Monitoring Tools in Robust CWM for the Analysis of Crime Data

Andrea Cappozzo[1], Luis Angel García-Escudero[2], Francesca Greselin[3(✉)], and Agustín Mayo-Iscar[2]

[1] Department of Mathematics, Politecnico di Milano, Milan, Italy
andrea.cappozzo@polimi.it
[2] Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain
{lagarcia,agustin.mayo.iscar}@uva.es
[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
francesca.greselin@unimib.it

Abstract. Robust inference for the Cluster Weighted Model requires the specification of a few hyper-parameters. Their role is crucial for increasing the quality of the estimators, while arbitrary decisions about their value could severely hamper inferential results. To guide the user in the delicate choice of such parameters, a monitoring approach has been introduced in the recent literature, yielding an adaptive method. The approach is here exemplified, via the analysis of a dataset on the effect of punishment regimes on crime rates.

1 Introduction and Notation

The purpose of the present paper is to demonstrate how to perform hyper-parameters selection when applying Cluster Weighted Modelling (CWM) to a real dataset. In detail, through the employment of graphical tools, we propose a two-stage monitoring procedure to sequentially detect the number of potential outliers and the thresholds used in constrained estimation.

The contribution advances the studies on the semi-automation of clustering techniques, which is a relevant topic in statistics and in real statistical applications. Our proposal has been developed along the lines of Cerioli et al. (2018) and Torti et al. (2021).

We begin by introducing very briefly the notation, then provide the main ideas of the monitoring methodology, and in Sect. 2 we present and discuss an application to Crime data. Final remarks end the paper in Sect. 3.

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. The linear gaussian CWM, introduced by Gershenfeld (1997), decomposes the

L. A. García-Escudero et al. (Eds.): SMPS 2022, AISC 1433, pp. 65–72, 2023.
https://doi.org/10.1007/978-3-031-15509-3_9

joint distribution of $(\mathbf{X}, Y)$, in each mixture component, as the product of the marginal and the conditional distributions, as follows:

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

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 (1), that is, $Y = \boldsymbol{b}_g' \mathbf{x} + b_g^0 + \varepsilon_g$ with $\varepsilon_g \sim N(0, \sigma_g^2)$, $\boldsymbol{b}_g \in \mathbb{R}^d$, $b_g^0 \in \mathbb{R}$, $\sigma_g^2 \in \mathbb{R}^+$, $\forall g = 1, \ldots, G$. Unfortunately, Maximum Likelihood inference on models based on normal assumptions suffers from two major drawbacks: (i) the likelihood is unbounded over the parameter space, hence its maximization is in an ill-posed mathematical problem; (ii) the resulting inference is strongly affected by outliers (see, e.g., Huber and Ronchetti 2009). To overcome both issues, García-Escudero et al. (2017) introduced the Cluster Weighted Robust Model (CWRM), where a fixed fraction α of the less plausible observations is trimmed out and the estimation of the scatter matrices and the regression errors is constrained to ensure robust inference. The first constraint is applied to the set of eigenvalues $\{\lambda_l(\boldsymbol{\Sigma}_g)\}_{l=1,\ldots,d}$ of the scatter matrices $\boldsymbol{\Sigma}_g$ by requiring

$$\lambda_{l_1}(\boldsymbol{\Sigma}_{g_1}) \leq c_X \lambda_{l_2}(\boldsymbol{\Sigma}_{g_2}) \qquad \text{for every } 1 \leq l_1 \neq l_2 \leq d \text{ and } 1 \leq g_1 \neq g_2 \leq G. \tag{2}$$

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

$$\sigma_{g_1}^2 \leq c_y \sigma_{g_2}^2 \qquad \text{for every } 1 \leq g_1 \neq g_2 \leq G. \tag{3}$$

The constants $c_X, c_y \geq 1$ prevent degenerate cases with $|\boldsymbol{\Sigma}_g| \to 0$ and $\sigma_g^2 \to 0$ leading to an unbounded likelihood or non interesting spurious solutions. Therefore, the percentage of trimmed data and the threshold for eigenvalues ratio play an important role and should be carefully set.

We will exemplify how the monitoring tools introduced in Cappozzo et al. (2021) can be applied to real data. A first monitoring step is devoted to screen the space of solutions for CWRM, in view of making an informed choice for the trimming level α, which is the most crucial parameter. Metrics such as the group proportion, the total sum of squares decomposition, the regression slopes and standard deviations, the cluster volumes, and the Adjusted Rand Index (ARI) between consecutive cluster allocations are monitored when varying α, to uncover the most sensible trimming level to be employed.

Afterwards, the second monitoring step screens the space of solutions $\mathscr{E}_0$ 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 $\mathscr{O}$ of "optimal" solutions, qualified by two features: their optimality in terms of a CWRM-specific BIC criterion (Cappozzo et al. 2021),

and their stability across hyper-parameter values. Stability of solution A means that, varying the values of the constraints, the estimation yields a partition B pretty close to A (the ARI between A and B is greater than a threshold η).

Finally, to explore the quality of the clustering obtained in the optimal solutions and to uncover the nature of the outliers, silhouette plots (Rousseeuw 1987) can be employed. Silhouette plots have been introduced for representing the quality of the clustering solution, and may be defined in the spirit of discriminant factors introduced in García-Escudero et al. (2011). Specific discriminant factors, tailored for the CWM characterization in (1) should be considered here. The first discriminant factor $DF(i)$ assesses the strength of the assignment, or the strength of the trimming decision for unit i in the joint modeling expressed by the CWM. On the other side, $DF_{Y|X}(i)$ and $DF_X(i)$ break down the overall mixture density in the contribution of the G regression hyperlanes and the component-wise random covariates, respectively.

2 Crime Dataset

The dataset originates from a study on the effect of punishment regimes on crime rates. We analyse aggregate data on 47 US states taken place in 1960 illustrated by Ehrlich (1973), available in the `MASS R` package. The crime rate, measured as the number of offenses per 100.000 population, is the response variable Y.

The goal is to infer whether the structure of dependence among covariates differently affects the crime rate Y depending on the geographical area.

Available predictors $\mathbf{X}$ for each of the 47 states are the following: *percentage of males aged 14–24*, *mean years of schooling*, *police expenditure*, *labour force participation rate*, *number of males per 1000 females*, *state population*, *number of non-whites per 1000 people*, *unemployment rate of urban males 14–24*, *unemployment rate of urban males 35–39*, *gross domestic product per head*, *income inequality*, *probability of imprisonment*, and *average time served in state prisons*. Finally, the indicator variable denoting the *16 Southern states* will be considered as the grouping variable hereafter. To this extent, robust CWM is applied on $(\mathbf{X}, Y)$ to uncover geographical grouping. The high number of variables involved in the estimation and the small sample size represent a challenge for the discriminating task.

The first step of our monitoring tools provides the outcome displayed in Fig. 1, where α takes values on a grid from 0 to 0.255. We opted for setting in advance $G = 2$ the number of geographical areas we would like to uncover. This choice is in line with the monitoring philosophy, for which any domain-related knowledge that may guide hyper-parameter selection shall be included in the analysis.

Fig. 1. Crime data: monitoring tools obtained in Step 1. Groups proportion (black bars denote the trimmed units), total sum of squares decomposition, regression coefficients, standard deviations, cluster volumes, ARI between consecutive cluster allocations are shown as a function of the trimming level α. G is kept fixed and equal to 2

We see that the choice of $\alpha = 0.064$ stabilizes the variance of the regression errors σ_1 and de-inflate the determinant of the scatter matrix $|\Sigma_1|^{1/d}$ in the first cluster (represented in green in Fig. 1), aligning them to the order of magnitude of the analogous quantities in the second cluster. Therefore, an estimation based on 44 observations seems sufficient to assure robustness without sacrificing efficiency.

In the second step, the monitoring procedure assesses the validity and stability of the solutions. Figure 2 reports the results, while the pairs plots for the first and second optimal solutions, based on a subset of explanatory variables, are displayed in Figs. 3 and 4, respectively. The first optimal solution remains best for $c_X = 10$ and for c_Y ranging from 1 to 8, while when $c_Y = 10$ the second optimal solution appears. On the one hand, the first solution shows a wide stability varying the values of the constraints; on the other hand, the second solution offers the highest classification accuracy. Indeed, the latter partition possesses only 2 misclassified units, maintained even after assigning the three trimmed units using the MAP rule: the proposed monitoring procedure provides a highly accurate classification for this dataset. In this regard, the best result present in the literature has been achieved by means of a competing method, introduced in Subedi et al. (2015), in which a variant of the CWM was developed considering a mixture of t-factor analyzers for modeling the marginal distribution of $\mathbf{X}$ and a t-distribution for the regression error. Such a model certainly has nice features in terms of explainability and parsimony, it nonetheless showcases 4 misclassified units when applied to the US crime dataset. The t distribution is known to be a very good option in presence of mild outliers, but hard trimming seems to achieve slightly better performance in this context. Note that, as already reported in Subedi et al. (2015), mixture of regression (DeSarbo and Cron 1988; De Veaux 1989) and mixtures of regression with concomitants (Dayton and Macready 1988) do not show good clustering performance whenever the distribution of the covariates plays a role on the cluster structure of the data (see the pairs plot in Figs. 3 and 4): such approaches are, by construction, unable to capture it.

Lastly, from the right panel in Fig. 3, the silhouette plots tell us that observations 4 and 29 are bad leverage points, having low values of the discriminant factors $DF_{Y|X}$ and DF_X respectively assessing the strength of the assignment/trimming for each unit in relation to the regression lines and the covariates. Without trimming such observations, inferential results on the regression parameters would have been biased. Observation 11 is instead a non-outlying point 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 (revealed by the high negative value for DF).

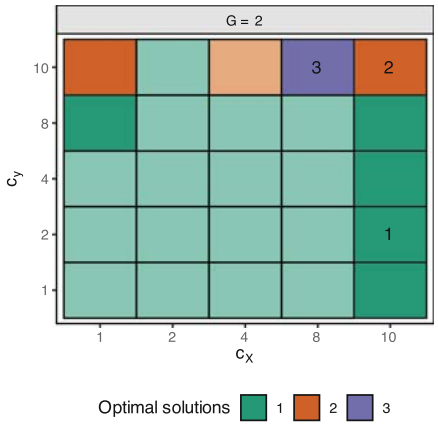

Fig. 2. Crime data. Step 2: monitoring the optimal solutions, indicated by the cells with ordinal numbers 1, 2, 3 and 4 ($\alpha = 0.064$). Each solution is featured by one color, showing the range of cases in which it is best (darker opacity cells), and stable (lighter opacity cells), varying c_X (horizontal axis) and c_Y (vertical axis) in $\mathscr{C}_0$. G is fixed and equal to 2

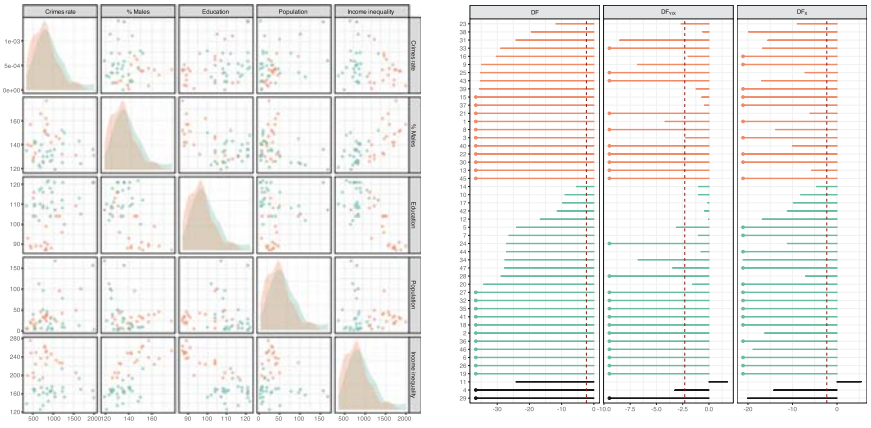

Fig. 3. Crime data. Pairs plot of the first optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM, trimmed units are denoted by × (left panel) and silhouette plots displaying $DF(i)$, $DF_{Y|X}(i)$ and $DF_X(i)$ for observation i in the dataset (right panel)

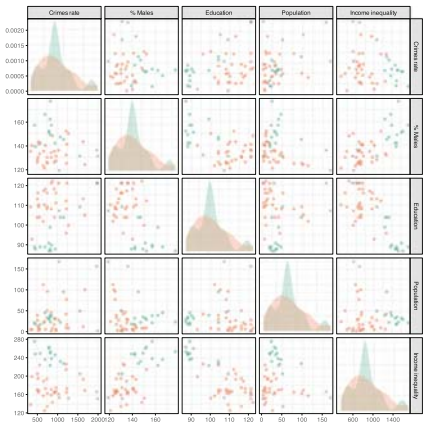

Fig. 4. Crime data. Pairs plot of the second optimal solution obtained in Step 2, different colors denote the partition induced by the CWRM. Trimmed units are denoted by $\times$

3 Conclusions

Robustness is the property of statistical methods to infer the generating process that originates the main body of the data in presence of contamination. A wide class of robust procedures for clustering is featured by constrained estimation and impartial trimming, for which specific hyper-parameters are introduced. In this paper, we presented an application of a recent contribution to the literature for the case of cluster-wise regression, where a monitoring approach is proposed. We have shown how the graphical and computational tools are able to assist the practitioner in the delicate task of setting hyper-parameters in the estimation of robust cluster weighted models, to analyze the effect of punishment regimes on crime rates.

The method relies on the combination of two exploratory steps. Sensible options for the trimming proportion α are identified in the first monitoring step. Afterwards, in the second monitoring step, the whole space of solutions is explored, varying the hyper-parameters governing the heterogeneity in the covariates and the regression error terms, as well as the number of groups.

In the analysis of the crime dataset, our semiautomatic procedure yields two final solutions, qualified by the interval of hyper-parameters values in which their optimality, stability and validity hold true. The first solution shows a wide stability; on the other hand, the second solution offers the highest classification accuracy for this dataset. New silhouette plots reveal the nature and the extent of the outlying observations, distinguishing between outliers with respect to the clustering of the covariate $\mathbf{X}$, and the local regression lines Y, following the nature of the Cluster Weighted model.

Further research can be devoted to reducing the computational burden of the proposed methodology, and extending it to other robust clustering models.

References

Cappozzo, A., García-Escudero, L.A., Greselin, F., Mayo-Iscar, A.: Parameter choice, stability and validity for robust cluster weighted modeling. Stats **4**(3), 602–615 (2021)

Cerioli, A., Riani, M., Atkinson, A.C., Corbellini, A.: The power of monitoring: how to make the most of a contaminated multivariate sample. Stat. Meth. Appl. **27**(4), 661–666 (2018)

Dayton, C.M., Macready, G.B.: Concomitant-variable latent-class models. J. Am. Stat. Assoc. **83**(401), 173–178 (1988)

DeSarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. J. Classif. **5**(2), 249–282 (1988)

De Veaux, R.D.: Mixtures of linear regressions. Comput. Stat. Data Anal. **8**(3), 227–245 (1989)

Ehrlich, I.: Participation in illegitimate activities: a theoretical and empirical investigation. J. Polit. Econ. **81**(3), 521–565 (1973)

García-Escudero, L.A., Gordaliza, A., Greselin, F., Ingrassia, S., Mayo-Iscar, A.: Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. Stat. Comput. **27**(2), 377–402 (2016). https://doi.org/10.1007/s11222-016-9628-3

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

Gershenfeld, N.: Nonlinear inference and cluster-weighted modeling. Ann. N. Y. Acad. Sci. **808**(1 Nonlinear Signal and Image Analysis), 18–24 (1997)

Huber, P.J., Ronchetti, E.M.: Robust Statistics. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, USA (2009)

Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(C), 53–65 (1987)

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P.D.: Cluster-weighted *t*-factor analyzers for robust model-based clustering and dimension reduction. Stat. Meth. Appl. **24**(4), 623–649 (2015)

Torti, F., Riani, M., Morelli, G.: Semiautomatic robust regression clustering of international trade data. Stat. Meth. Appl. **30**(3), 863–894 (2021). https://doi.org/10.1007/s10260-021-00569-3