EXPLORING SOLUTIONS VIA MONITORING FOR CLUSTER WEIGHTED ROBUST MODELS

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

[1] Department of Mathematics, Politecnico di Milano, (e-mail: andrea.cappozzo@polimi.it)

[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, (e-mail: lagarcia@uva.es, agustin.mayo.iscar@uva.es)

[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: francesca.greselin@unimib.it)

ABSTRACT: Depending on the selected hyper-parameters, cluster weighted modeling may produce a set of diverse solutions. Particularly, the user can manually specify the number of mixture components, the degree of heteroscedasticity of the clusters in the explanatory variables and of the errors around the regression lines. In addition, when performing robust inference, the level of impartial trimming enforced in the estimation needs to be selected. This flexibility gives rise to a variety of “legitimate” solutions. To mitigate the problem of model selection, we propose a two stage monitoring procedure to identify a set of “good models”. An application to the benchmark tone perception data showcases the benefits of the approach.

KEYWORDS: Cluster-weighted modeling, Outliers, Trimmed BIC, Eigenvalue constraint, Monitoring, Constrained estimation, Model-based clustering.

1 Introduction and model preliminaries

Assume to have observed a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ of n i.i.d. samples, where the regression on Y varies across G groups, based on a vector $\mathbf{X}$ of explanatory variables with values in $\mathbb{R}^d$. Within this framework, the Gaussian Cluster Weighted Robust Model (García-Escudero *et al.*, 2017) is based on the constrained maximization of the *trimmed* log-likelihood:

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

subject to: $\lambda_{l_1}(\Sigma_{g_1}) \leq c_X \lambda_{l_2}(\Sigma_{g_2})$ for every $1 \leq l_1 \neq l_2 \leq d$, $1 \leq g_1 \neq g_2 \leq G$ and $\sigma^2_{g_1} \leq c_y \sigma^2_{g_2}$ for every $1 \leq g_1 \neq g_2 \leq G$. The 0-1 trimming indicator function $z(\cdot,\cdot)$ tells us whether observation $(\mathbf{x}_i, y_i)$ is trimmed off, with trimming level α% of observations being left unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i, y_i) = \lfloor n(1-\alpha) \rfloor$. The set $\{\lambda_l(\Sigma_g)\}_{l=1,\ldots,d}$ denotes the eigenvalues of the scatter matrices Σ_g and the constants c_X and c_y are respectively finite real numbers such that $c_X \geq 1$ and $c_y \geq 1$.

2 Tone perception data application

The tone perception dataset (De Veaux, 1989) is employed as a case study to illustrate the proposed two-step monitoring procedure. In the first step, dedicated graphical and exploratory tools are employed for determining one or more plausible values for the trimming level α. Specifically, group proportion (black bars denote the trimmed units), total sum of squares decomposition (Ingrassia & Punzo, 2020), regression coefficients, standard deviations, cluster volumes and Adjusted Rand Index (ARI) between consecutive cluster allocations are monitored within a grid of αs, as reported in Figure 1. For each trimming level, the best model is selected according to a novel penalized likelihood criterion tailored for the CWRM framework, building upon the proposal developed in Cerioli *et al.*, 2018 for Gaussian mixtures. As it is clearly visible for the plots in Figure 1, model parameters stabilize as soon as α is set higher than 0.08, a value sufficient to trim off the level of contamination known to be present in this dataset (García-Escudero *et al.*, 2017).

In the second stage, conditioning on the α selected in the previous step, solutions stability and validity are fully investigated varying hyper-parameters in $\mathcal{E}_0 = \{(G, c_X, c_y) : G = 1,\ldots,4, c_X, c_y = 2^1,\ldots,2^5\}$, as reported in Figure 2. Darker and lighter opacity cells respectively indicate the sets of $\mathcal{B}_t$ best and $\mathcal{S}_t$ stable solutions, for each optimal solution $t, t = 1\ldots,4$, where optimality is in the sense of the penalized criterion. The former set includes solutions ARI-similar to the optimal and not worse than the next optimal, while the latter encompasses all solutions ARI-similar to the optimal, such that $\mathcal{B}_t \subseteq \mathcal{S}_t$. In this example, solutions are assumed to be ARI-similar if the ARI between the estimated partitions is higher than 0.7. It is interesting to notice that the CWRM favors models with higher number of clusters with respect to the accepted truth of $G = 2$ (fourth optimal solution, stable in the entire grid of c_X and c_y). The reason being that, contrarily to the standard mixture of regression, the CWRM treats the covariate as random, thus allowing the learning of group-wise different distributions in the explanatory variable (Figure 3).

We have demonstrated the adequacy of our monitoring procedure in aiding

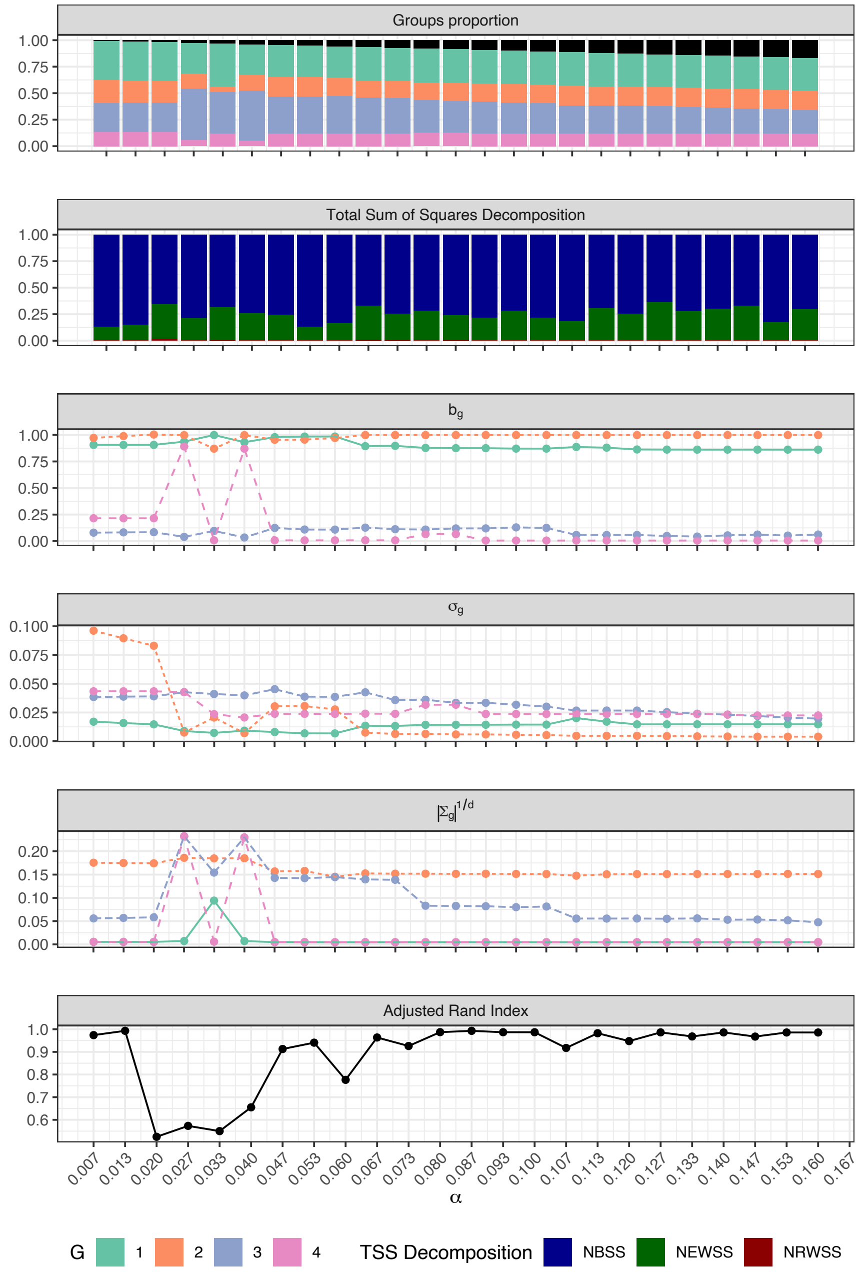

Figure 1: Step 1, monitoring the choice of a plausible trimming level α, tone perception data.

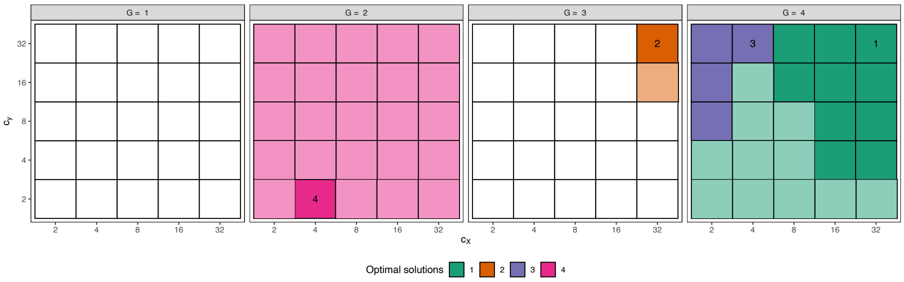

Figure 2: Step 2: monitoring optimal solutions, in terms of validity and stability. Trimming level $\alpha = 0.08$, tone perception data.

practitioners in the hyper-parameters selection when fitting CWRM. Furthermore, by exploring the space of solutions a deeper understanding of the data structure is achieved, uncovering sometimes unexpected yet valuable results.

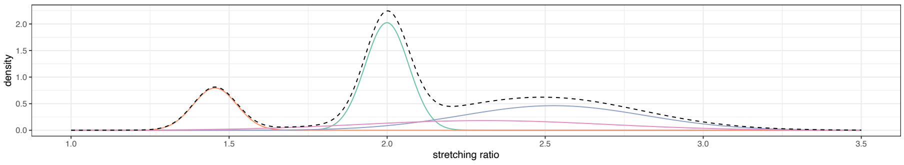

Figure 3: Estimated density on the explanatory variable, first optimal solution. Trimming level $\alpha = 0.08$, tone perception data.

References

CERIOLI, ANDREA, GARCÍA-ESCUDERO, LUIS ANGEL, MAYO-ISCAR, AGUSTÍN, & RIANI, MARCO. 2018. Finding the number of normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics*, **27**(2), 404–416.

DE VEAUX, RICHARD D. 1989. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, **8**(3), 227–245.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ISCAR, A. 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, **27**(2), 377–402.

INGRASSIA, SALVATORE, & PUNZO, ANTONIO. 2020. Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition. *Journal of Classification*, **37**(2), 526–547.