Monitoring tools for robust estimation of Cluster Weighted models

Strumenti di monitoring per la stima robusta del modello Cluster Weighted

Andrea Cappozzo and Francesca Greselin

Abstract In a robust approach to model fitting for the cluster weighted model, many choices are to be made by the statistician: specifying the shape of the clusters in the explanatory variables, assuming (or not) equal variance for the errors in the regression lines, and setting hyper-parameter values for the robust estimation to be protected from outliers and contamination. The most delicate hyper-parameter to specify is perhaps the percentage of trimming, or the amount of data to be excluded from the estimate, to ensure reliable inference. In this work we introduce diagnostic tools to help the professional, or the scientist who needs to group the data, to make an educated choice about this hyper-parameter, after a first exploration of the resulting model space.

Abstract *Nella stima robusta di un cluster weighted model, lo statistico deve fare molte scelte: specificare la forma dei cluster nelle variabili esplicative, assumere (o meno) varianza uguale per gli errori nelle linee di regressione e impostare i valori degli iper-parametri per la stima robusta, per evitare la distorsione generata da valori anomali e contaminazione. L'iper-parametro più delicato da specificare è la percentuale di trimming, ovvero la quantità di dati da escludere nella stima per garantirne l'affidabilità. In questo lavoro introduciamo specifici strumenti diagnostici per aiutare il professionista, o lo scienziato che ha bisogno di classificare i dati, a compiere una scelta ragionata a riguardo di tale iper-parametro, anche in base ad una prima esplorazione dello spazio delle soluzioni.*

Key words: Cluster-weighted modeling, Outliers, Trimmed BIC, Eigenvalue constraint, Monitoring

---

Andrea Cappozzo
University of Milano Bicocca, Department of Statistics and Quantitative Methods e-mail: a.cappozzo@unimib.it

Francesca Greselin
University of Milano Bicocca, Department of Statistics and Quantitative Methods e-mail: francesca.greselin@unimib.it

1 Introduction

Clustering is a well known ill-posed problem, where the number of groups, their shape, and their parameters depend, in general, on a multiplicity of subjective choices [4]. Generally, selecting the unknown number of groups G defines the most challenging task. The most popular method adopted in model-based clustering for tackling the problem is based on penalized likelihoods, but the presence of data contamination and outliers could severely undermine such powerful criteria. In addition, when it comes to cluster weighted modeling, many other choices need to be performed: whether to constrain the cluster shapes in the explanatory variables, to impose or not equal variances in the regression errors, how to set hyper-parameters for discarding spurious solutions and how to protect against outliers.

We introduce here a semiautomatic procedure for selecting a reduced set of solutions, extending to the cluster weighted model the methodology developed in [1] for the Gaussian mixture models. Such an extension is far from being straightforward. A new penalized likelihood criterion will be devised to account for the constraint imposed on the regression term and on the covariates, varying trimming levels and number of cluster. The remainder of the article proceeds as follows. Section 2 provides a brief overview of the Cluster Weighted Model (CWM) and its robust estimation. Section 3 reports the two-stage monitoring strategy, based on (i) a first exploration of the model space with a dedicated information criterion and (ii) usage of new "trimming-based" tools, tailored for CWM. Section 4 concludes the paper by showcasing the validity of our proposal within a controlled experiment.

2 The cluster weighted model

Let $\mathbf{X}$ be a vector of *explanatory* variables with values in $\mathbb{R}^d$, and let Y be a *response* or *outcome* variable, with values in $\mathbb{R}$. Suppose that the regression of Y on $\mathbf{X}$ varies across the G levels (group or clusters) of a categorical latent variable. The CWM, introduced in [3], decomposes the joint p.d.f. of $(\mathbf{X}, Y)$ in each component of the mixture as the product of the marginal and the conditional distributions as follows

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^{G} \pi_g p(y|\mathbf{x}; \xi_g) p(\mathbf{x}; \psi_g). \tag{1}$$

In the cluster-weighted approach the marginal distribution of $\mathbf{X}$ and the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ may have different scatter structures in each group. In this work, we focus on the *linear Gaussian CWM*:

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^{G} \pi_g \phi(y; \mathbf{b}_g' \mathbf{x} + b_g^0, \sigma_g) \phi_d(\mathbf{x}; \mu_g, \Sigma_g), \tag{2}$$

where $\phi_d(\cdot; \mu_g, \Sigma_g)$ denotes the density of the d-variate Gaussian distribution with mean vector μ_g and covariance matrix Σ_g, and Y is related to $\mathbf{X}$ by a linear model,

that is, $Y = \mathbf{b}_g'\mathbf{x} + b_g^0 + \varepsilon_g$ with $\varepsilon_g \sim N(0, \sigma_g^2)$, $\mathbf{b}_g \in \mathbb{R}^d$, $b_g^0 \in \mathbb{R}$, $\forall g = 1, \ldots, G$. Given a sample of n i.i.d. pairs drawn from $(Y, \mathbf{X})$, the ML estimation of the linear Gaussian CWM is based on the maximization of the following log-likelihood function

$$\mathscr{L} = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \pi_g \phi(y_i; \mathbf{b}_g'\mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \right]. \tag{3}$$

Unfortunately, ML inference on models based on normal assumptions suffers from lack of robustness. Another important concern is the unboundedness of the likelihood function to be maximized. To overcome these issues, a robust version of the CWM has been presented in the literature by considering impartial trimming and constrained estimation of the scatter variances [2]. The robust approach to CWM (CWRM) is based on the maximization of the *trimmed* log-likelihood function [6]

$$\mathscr{L}_{trimmed} = \sum_{i=1}^{n} z(\mathbf{x}_i, y_i) \log \left[\sum_{g=1}^{G} \pi_g \phi(y_i; \mathbf{b}_g'\mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \right], \tag{4}$$

where $z(\cdot, \cdot)$ is a 0-1 trimming indicator function that tells us 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 unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i, y_i) = \lfloor n(1-\alpha) \rfloor$, and the parameter α denotes the trimming level.

We introduce two constraints on the maximization in (4). The first one concerns the set of eigenvalues $\{\lambda_l(\Sigma_g)\}_{l=1,\ldots,d}$ of the scatter matrices Σ_g by imposing

$$\lambda_{l_1}(\Sigma_{g_1}) \leq c_X \lambda_{l_2}(\Sigma_{g_2}) \qquad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G. \tag{5}$$

The second constraint refers to the variances σ_g^2 of the regression error terms, by requiring

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G. \tag{6}$$

The constants c_X and c_ε, in (5) and (6) are finite (not necessarily equal) real numbers, such that $c_X \geq 1$ and $c_\varepsilon \geq 1$. They automatically guarantee that we are avoiding the $|\Sigma_g| \to 0$ and $\sigma_g^2 \to 0$ degenerate cases.

3 Monitoring the setting of CWM hyper-parameters

We propose a semi-automatic approach to provide adaptive values for the hyper-parameter α involved in the robust fitting of CWMs. By building upon previous work developed for robust clustering [7], a two-stage monitoring procedure is devised. First off, for each trimming level $\alpha \in \{0, \ldots, \alpha_{MAX}\}$ ($\alpha_{MAX} = 0.15$ in the analysis of Section 3) the most appropriate model, varying G, c_X and c_ε, is determined. Secondly, exploratory tools are employed to compare solutions for different levels of α, providing aid in assessing the true contamination level present in a dataset.

In details, in the first phase a constrained estimation criterion is devised for comparing models when α is kept fixed. As in the well known Bayesian Information Criterion ($BIC = -2\mathscr{L} + \nu_G$) and along the lines of [1], the dedicated penalty term ν_G depends on the number of free parameters in the model:

$$\begin{aligned}\nu_G = &\{(G-1) + Gp + G(p+1) + \\ &1 + ((Gp-1) + Gp(p-1)/2)(1-1/c_X) + \\ &1 + (G-1)(1-1/c_\varepsilon)\}\log(\lceil n(1-\alpha)\rceil).\end{aligned} \tag{7}$$

The first three terms in (7) respectively refer to the $(G-1)$ mixture weights, the Gp cluster means of the covariates, and the $G(p+1)$ beta coefficients for the regression $\mathbf{b}_g + b_g^0$, $g = 1,\ldots,G$. The second group of terms is related to the modelling of $\mathbf{X}$, where we have 1 free eigenvalue, $Gp-1$ constrained eigenvalues and $Gp(p-1)/2$ rotation matrices for Σ_g. Except the first one, all terms are multiplied by $(1-1/c_X)$ to take into account the enforced constrained estimation. Lastly, in the third line of (7), the part relative to modelling $Y|\mathbf{X}$ induces one free σ_g^2 and $G-1$ constrained σ_g^2. Notice that, while in [1] the authors distinguish between rotation and eigenvalue parameters multiplying only the latter by the factor $(1-1/c_X)$, we opt here for penalizing all the variance parameters, as rotation looses its meaning for $c_X \to 1$.

In the second phase, we extend the monitoring introduced in [7], where a plot of the Adjusted Rand Index (ARI) between consecutive cluster allocations for a grid of α values is proposed, to determine an optimum trimming level. This tool can be effective in detecting noise in the form of bridges, where only a correct level of trimming uncovers the true underlying structure. In the case of scattered noise, however, the clustering structure could evolve very smoothly from an initial partition, obtained without trimming, and a pretty different final partition, yielding an ARI pattern between consecutive solutions with no apparent abrupt change. Motivated by this argument, we widen the monitoring tools accompanying the ARI plot with regression coefficients and mixture weights paths, to highlight specific CWM features. Further, we are interested in monitoring the CWM validation measure based on the decomposition of the total sum of squares $TSS = BSS + RWSS + EWSS$ [5]. BSS is the (soft) between-group sum of squares, while $EWSS$ is the portion of the (soft) within-group sum of squares WSS explained by the model, thanks to the covariate, and $RWSS$ is the residual portion of WSS. In terms of cluster validation, therefore, BSS can be seen as a separation measure on the Y-axis, and WSS can be seen as a cluster compactness measure. To overcome the non-identifiability issue due to invariance of mixture components, a relabeling strategy based on data depth [8] is adopted. In this way, component-dependent metrics, estimated varying trimming levels, are directly comparable: an application is provided in the next section.

4 Illustrative experiment

A dataset with 180 genuine samples is generated according to (2) with the following parameters:

$$\pi = (0.5, 0.5)', \quad \mu_1 = (2,2)', \quad \mu_2 = (5,5)', \quad \Sigma_1 = \Sigma_2 = I_2$$
$$b_1^0 = 30, \quad b_2^0 = 50, \quad \mathbf{b}_1 = (-1,-1)' \quad \mathbf{b}_2 = (10,10)' \quad \sigma_1^2 = \sigma_2^2 = 1, \quad (8)$$

in addition, 20 uniformly distributed outliers are appended to the uncontaminated observations, resulting in a total of $n = 200$ data units with a true contamination level equal to 0.1. In the first phase, models with $c_X, c_\varepsilon \in \{1,4,16,64\}$ and $G = \{2,3,4\}$ are fitted to the considered dataset: Table 1 reports the best model, selected by minimizing the information criterion introduced in the previous section (denoted with TBIC in the table), conditioning on the trimming value α. Notice that, whenever α is set below the true contamination rate, some erroneous solutions are preferred: G is selected to be greater than 2, with spurious groups fitting the portion of untrimmed noise. The second phase of our procedure encompasses the plots reported in Figure

Table 1 Best models, as a function of G, c_X and c_ε, selected via TBIC minimization conditioning on the trimming value α (only a subset of the entire α grid considered in the experiment is reported) for the first phase of the monitoring procedure.

α	0.00	0.03	0.06	0.09	0.10	0.11	0.12	0.13	0.14	0.15
G	4	4	4	3	2	2	2	2	2	2
c_X	4	4	64	4	4	4	4	4	4	4
c_ε	64	64	64	64	1	1	1	1	1	1
TBIC	2801.82	2363.97	2157.08	1998.03	1940.30	1885.77	1848.96	1812.12	1776.44	1741.11

1: by monitoring the changes in mixing proportions, regression parameters, total sum of squares and ARI between consecutive cluster allocations the analyst may reasonably observe how the solutions stabilize as soon as α is higher than the true contamination level 0.1. Particularly, given the ARI almost constant high value (bottom right plot), this metric alone would not have been sufficient to properly address the complexity of the problem.

5 Conclusions

The present article provides a two-stage monitoring procedure for aiding in the hyper-parameters selection when fitting robust CWM to contaminated datasets. We opted for providing the user with sensible information to make the required tuning decisions: ultimately an optimal tuning of model parameters should also depend on knowledge about the subject matter background and the aim of clustering. The procedure takes over and extends the state-of-the-art methods proposed for robust clustering by including a wider range and component-dependent metrics, essential for thoroughly understanding the true data generating mechanism.

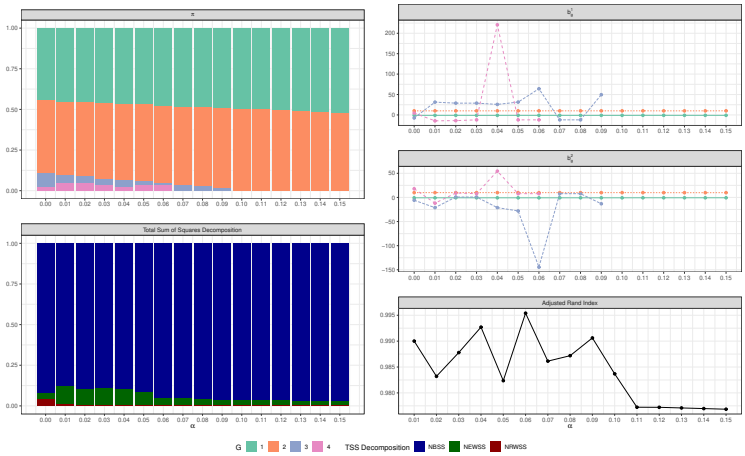

Fig. 1 Monitoring the mixing proportions (top left plot), regression parameters (top right plots), total sum of squares decomposition (bottom left plot) and ARI between consecutive cluster allocations (bottom right plot, please be aware of the Y axis range) as a function of the trimming proportion α.

References

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

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

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

[4] C. Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.

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

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

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

[8] K. Singh, J. M. Parelius, and R. Y. Liu. 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, jun 1999.