution of observation i to the likelihood in the model specification, i.e., unit i is discarded according to the CWM density, but it may not have been the case if we were to individually look at the two terms in (2). To illustrate the idea, let us go back to the toy example employed all over these sections: observation 98 is a vertical outlier, meaning that the reason for it to be trimmed is due to its poor fitting to the regression term, whilst if we were to evaluate its marginal density on X only we would not have discarded it. The reverse argument can be made for unit 99, a group-specific good leverage point. Silhouette plots (Rousseeuw, 1987b) can be employed for visually exploring the defined discriminant factors. Figure 5 reports 3 panels in which $\mathrm{DF}(i)$, $\mathrm{DF}_{Y|X}(i)$ and $\mathrm{DF}_X(i)$ are displayed for the first optimal solution recovered for the toy dataset, alongside the resulting scatter plot. We notice that no observation is doubtfully assigned according to the CWM discriminant factor because $\mathrm{DF}(i)$ is lower than $\log(1/10)$ (dashed red line) for all $i = 1, \dots, N$. On the other hand, by inspecting $\mathrm{DF}_{Y|X}$ and DF_X for the 4 trimmed units we can easily uncover more details on why such observations were trimmed. While unit 100 (bad leverage point) showcases very low values for both $\mathrm{DF}_{Y|X}$ and DF_X, $\mathrm{DF}_{Y|X}(99)$ is very close to 0, indicating that unit 99 does not seem to be an outlier according to the regression line. Indeed, unit 99 is a good leverage point. Conversely, $\mathrm{DF}_X(98)$ is positive: observation 98 is a vertical outlier and it should not have been trimmed according to the marginal

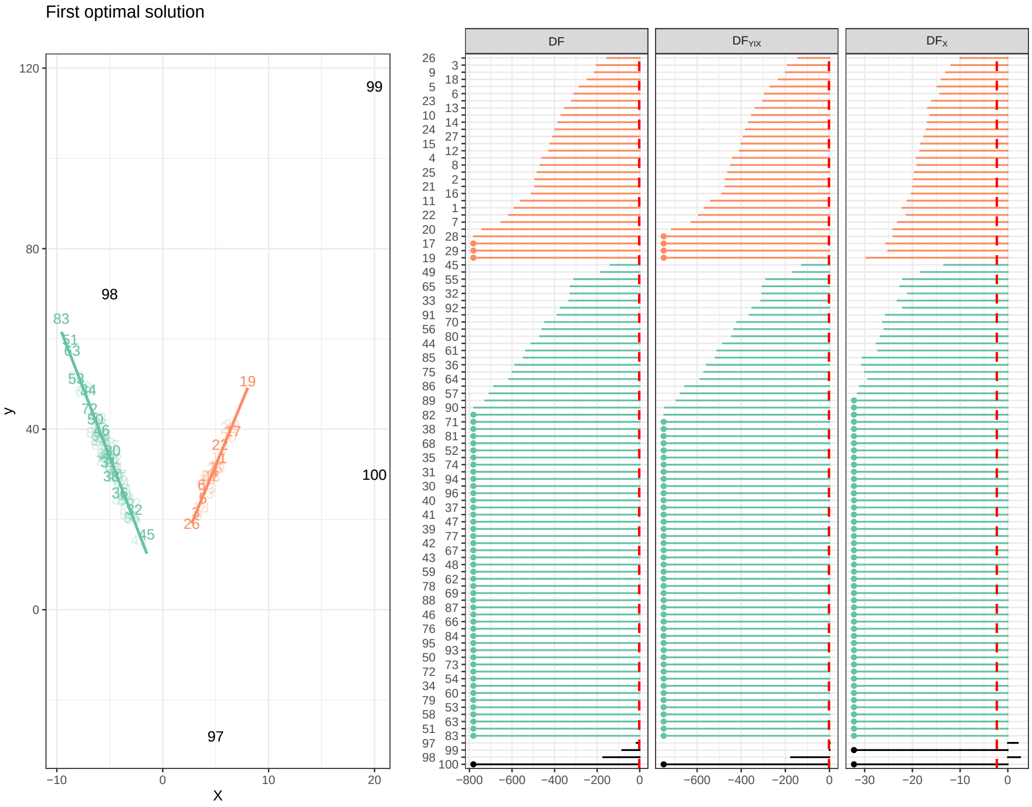

Figure 5: Toy dataset, first optimal solution. Scatter plot of the estimated model (left panel) and silhouette plots displaying $\mathrm{DF}(i)$, $\mathrm{DF}_{Y|X}(i)$ and $\mathrm{DF}_X(i)$, $i = 1, \ldots, N$ (right panel). Row indexes, ordered according to DF, are reported on the y-axis. Dashed red line superimposed at $\log(1/10)$. Solid dots at the left-most part of the bar plots indicate a resulting smaller value, not displayed for visualization purposes. Trimmed units are colored in black.

distribution of X only. Lastly, unit 97 is, as previously defined, a CWM outlier since its outlyingness results from the CWM density, and points as such showcase positive values for both $\text{DF}_{Y|X}$ and DF_X. The Discriminant Factors can also be employed to evaluate which optimal solution shall be preferred. Along with it, the *CWM decomposition of the total sum of squares* (Ingrassia and Punzo, 2020) provides a validation tool, specifically tailored for CWM, in which a three-term decomposition of the total variability on Y is produced. Figure 6 displays a ternary diagram showcasing an example of such decomposition for the top three solutions of the Toy dataset. NBSS represents the proportion of variability explained by the weighted differences between the weighted group means and the overall mean (i.e., the variability of Y explained by the latent group variable G), NEWSS is the proportion explained by the inclusion of the covariates $\mathbf{X}$ via the slope(s) of the local regressions while NRWSS accounts for the proportion of unexplained variability. Such a measure provides further insights into the set of optimal solutions retained by our monitoring procedure. A detailed interpretation of the plot and its usage for validating the optimal solutions found via the monitoring procedure is given in the Supplementary materials.

Some final considerations are due before focusing on applications to real multivariate data, to show the effectiveness of our proposal, as they are the main contribution of the present paper. A reduced version of the proposed methodology was briefly introduced in Cappozzo et al. (2021), where neither real data applications nor the novel definitions of Discriminant Factors for CWRM were included. Along the same research line, it is worth noting that Torti et al. (2021) recently presented a semiautomatic monitoring procedure for robust regression clustering, with a focus on international trade data. While the two approaches aim at solving a similar problem, the resulting methodologies are quite different: a discussion on the matter is reported in the last section of Cappozzo et al. (2021).

4 Application to real datasets

All the subsequent analyses are carried out by means of the `R` environment for statistical computing (R Core Team, 2022): an R package implementing the monitoring procedure is available at github.com/AndreaCappozzo/CWRMmonitor, while an R script providing a short tutorial on how to use the package can be found in the Supplementary Materials.

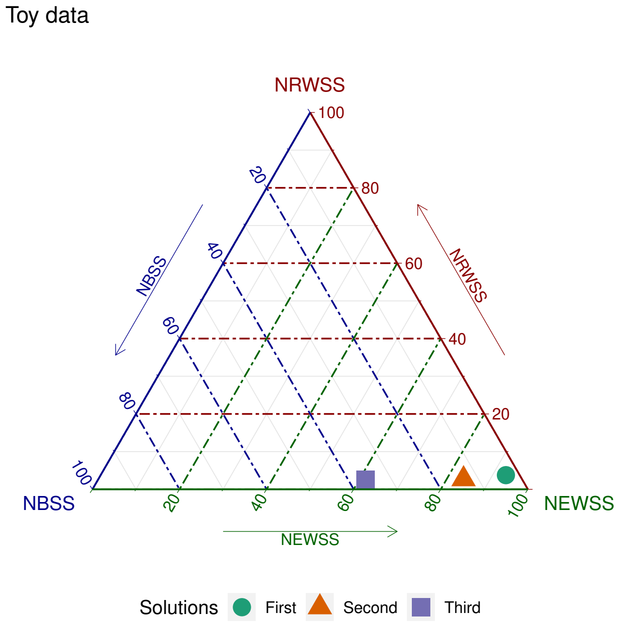

Figure 6: Toy data. Example of ternary diagram from the Total Sum of Squares Decomposition for the top three optimal solutions identified in Figure 4.

4.1 Tourism dataset

The effectiveness of cultural attractions in enhancing tourism flows is a widely debated issue, both in cultural economics and in tourism economics. Quite surprisingly, however, the evidence about the relationships between attendance at cultural attractions and tourist flows is restricted to specific, albeit interesting, case studies. In this analysis, we focus on the correlation between tourism flows and attendance at museums and monuments. In addition, we aim to study whether there could be a different forecast on museum attendances, based on tourist overnights, across the different seasons of the year. We analyze Italian data with a monthly frequency from January 1996 to December 2017 (Source: Italian Central Statistics Office, and Ministry of Cultural Heritage and Tourism). The data comprises $N = 264$ monthly bivariate observations: *attendance at museums and monuments* (Y, data in millions), and *tourist overnights* (X, data in millions).

Part of this dataset (subset of data from January 1996 to December 2010) has been

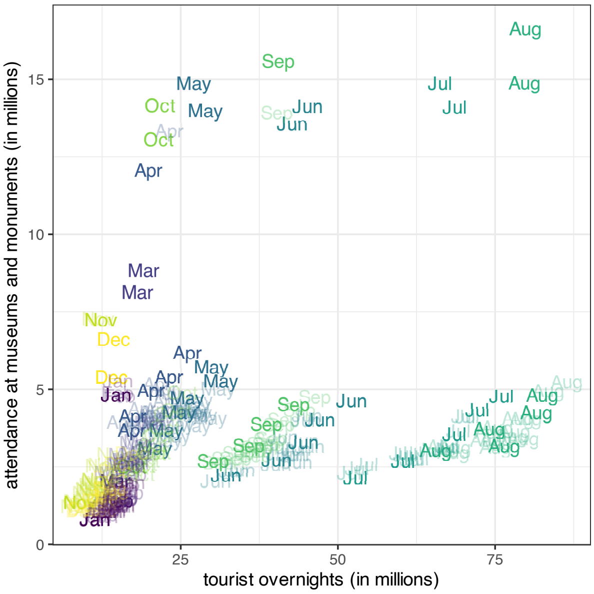

Figure 7: Tourism data: Plot of tourist overnights (X, in millions) and attendance at museums and monuments (Y, in millions) in Italy over the period from January 1996 to December 2017 ($N = 264$). Month abbreviations are used as labels in the scatter plot.

analyzed in Ingrassia et al. (2014) by applying the CWM with interesting results, demonstrating the differences in the intra-group marginal distributions and the linear models. In the following years, some manifest outliers appeared in the data, hence a robust approach is required. Figure 7 displays a scatterplot encompassing the entire 22-year period: few scattered units are visible on the top of the graph.

The two-step approach described in Section 3 has been applied to the Tourism dataset. The first monitoring phase generate the plots in Figure 8. Moving from left to right along the horizontal grid, an additional observation is sequentially trimmed from the data until a level of trimming equal to 0.095% is reached. Conditioning on the trimming level, the best solution minimizing (8) is retained considering up to 8 components in the mixture.

First of all, we see a dramatic drop in the volume of the scatter matrix $\sqrt[d]{|\hat{\boldsymbol{\Sigma}}_5|}$ (bottom right plot), after trimming the four more implausible observations. Presumably, some noise was previously fitted, increasing its variability. This conjecture is also confirmed by the monitored values of the ARI. While the majority of the groups have stable estimations for the regression coefficients and regression errors, it is apparent that the sixth group,

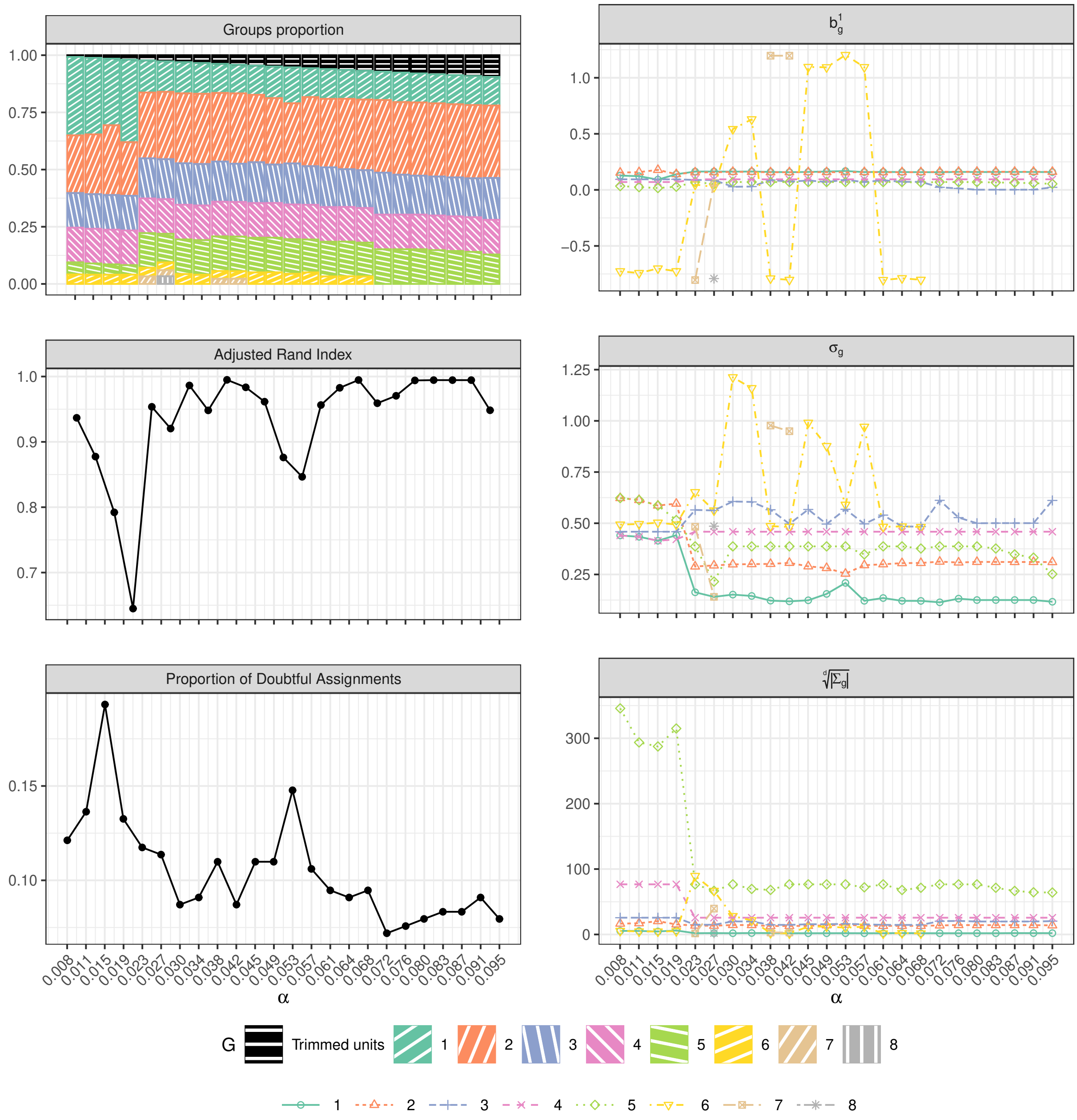

Figure 8: Monitoring tools in Step 1 for the Tourism dataset. The considered metrics are monitored as a function of the trimming level α. Unstable components disappear when $\alpha \geq 0.072$.

depicted in yellow in Figure 8, is quite an unstable component. Its regression coefficients and errors are very volatile across increasing values of α. In an attempt to match all these considerations, and looking at the minimum proportion of doubtful assignments, the information obtained by the monitoring approach suggests selecting $\alpha = 0.072$. With this choice, the 17 most scattered observations in Figure 7 are trimmed out and the tiny sixth group is discarded from contributing to the final model. It is worth noting that all the trimmed observations refer to the period January 2011-December 2017, which is characterized by a sudden increase of the volumes in terms of attendance at museums and monuments in comparison with the previous decades.

After Step 1, Step 2 focuses on the exploration of the model space when $\alpha = 0.072$. We aim to monitor the optimal solutions, considering the entire range of modeling choices, varying the constraints and the number of groups. Figure 9 displays the result of the second step. *Solution 1* involves 5 groups, and has good properties of optimality and stability. It is plotted in the leftmost panel in Figure 10. The local regression lines are also indicated, providing a different forecast on museum attendances, based on tourist overnights. By comparing with the leftmost panel in Figure 11 (referred to the same solution), we see that such forecasts depend on the different seasons of the year.

Solution 2 comprises 4 groups in data, and is stable for all values of the hyper-parameters c_X and c_y, for $G = 4$. The corresponding clustering representation is given in the central panels of Figures 10 and 11. Finally, *Solution 3* is again involving 5 groups, but its range of being best and stable is quite poor. The rightmost panels of Figure 10 and 11 display *Solution 3*.

A final consideration applies when comparing the three optimal solutions in Figure 10: there is a substantial agreement on the different forecast of museum attendances, based on tourist overnights, across the different seasons of the year. In addition, notice that the trimmed units entirely agree for the three optimal models, clearly highlighting the peculiar behavior of those data points and the necessity of discarding them from the estimation procedure. To interpret the role of the outlying units, the silhouette plot of the discriminant factors DF, $\mathrm{DF}_{Y|X}$ and DF_X for the trimmed units obtained for *Solution 1* is given in Figure 12. We see that trimmed units are mostly vertical outliers, having positive DF_X

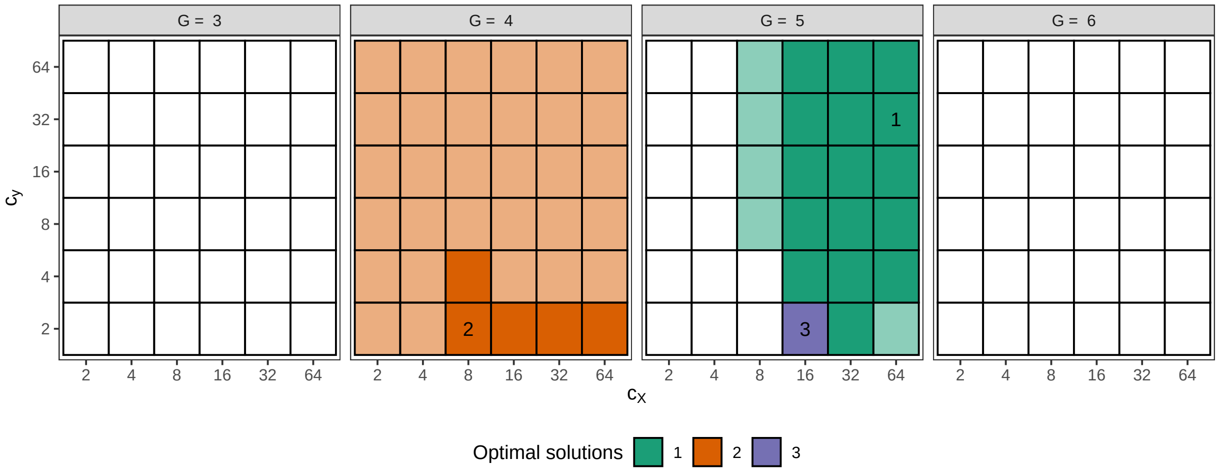

Figure 9: Monitoring tools in Step 2 for the Tourism dataset. The optimal solutions are indicated by the cells with ordinal numbers 1, 2 and 3 ($\alpha = 0.072$). Each solution is featured by one color, showing the range of cases in which it is best (darker opacity cells), and stable (lighter opacity cells), varying G, c_X (horizontal axis) and c_y (vertical axis) in $\mathcal{E}_0$. No optimal solutions found for $G = 2$ and for $G \geq 7$ (not displayed).

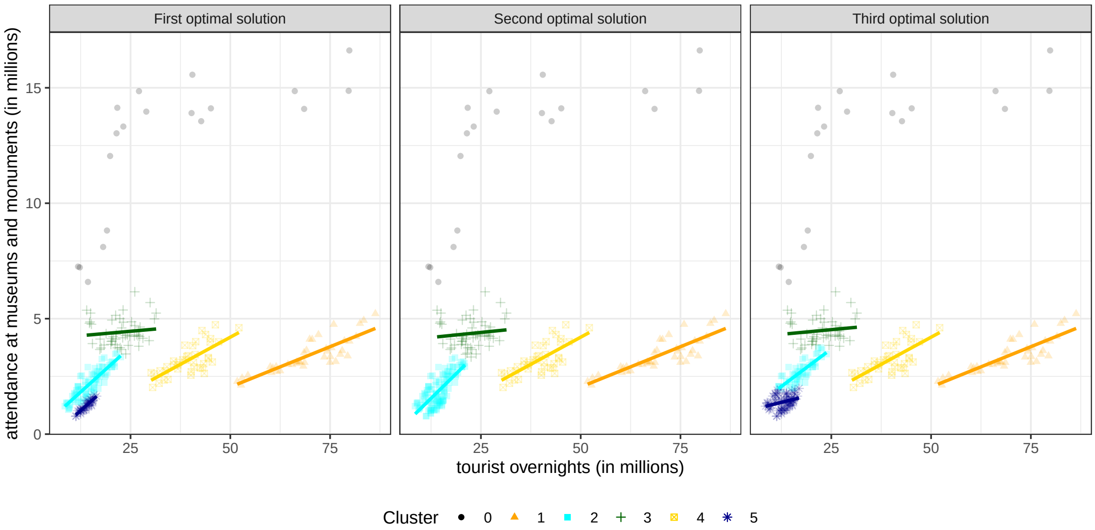

Figure 10: Tourism data. The top three best solutions obtained in Step 2, with the linear models estimated within the clusters.

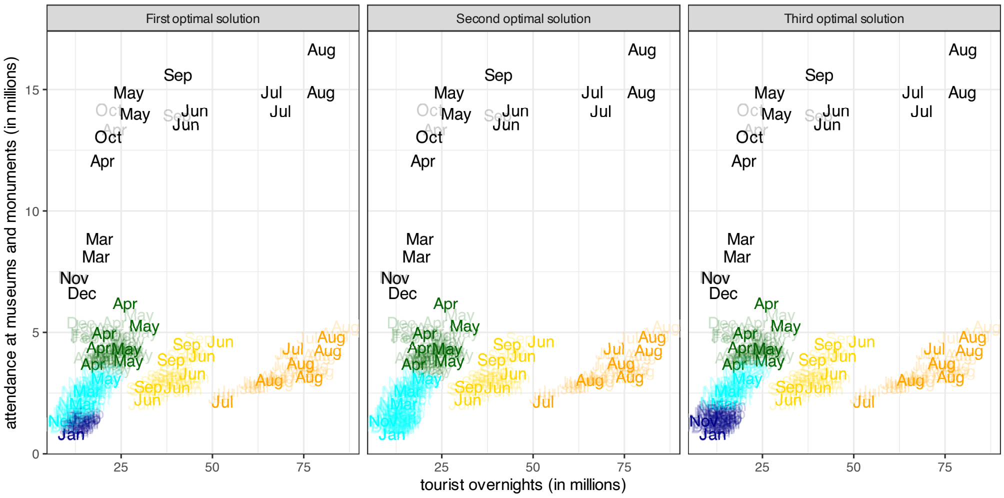

Figure 11: Tourism data. The top three best solutions obtained in Step 2, with respect to month labels.

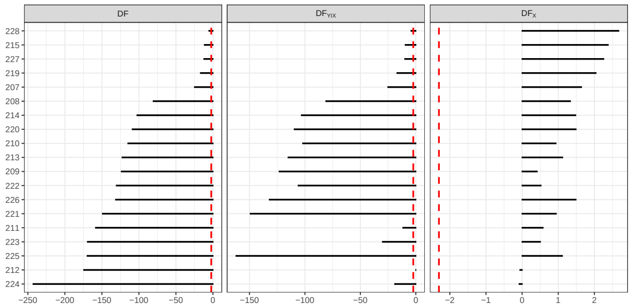

Figure 12: Tourism data. Silhouette plots of the discriminant factors DF, DF$_{Y|X}$ and DF$_X$ for the trimmed units in *Solution 1*.

and very low values of DF$_{Y|X}$. Although the reasoning above suggests that observations characterized by an unexpected volume of attendance at museums and monuments should be trimmed, one may nonetheless be interested in modeling them. Such an outcome can be obtained by setting a trimming level that is lower than $\alpha = 0.072$: an analysis on this regard is reported in the Supplementary Materials.

4.2 Female vole dataset

The next dataset encompasses $N = 86$ observations of 7 measurements from females of two species of voles, Microtus californicus and M. ochrogaster. The original dataset is described in Table 5.3.7 of Flury (1997) and it is available in the `Flury` `R` package. As previously done in the CWM literature (Subedi et al., 2013, 2015), we aim at regressing the *age* variable (measured in days) on $p = 6$ characteristics of the voles, namely *condylo incisive length*, *incisive foramen length*, *alveolar length of upper molar tooth row*, *zygomatic width*, *interorbital width* and *skull height*. The *species* variable acts as a grouping factor, and it is assumed unknown.

Results for the first monitoring step of Section 3.1 are displayed in Figure 13. We immediately notice that, conditioning on each trimming level, the model selection criterion in (8) always selects $G = 2$, identifying the definite existence of two clusters. Even though no extreme outliers are known to be present in this dataset, it seems that a small proportion of trimming induces a higher difference in the regression parameters of the two groups, particularly in the second and third element of the $\hat{\boldsymbol{b}}_g$ vectors. On the other hand, an α value higher than 0.10 seems to produce some abrupt changes in the line-plot patterns, indicating that a moderate trimming level may be sufficient to account for some mild outliers. Lastly, notice the drop in the proportion of doubtful assignments for $\alpha = 0.07$ and the stability in the ARI around such value. For all these reasons, the first monitoring step seems to suggest $\alpha = 0.07$ as a reasonable value for carrying out the subsequent analysis.

The second step of our monitoring procedure investigates the validity and stability of solutions for a given trimming level ($\alpha = 0.07$ in this case): results are graphically reported in Figure 14. As expected, the first optimal solution is obtained when $G = 2$, for which several solutions remain best when c_X and c_y is set higher than 16. The other optimal solutions are attained with $G = 3$, with small best and stable sets. The extra group arises as the species of M. ochrogaster voles is split in two sub-clusters: the same behavior has been previously observed with the linear Gaussian cluster weighted factor analyzers model (Subedi et al., 2013). On the other hand, the first optimal solution not only identifies the correct number of groups but it also recovers the highest classification accuracy obtained for this dataset (see Table 8 in Subedi et al. (2015)). After having a-posteriori assigned the trimmed units via the MAP rule, only two ochrogaster observations are misclassified. The very same results, not displayed here, were also obtained employing a smaller trimming level ($\alpha = 0.023$). Still, it was not possible to retrieve the same classification performance

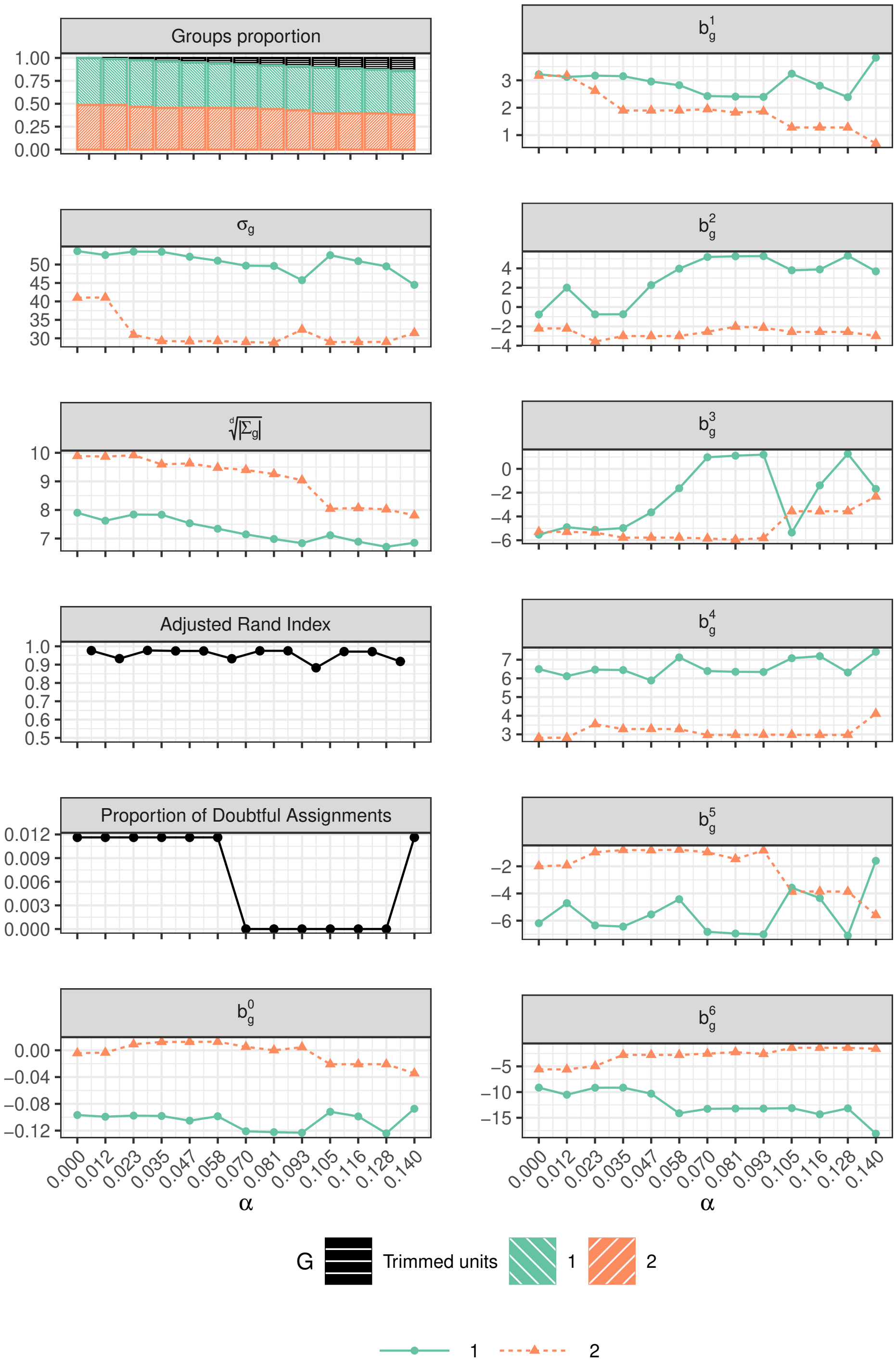

Figure 13: Monitoring tools in Step 1 for the Female vole dataset. The considered metrics are monitored as a function of the trimming level α. The proportion of doubtful assignment is minimized for $\alpha = 0.07$, and line plots for regression slopes and standard deviations remain stable for adjacent increasing values of $\alpha = 0.07$.

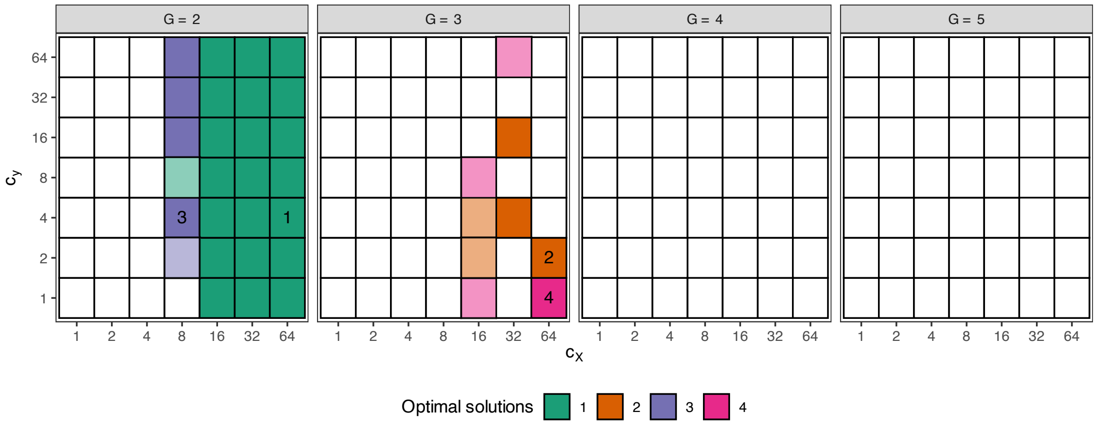

Figure 14: Monitoring tools in Step 2 for the Female vole dataset. The optimal solutions are indicated by the cells with ordinal numbers 1, 2, 3 and 4 ($\alpha = 0.07$). Each solution is featured by one color, showing the range of cases in which it is best (darker opacity cells), and stable (lighter opacity cells), varying G, c_X (horizontal axis) and c_y (vertical axis) in $\mathcal{E}_0$.

if no trimming is applied.

The pairs plots for the first and second optimal solutions are respectively displayed in Figure 15 (along with the three panels of silhouette plots) and Figure 16. From the silhouette plots we see that no observation is doubtfully assigned according to the CWM discriminant factor criterion: $\text{DF}(i) \leq \log(1/10)$ for all $i = 1, \ldots, 86$. Moreover, by inspecting $\text{DF}_{Y|X}$ and DF_X for the 6 trimmed units we can easily observe that unit 24 has been trimmed because it is a good leverage point, units 1 and 35 are mild vertical outliers.

4.3 AIS dataset

The third application deals with the AIS dataset included in the `sn` package (Azzalini, 2021), and recently analyzed in the mixture of regression literature (Soffritti and Galimberti, 2011; Dang et al., 2017). It contains measurements of 102 male and 100 female athletes collected at the Australian Institute of Sport (Cook and Weisberg, 1994). We consider a subset of five variables, namely *lean body mass* (LBM), *body mass index* (BMI), *sum of skin folds* (SSF), *percentage body fat* (PBF) and *hemoglobin concentration* (y). The

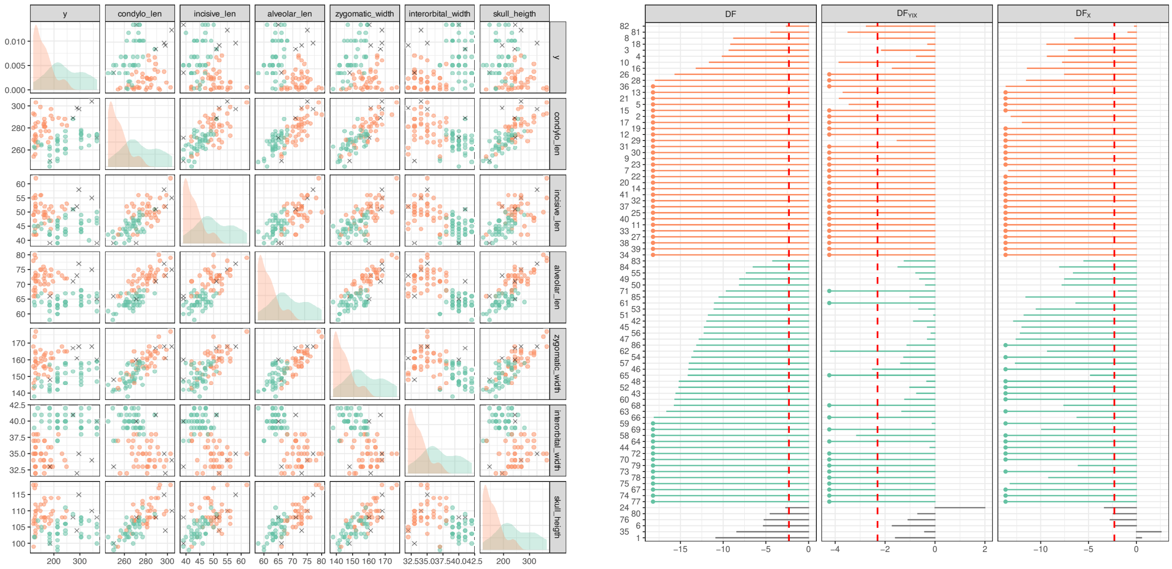

Figure 15: Female vole data. Pairs plot of the first optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM, where trimmed units are denoted by “×” (left panel) and silhouette plots displaying DF, $\text{DF}_{Y|X}$ and DF_X for all observations in the dataset (right panel).

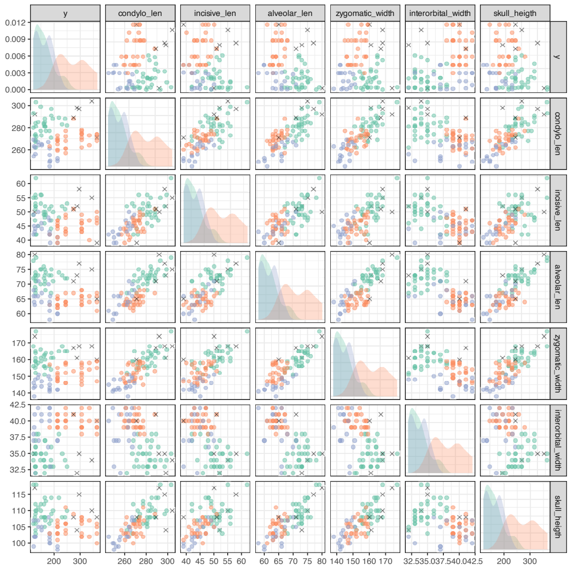

Figure 16: Female vole data. Pairs plot of the second optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM. Trimmed units are denoted by “×”.

Table 1: AIS data. Confusion matrix and Adjusted Rand Index for the top two optimal solutions obtained with the partition induced by the CWRMs (trimmed units have been a-posteriori assigned using the MAP rule).

(a) First optimal solution (ARI=0.758)

	1	2
Female	100	0
Male	13	89

(b) Second optimal solution (ARI=0.646)

	1	2	3
Female	1	72	27
Male	90	1	11

blood composition variable y acts as the response variable, while the remaining biometrical measures are the covariates. We aim at uncovering the cluster-wise linear structure in the data by exploring the space of solutions induced by the CWRM. The *gender* of the athlete will be the grouping variable thereafter.

The first step of the proposed monitoring procedure is displayed in Figure 17. It can be observed that, when no or little trimming is considered, the selection criterion in (8) suggests that a mixture of 3 components is preferred to capture the heterogeneity in the data. Conversely, for $\alpha \geq 0.035$ the number of clusters settles to $G = 2$. This behavior is due to the presence of some extreme observations and to the skewness showcased by some biometrical variables, as discussed in Azzalini and Capitanio (1999). After the disappearance of the third component, the line-plot patterns seem to stabilize and the proportion of doubtful assignments decreases, suggesting that a trimming level $\alpha = 0.035$ is adequate to achieve robustness in the solutions.

The second step, obtained conditioning on the selected trimming level, yields results shown in Figure 18. The first optimal solution is a cluster weighted model with 2 components and homoscedastic regression errors ($c_y = 1$). It is stable in a wide range of values on the (c_X, c_y) grid. The second and third optimal solutions appear when $G = 3$, indicating that an extra component may be as well included when modeling the AIS dataset. With respect to the estimated partition, the first optimal solution agrees with the true underlying male/female subdivision, correctly classifying 189 athletes by their gender: see Figure 19 and the confusion matrix reported in Table 1 (a). The three silhouette plots inform us that the outliers are all good leverage ones, say points scattered far from the clusters

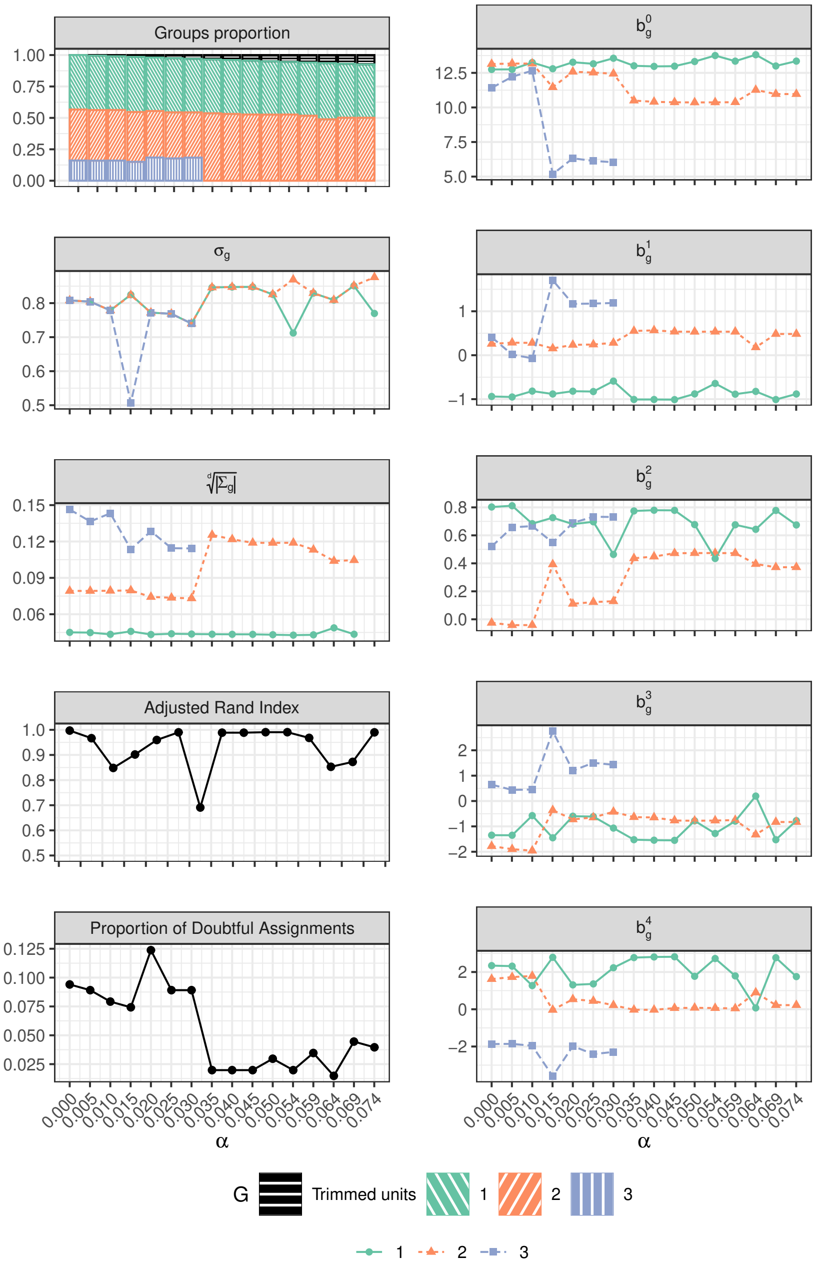

Figure 17: Monitoring tools in Step 1 for the AIS dataset. The considered metrics are monitored as a function of the trimming level α. The proportion of doubtful assignment is minimized for $\alpha = 0.035$, the estimated parameters are stable for contiguous values of $\alpha = 0.035$ and the unstable component also disappears when $\alpha \geq 0.035$.

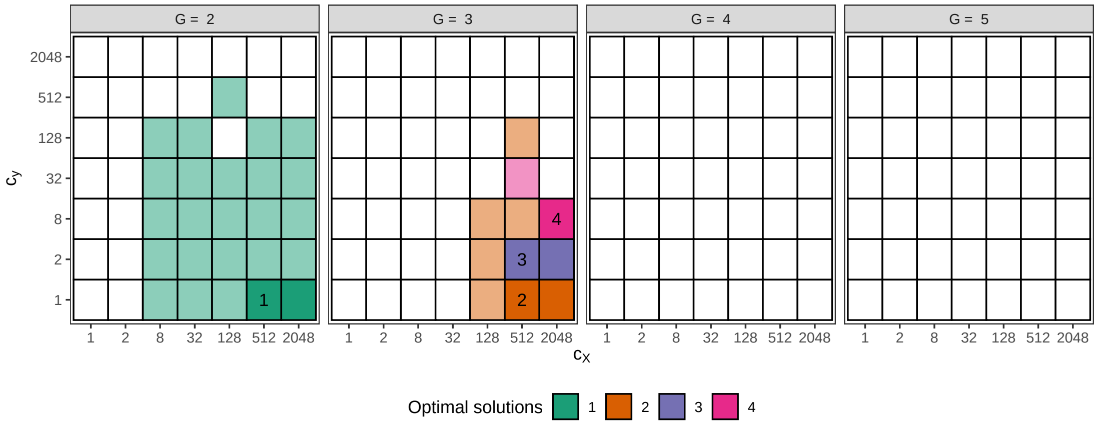

Figure 18: Monitoring tools in Step 2 for the AIS dataset. The optimal solutions are indicated by the cells with ordinal numbers 1, 2, 3 and 4 ($\alpha = 0.035$). Each solution is featured by one color, showing the range of cases in which it is best (darker opacity cells), and stable (lighter opacity cells), varying c_X (horizontal axis) and c_y (vertical axis) in $\mathcal{E}_0$.

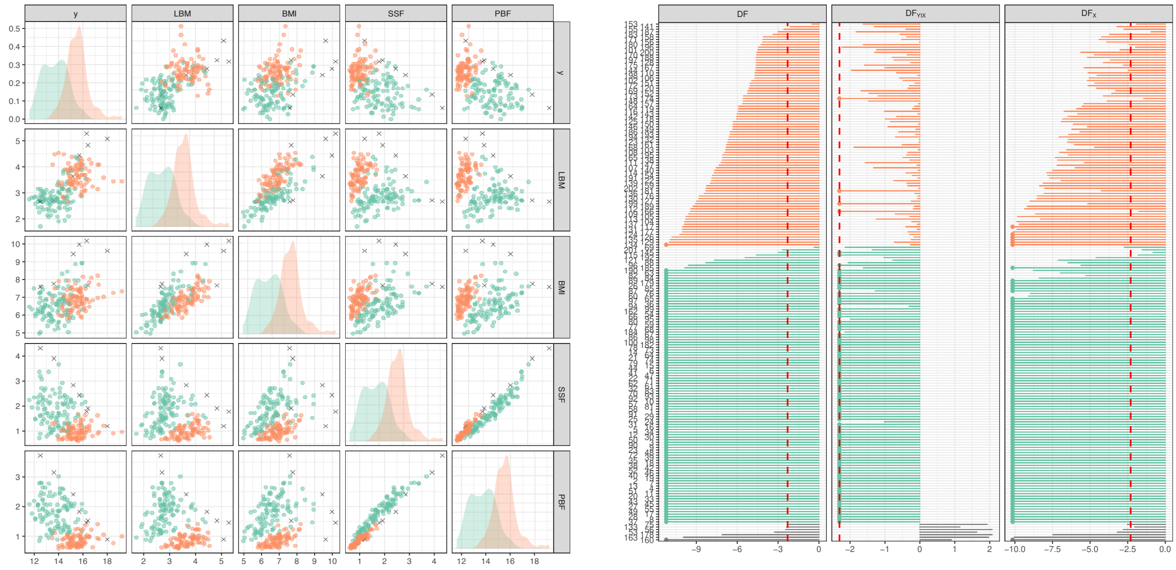

Figure 19: AIS data. Pairs plot of the first optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM, trimmed units are denoted by “×” (left panel) and silhouette plots displaying DF, $\mathrm{DF}_{Y|X}$ and DF_X for all observations in the dataset (right panel).

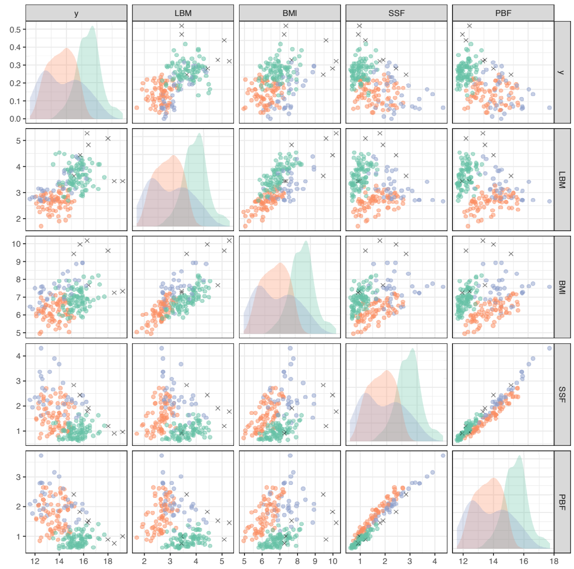

Figure 20: AIS data. Pairs plot of the second optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM. Trimmed units are denoted by "×".

in the covariates ($\text{DF}_{Y|X}$ over the threshold). Alike, the clustering induced by the second result uncovers the male/female separation, yet the latter class is further divided into two sub-populations, as reported in Figure 20 and in Table 1 (b). By looking at the pairs plot, we notice that the third group, smaller in size in comparison to the two main ones, captures women with a peculiar pattern in some of the biometrical variables, particularly in the sum of skin folds (SSF) and percentage body fat (PBF). Furthermore, this difference is also reflected in the relationship with the response variable y (hemoglobin concentration), for which the associated regression parameters are very different in the two clusters, with even opposite signs. While a clinical interpretation of this result is out of the scope of the present manuscript, it is apparent that sport scientists and sport analysts could fruitfully employ the proposed methodology.

5 Conclusions

It is now widely acknowledged in the literature that unsupervised classification problems have to be based on robust approaches, like those featured on impartial trimming - to protect from the harmful effects of outliers, and constrained estimation - to set a well-defined

Maximum Likelihood problem. To this extent, hyper-parameters tuning is required within the inferential process, and the debate about how to set them needs a fresh rethinking. A methodology for selecting the optimal level of trimming, in particular, is still missing, and it is the object of ongoing research. This is the most critical choice for the robust inferential procedure: masking issues can arise when adopting a lower than needed level of trimming, while excessive trimming can bias the estimate of the number of clusters and reduces the efficiency of the statistical method.

In this paper, we contributed to the literature by introducing graphical and computational tools to assist the practitioner in the delicate task of setting hyper-parameters in the estimation of robust cluster weighted models. The method relies on the combination of two exploratory steps. In the first monitoring step, the crucial assessment of the percentage of trimming is addressed. A wide exploration of the model space is made, and the graphical representation of some cluster-dependent metrics, coupled with the evolution of the estimated model parameters, allows to single out a small set of sensible options for the trimming proportion α. Afterward, for each plausible value of α, the whole space of solutions is explored, varying the hyper-parameters governing the heterogeneity on the covariates and the regression error terms, as well as the number of groups. The final output offers a set of optimal solutions, featured by the interval of hyper-parameter values in which their optimality, stability and validity hold.

An assessment of the role and extent of the outlying observations has been provided, introducing three new silhouette plots. The purpose is to understand the possible effects of the contaminated observations, referring to the clustering of the $\mathbf{X}$-covariate, and the local regression lines $Y|\mathbf{X}$, following the nature of the Cluster Weighted model.

The proposed monitoring techniques perform satisfactorily well in all the considered real data examples, providing valuable insight for the resulting model fitting. We have chosen datasets with covariates ranging from dimension 1 up to 6, and analyzed phenomena in the field of tourism, biology and sport analysis. On the light of the obtained results, the researcher is advised to resort to the monitoring strategy for an effective tuning, by conjugating it with any domain-specific knowledge that could be available. Such information can be very easily incorporated in the proposed procedure. To sum up, one of the main aspects of our methodology concerns the selection of the trimming parameter, related to the efficiency/robustness trade-off in finite samples.

Some possible directions for research concern strategies to reduce the computational burden of the proposed method. While our procedure greatly benefits from parallelization (a discussion on computing times for the case studies of Section 4 is reported in the Supplementary Materials), the exhaustive search for the best model among all the combinations of α, G, c_X, c_y may be further reduced considering conditional search and/or ad-hoc criteria for sensibly exploring the model space. In addition, to speed up ML estimation, initial values for the parameters can be inherited by contiguous solutions already obtained through the search. Furthermore, there may be interest in extending the monitoring process to more challenging scenarios, e.g., robust mixture of factor analyzers (García-Escudero et al., 2016), that were out of the scope of the present paper. Lastly, the proposed methodology can be further expanded to account for mixtures of cluster-weighted generalized linear models, with univariate as well as multivariate response, along the lines of Dang et al. (2017). All the aforementioned ideas are currently being explored and they will be the object of future research.

Acknowledgments

The authors wish to thank two anonymous reviewers and the Associate Editor for their helpful comments. Andrea Cappozzo's work is supported by the Research Programme: "Integration between study design and data analytics for generating credible evidence in the field of healthcare from heterogeneous sources of structured and unstructured data". Francesca Greselin's work is supported by Milano-Bicocca University Fund for Scientific Research, 2019-ATE-0076. Luis A. García-Escudero and Agustín Mayo Iscar work is supported by the Spanish Ministerio de Ciencia e Innovación, grant PID2021-128314NB-I00.

Supplementary materials

README: the supplemental files include a README describing the content of the supplementary materials

Appendix: the supplemental files include a further analysis of the tourism dataset, validation of optimal solutions via the Total Sum of Squares Decomposition and additional details on computing times.

R code: the supplemental files include an R script providing a short tutorial on how to use the CWRMmonitor package (github.com/AndreaCappozzo/CWRMmonitor) implementing the monitoring procedure described in the paper.

Rds file: the supplemental files include an .Rds object containing the CWRM models fitted on the AIS data to which apply the monitoring procedure, recovering the results reported in Section 4.3

References

Azzalini, A. (2021). *The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-t and the SUN (version 2.0.0).* Università di Padova, Italia.

Azzalini, A. and A. Capitanio (1999, aug). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 579–602.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Cappozzo, A., L. A. G. García Escudero, F. Greselin, and A. Mayo-Iscar (2021, jul). Parameter Choice, Stability and Validity for Robust Cluster Weighted Modeling. *Stats 4*(3), 602–615.

Cerioli, A., L. A. García-Escudero, A. Mayo-Iscar, and M. Riani (2018, apr). Finding the number of normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics 27*(2), 404–416.

Claeskens, G. and N. L. Hjort (2008, jan). *Model Selection and Model Averaging.* Cambridge University Press.

Cook, R. D. and S. Weisberg (1994). *An introduction to regression graphics*, Volume 405. John Wiley & Sons.

Dang, U. J., A. Punzo, P. D. McNicholas, S. Ingrassia, and R. P. Browne (2017, apr). Multivariate Response and Parsimony for Gaussian Cluster-Weighted Models. *Journal of Classification 34*(1), 4–34.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter 4*(1), 65–75.

Flury, B. (1997). *A first course in multivariate statistics.* Springer Science & Business Media.

Fritz, H., L. A. García-Escudero, and A. Mayo-Iscar (2012). tclust : An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software 47*(12), 1–26.

García-Escudero, L. A., A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar (2016, jul). The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics & Data Analysis 99*, 131–147.

García-Escudero, L. A., A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar (2017, mar). Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing 27*(2), 377–402.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2011, oct). Exploring the number of groups in robust model-based clustering. *Statistics and Computing 21*(4), 585–599.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2015). Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing 25*(3), 619–633.

García-Escudero, L. A., A. Mayo-Iscar, and M. Riani (2020, sep). Model-based clustering with determinant-and-shape constraint. *Statistics and Computing 30*(5), 1363–1380.

García-Escudero, L. A., A. Mayo-Iscar, and M. Riani (2022, feb). Constrained parsimonious model-based clustering. *Statistics and Computing 32*(1), 2.

Gershenfeld, N. (1997, jan). Nonlinear Inference and Cluster-Weighted Modeling. *Annals of the New York Academy of Sciences 808*(1 Nonlinear Sig), 18–24.

Hathaway, R. J. (1985, jun). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics 13*(2), 795–800.

Hennig, C. (2004, aug). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics 32*(4), 1313–1340.

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

Hennig, C. and T. F. Liao (2013, may). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(3), 309–369.

Huber, P. J. and E. M. Ronchetti (2009, jan). *Robust Statistics.* Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Ingrassia, S., S. C. Minotti, and A. Punzo (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis 71*, 159–182.

Ingrassia, S., S. C. Minotti, and G. Vittadini (2012, oct). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification 29*(3), 363–401.

Ingrassia, S. and A. Punzo (2020, jul). Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition. *Journal of Classification 37*(2), 526–547.

Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*(2), 159–179.

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Riani, M., A. C. Atkinson, A. Cerioli, and A. Corbellini (2019, apr). Efficient robust methods via monitoring for clustering and multivariate data analysis. *Pattern Recognition 88*, 246–260.

Rousseeuw, P. J. (1987a). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Rousseeuw, P. J. (1987b). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*(C), 53–65.

Rousseeuw, P. J. and K. V. Driessen (1999, aug). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*(3), 212–223.

Schwarz, G. (1978, mar). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Singh, K., J. M. Parelius, and R. Y. Liu (1999, jun). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics 27*(3), 783–858.

Soffritti, G. and G. Galimberti (2011, oct). Multivariate linear regression with non-normal errors: a solution based on mixture models. *Statistics and Computing 21*(4), 523–536.

Subedi, S., A. Punzo, S. Ingrassia, and P. D. McNicholas (2013, mar). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification 7*(1), 5–40.

Subedi, S., A. Punzo, S. Ingrassia, and P. D. McNicholas (2015, nov). Cluster-weighted t -factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications 24*(4), 623–649.

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

Torti, F., M. Riani, and G. Morelli (2021, jun). Semiautomatic robust regression clustering of international trade data. *Statistical Methods & Applications 0123456789*.

Van Aelst, S., X. (Steven) Wang, R. H. Zamar, and R. Zhu (2006, mar). Linear grouping using orthogonal regression. *Computational Statistics & Data Analysis 50*(5), 1287–1312.

von Luxburg, U., R. C. Williamson BobWilliamson, and I. Guyon (2012). Clustering: Science or Art? *JMLR: Workshop and Conference Proceedings 27*, 6579.