An adaptive method to robustify ML estimation in Cluster Weighted Modeling

Stima robusta per Cluster Weighted Model

L.A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar

Abstract Cluster-Weighted Models are a wide family of mixture distributions for modeling the joint probability of data coming from a heterogeneous population, and includes mixtures of distributions and mixtures of regressions as special cases. Unfortunately, they suffer from non-regular maximum likelihood issues, due to possible spikes and unboundedness in the target function. We propose an improved version of the Gaussian Cluster-Weighted estimation methodology, by trimming a portion α of the data and imposing constraints to the estimated variances. Trimming provides robustness properties to the estimators and constraints move the maximization problem to a well-posed setting and allow to avoid spurious solutions, i.e. fitting a small localized random pattern in the data rather than a proper underlying cluster structure. Theoretical results are illustrated using a few empirical studies.

Abstract *I modelli Cluster Weighted sono una ampia famiglia di misture che comprendono, come casi particolari, le misture di distribuzioni e di regressione e consentono di modellizzare dati eterogenei. Per questi modelli si osservano problemi di non regolarità della funzione di verosimiglianza, che può essere illimitata e avere spikes. In questo lavoro si introduce un nuovo metodo di stima del Cluster Weighted Gaussiano, che consente di eliminare una porzione α di dati contaminati e di imporre vincoli alla stima delle matrici di covarianza. Le proprietà di robustezza degli stimatori sono assicurate dall'eliminazione degli outliers, e la stima vincolata fa sì che il problema di massimizzazione sia ben posto e riduce le soluzioni spurie, ovvero l'adattarsi del modello ad un piccolo raggruppamento casuale di dati invece che ad una vera e propria componente della mistura. I risultati teorici sono illustrati anche mediante alcune analisi empiriche.*

Luis Angel García-Escudero · Alfonso Gordaliza · Agustín Mayo-Iscar
Departamento de Estadística e Investigación Operativa, Universidad de Valladolid (Spain) e-mail: lagarcia@eio.uva.es, alfonsog@eio.uva.es, agustinm@eio.uva.es

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

Salvatore Ingrassia
Department of Economics and Business, University of Catania (Italy). e-mail: s.ingrassia@unict.it

Key words: Constrained estimation, Cluster Weighted Modeling, Mixture of regression, Model-Based Clustering.

1 Introduction and Motivation

The analysis of mixture models is a fertile source of non-regular maximum likelihood problems. For instance, a two-component normal mixture incurs the problem of unbounded likelihood if the mean parameter of the first component is set to be one of the data values and the standard deviation σ is allowed to tend to 0. However, the singularity does not impose itself until σ becomes extremely small. In many normal mixture problems susceptible to unbounded likelihood, there is also, an asymptotically consistent local maximum (Redner and Walker, 1984), but still spurious solutions could drive the maximization of the target function far away from the true value of the parameter. Moreover, it is well known that a small fraction of outlying observations (background noise, pointwise contamination, unexpected minority patterns, etc.) could severely affect ML parameter estimation. With these considerations in mind, we approach Cluster-Weighted Models (CWMs), introduced in Gershenfeld (1997). They are a flexible family of mixture models for fitting the joint density of a pair $(\mathbf{X};Y)$ composed by a response variable Y and by a vector of covariates $\mathbf{X}$, assuming that data are coming from a heterogenous population. Ingrassia *et al.* (2012) show that Gaussian CWM includes, as special cases, mixtures of distributions and finite mixture of regression models. Our purpose is to modify the classical ML method, by adding trimming and constraints in such a way to make robust and free from non-regularity conditions the model estimation. We have organized the rest of the paper as follows. In Section 2 we recall the main ideas about the CWM and we discuss issues in EM estimation. In Section 3 we present the trimmed CWRM and a feasible algorithm for its implementation. Concluding remarks will end the paper.

2 Cluster Weighted Modeling

Let $p(\mathbf{x},y)$ be the joint density of $(\mathbf{X},Y)$. Suppose that Ω can be partitioned into G groups, say $\Omega_1,\ldots,\Omega_G$. In this work, we will focus on CWM linear models with Gaussian components: assuming $\mathbf{X}|\Omega_g \sim N_d(\mu_g,\Sigma_g)$, a linear relationship between Y and $\mathbf{x}$ in the g-th group written as $Y = \mathbf{b}_g'\mathbf{x} + b_g^0 + \varepsilon_g$ where $\varepsilon_g \sim N(0,\sigma_g^2)$, and $Y|\mathbf{x},\Omega_g \sim N(\mathbf{b}_g'\mathbf{x} + b_g^0, \sigma_g^2)$, the *linear Gaussian CWM* has the following density $p(\mathbf{x},y;\theta) = \sum_{g=1}^{G} \phi(y;\mathbf{b}_g'\mathbf{x} + b_g^0,\sigma_g)\phi_d(\mathbf{x};\mu_g,\Sigma_g)\pi_g$.As usual, $\phi_d(\cdot;\mu_g,\Sigma_g)$ is the density of the d-variate Gaussian distribution with mean vector μ_g and covariance matrix Σ_g, and π_g is the weight of Ω_g in the mixture.The ML estimation of the Gaussian CWM suffers from a serious lack of robustness, which should be taken into account due to the common presence of noise sources in data. To illustrate this problem, Figure 1(a) shows a simulated data set, *Simdata1*, generated from

a linear Gaussian CWM with $G = 2$, 90 observations from each component. 20 contaminating observations have been added as either background noise, or pointwise contamination around the point $(15, 20)$. The true underlying regression lines are represented with dotted lines in Figure 1, which shows that contaminating data points seriously affected the estimation.

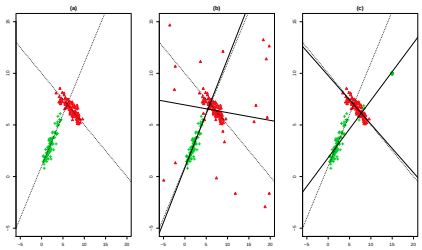

Fig. 1 *Simdata1*: (a) original data and Cluster Weighted Model fitting; (b) original data plus background noise and fitted model; (c) original data plus pointwise contamination and fitted model.

Another important issue concerns the unboundedness of the target function $\sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \phi(y_i; \mathbf{b}_g' \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \pi_g \right]$, when no constraints are imposed on the scatter parameters. In this case, the defining problem is ill-posed because the target function tends to ∞ when either $\mu_g = \mathbf{x}_i$ and $|\Sigma_g| \to 0$ or $y_i = \mathbf{b}_g' \mathbf{x}_i + b_g^0$ and $\sigma_g^2 \to 0$. Moreover, as a trivial consequence of the unboundedness, the EM algorithms often applied to fit a CWM can be trapped into non-interesting local maximizers and the result of the EM algorithm strongly depends on the initialization of the algorithm. Spurious solutions may be due to very localized patterns in the explanatory variables, as it will be shown by a second simulated data set, *Simdata2*. In Figure 2, two sets of 90 observations for $\mathbf{X}$ were drawn from two bivariate spherical normal distributions centered, respectively, at $(2, 2)$ and $(4, 4)$. Furher, 20 almost collinear observations, were added close to the second group, centered at $(4, 4)$. The same linear functions with equally distributed error terms have been considered, to generate the response variable Y. We can see in Figure 2 that the standard fit of the CWM yields to the determination of a "spurious" component with the 20 almost collinear observations and a second component joining together the two groups, with 90% of the observations. To overcome the issues we illustrated in these examples, in the next section we will propose an alternative methodology which incorporates trimming and constraints to the CWM.

3 Trimmed Cluster Weighted Restricted Modeling

For a given sample of n observations, the trimmed CWRM methodology is based on the maximization of the following log-likelihood function

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

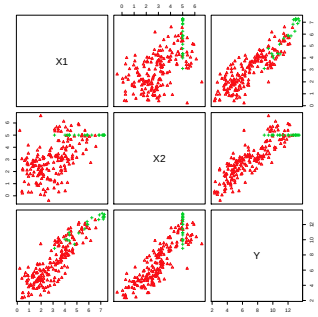

Fig. 2 *Simdata2*: Scatter plot matrix. Dots in green concern collinear observations.

where $z(\cdot,\cdot)$ is a 0-1 trimming indicator function that tell us whether observation $(y_i,\mathbf{x}_i)$ is trimmed off ($z(\mathbf{x}_i,y_i)$=0), or not ($z(\mathbf{x}_i,y_i)$=1). A fixed fraction α(trimming level) of observations can be unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i,y_i) = [n(1-\alpha)]$. Analogous approaches based on trimmed mixture likelihoods can be found in Neykov *et al.* (2007), Gallegos and Ritter (2009) and García-Escudero *et al.* (2013). Moreover, we introduced two further constraints, on the set of eigenvalues $\{\lambda_l(\Sigma_g)\}_{l=1,\ldots,d}$ of the scatter matrices Σ_g

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

and on the variances σ_g^2 of the regression error terms,

$$\sigma_{g_1}^2 \leq c_\varepsilon \sigma_{g_2}^2 \text{ for every } 1 \leq g_1 \neq g_2 \leq G \quad \text{ with } 0 < c_X, c_\varepsilon < +\infty.$$

These constraints can be seen as an extension to CWMs of those introduced in Ingrassia and Rocci (2007), García-Escudero *et al.* (2008) and Greselin and Ingrassia (2010) and go back to Hathaway (1985). Here, they allow for a specific treatment when modeling the marginal distribution of $\mathbf{X}$ and the regression error term, giving a high flexibility to the model.

Let us consider now the effects of trimming in the two data sets derived from *Simdata1*. In Figure 3 we can see that setting $\alpha = 0.1$ allows to restore the true structure of the data, by discarding the outlying observations, both in the case of background noise and huge pointwise contamination. Hence, trimming modifies the ML estimation in such a way that it is no more influenced by potential outliers and drives it far from the previous bad results shown in Figure 1. Commenting the use of constraints for *Simdata2*, we can see in Figure 3 how a moderate choice for c_X and c_ε allows to correctly detect the $G = 2$ main groups and to avoid the disturbing effect of the "spurious" pattern in the data. More information about the role played by the parameters α, c_X and c_ε could be given, omitted here for the sake of space.

3.1 Algorithm

The maximization of the target function on its parameters under the bounds given by c_X and c_ε is not an easy task, obviously. We will give a feasible algorithm ob-

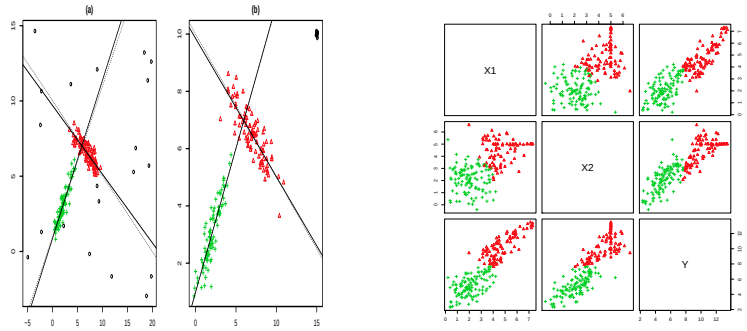

Fig. 3 Left panels:*Simdata1*: Results of fitting the trimmed CWRM with $\alpha = 0.1$, $c_X = c_\varepsilon = 20$ for the two data sets in Figure 1, panels (b) and (c). Trimmed points are denoted by black circles. Right panel: *Simdata2*: Results of fitting the trimmed CWRM with $\alpha = 0$, $c_X = c_\varepsilon = 20$ for data in Figure 2.

tained by combining the EM algorithm for CWM with that (with trimming and constraints) introduced in García-Escudero *et al.* (2013). The algorithm is initialized *nstart* times, by selecting different values of the initial parameter vector θ_0. Constraints on scatter matrices and variances of the error terms should be enforced (as described below, in the M-step). We will have adapted EM steps, alternatively executed until convergence. During the E-step:

- the current mixture density has to be evaluated at each observation in the sample, following the CWM methodology
- the proportion $1-\alpha$ of observations with highest values of the density is retained, so giving the subset I_α of untrimmed observations
- the posterior probabilities of the observations in I_α are computed, for trimmed observations they are set to 0.

During the M-step the parameters are updated, taking into account only the observations in I_α, i.e. tentatively discarding the observations suspicious to be outliers. Along the iterations, due to the updates, it may happen that the estimated scatter matrices T_g and the estimated variances of the error terms s_g^2 do not satisfy the constraints. To enforce them, the singular-value decomposition of $T_g = U_g' E_g U_g$ is considered, with U_g being an orthogonal matrix and $E_g = \mathrm{diag}(e_{g1}, e_{g2}, \ldots, e_{gd})$ a diagonal matrix. The truncated eigenvalues are defined as $[e_{gl}]_m = \min(c_X \cdot m, (\max(e_{gl}, m))$, with m being some threshold value. The scatter matrices are finally updated as $\Sigma_g^{(l+1)} = U_g' E_g^* U_g$, with $E_g^* = \mathrm{diag}\left([e_{g1}]_{m_{\mathrm{opt}}^X}, [e_{g2}]_{m_{\mathrm{opt}}^X}, \ldots, [e_{gp}]_{m_{\mathrm{opt}}^X}\right)$ and m_{opt}^X minimizing the real valued function

$$m \mapsto \sum_{g=1}^{G} \pi_g^{(l+1)} \sum_{l=1}^{d} \left(\log\left([e_{gl}]_m\right) + \frac{e_{gl}}{[e_{gl}]_m} \right).$$

Analogously, we introduce the truncated variances $[s_g^2]_m = \min(c_\varepsilon \cdot m, (\max(s_g^2, m))$. The variance of the error terms are finally updated as $\sigma_g^{2(l+1)} = [s_g^2]_{m_{\mathrm{opt}}^\varepsilon}$, with $m_{\mathrm{opt}}^\varepsilon$ minimizing the real valued function

$$m \mapsto \sum_{g=1}^{G} \pi_g^{(l+1)} \left(\log\left([s_g^2]_m\right) + \frac{s_g^2}{[s_g^2]_m} \right).$$

Proposition 3.2 in Fritz *et al.* (2013) shows that m_{opt}^X and $m_{\text{opt}}^\varepsilon$ can be obtained, respectively, by evaluating $2pG+1$ times (respectively $2G+1$ times) the corresponding real valued function.

Finally, at convergence, the set of parameters yielding the highest value of the target function and the associated set I_α of untrimmed observations are returned as the final algorithm output.

In this work, we have presented a methodology based on trimming and constraints to robustify and control variabilities in a linear Gaussian CWM, moving the likelihood maximization to a well-posed setting. An algorithm, with an affordable increase in computing time, has been also given for its practical implementation. We have seen that the proposed methodology drives the estimation procedure to identify and discard sparse outliers, and even strongly concentrated contaminating observations, acting as leverage points, which are so harmful in the framework of regression mixtures. At the same time, the constraints serve to avoid the likelihood singularities and reduce the detection of spurious solutions.

Further research is needed to tune the choice of the involved parameters, and this is not an easy task, as these parameters are clearly interrelated. First attempts to extract such information from the observed sample are currently under study.

References

Fritz, H., García-Escudero, L., and Mayo-Iscar, A. (2013). A fast algorithm for robust constrained clustering. *Computational Statistics & Data Analysis*, (61), 124–136.

Gallegos, M. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixtures. *Sankhya (Ser. A)*, (71), 164–220.

García-Escudero, L., Gordaliza, A., and Mayo-Iscar, A. (2013). A constrained robust proposal for mixture modeling avoiding spurious solutions. Accepted for publication in *Advances in Data Analysis and Classification*.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

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

Greselin, F. and Ingrassia, S. (2010). Constrained monotone EM algorithms for mixtures of multivariate *t* distributions. *Statistics and computing*, **20**(1), 9–22.

Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

Ingrassia, S. and Rocci, R. (2007). Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Comput. Stat. Data Anal.*, (51), 5339–5351.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via the Cluster-Weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.

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

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, pages 195–239.