Penalized Versus Constrained Approaches for Clusterwise Linear Regression Modeling

Roberto Di Mari, Stefano Antonio Gattone, and Roberto Rocci

Abstract Several approaches exist to avoid singular and spurious solutions in maximum likelihood (ML) estimation of clusterwise linear regression models. We propose to solve the degeneracy problem by using a penalized approach: this is done by adding a penalty term to the log-likelihood function which increasingly penalizes smaller values of the scale parameters, and the tuning of the penalty term is done based on the data. Another traditional solution to degeneracy consists in imposing constraints on the variances of the regression error terms (constrained approach). We will compare the penalized approach to the constrained approach in a simulation study, providing practical guidelines on which approach to use under different circumstances.

Keywords Clusterwise linear regression · Penalized likelihood · Scale constraints

1 Introduction

Let $\mathrm{y}_1, \ldots, \mathrm{y}_n$ be a sample of independent observations drawn from the response random variable Y_i, each observed alongside with a vector of J explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let us assume $\mathrm{Y}_i|\mathbf{x}_i$ to be distributed as a finite mixture of linear regression models, that is

R. Di Mari (✉)
Department of Economics and Business, University of Catania, Catania, Italy
e-mail: roberto.dimari@unict.it

S. A. Gattone
Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, Italy
e-mail: gattone@unich.it

R. Rocci
Department of Statistical Sciences, University of Rome La Sapienza, Rome, Italy
e-mail: roberto.rocci@uniroma1.it

S. Balzano et al. (eds.), *Statistical Learning and Modeling in Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-69944-4_10

$$f(\mathrm{y}_i|\mathbf{x}_i;\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g \phi_g(\mathrm{y}_i|\mathbf{x}_i, \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right], \tag{1}$$

where G is the number of clusters and p_g, $\boldsymbol{\beta}_g$, and σ_g^2 are the mixing proportion, the vector of $J+1$ regression coefficients that includes an intercept, and the variance term for the gth cluster. The set of all model parameters is given by $\boldsymbol{\psi} = \{(p_1, \ldots, p_G; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G; \sigma_1^2, \ldots, \sigma_G^2) \in \mathbb{R}^{(G-1)+(J+1)G+G} : p_1 + \cdots + p_G = 1, p_g > 0, \sigma_g^2 > 0, \text{ for } g = 1, \ldots, G\}$.

The likelihood function can be specified as

$$\mathscr{L}(\boldsymbol{\psi}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g{}^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right]\right\}, \tag{2}$$

which we maximize to estimate $\boldsymbol{\psi}$ either by means of direct maximization or with the perhaps more popular EM algorithm [5]. However, there is a well-known complication in ML estimation of this class of models: the likelihood function of mixtures of (conditional) normals with cluster-specific variances is unbounded [4, 11].

A traditional solution to the problem of unboundedness is based on the seminal work of [7] which, for univariate mixtures of normals, suggested imposing a lower bound to the ratios of the scale parameters in the maximization step. The method is equivariant under linear affine transformations of the data. That is, if the data are linearly transformed, the estimated posterior probabilities do not change and the clustering remains unaltered. Recently, in the multivariate case, [12] incorporated constraints on the eigenvalues of the component covariances matrices of Gaussian mixtures that are tuned on the data based on a cross-validation strategy. These constraints are built upon [9]'s reformulation and are an equivariant sufficient condition for Hathaway's constraints. Estimation is done in a familiar ML environment [10], with a data-driven selection of the scale balance. Di Mari et al. [6] adapted [12]'s method to clusterwise linear regression, further investigating its properties.

Another possible approach for handling unboundedness is to modify the log-likelihood function by adding a penalty term, in which smaller values of the scale parameters are increasingly penalized. Representative examples can be found in [1–3].

In this work, we review the constrained approach of [6] and develop a data-driven equivariant penalized approach for ML estimation. In Sect. 2, we sketch the bulk of the methodologies; in Sect. 3 we report the results from the simulation study and then draw some conclusions (Sect. 4).

2 The Methodology

2.1 The Constrained Approach

Di Mari et al. [6] proposed relative constraints on the group conditional variances σ_g^2 of the kind

$$\sqrt{c} \leq \frac{\sigma_g^2}{\bar{\sigma}^2} \leq \frac{1}{\sqrt{c}}, \tag{3}$$

or equivalently

$$\bar{\sigma}^2\sqrt{c} \leq \sigma_g^2 \leq \bar{\sigma}^2\frac{1}{\sqrt{c}}. \tag{4}$$

The above constraints are equivariant and have the effect of shrinking the variances to a suitably chosen $\bar{\boldsymbol{\sigma}}^2$, the *target* variance term, and the level of shrinkage is given by the value of c. These constraints are easily implementable within the EM algorithm [9, 10], which is fully available in closed form, and the selection of c is based on the data.

2.2 The Penalized Approach

An alternative to the constrained estimator is the penalized approach, in which a penalty $s_n(\sigma_1^2, \ldots, \sigma_G^2)$ is put on the component variances and it is added to the log-likelihood. Under certain conditions on the penalty function, the penalized estimator is know to be consistent [1]. A function s_n that satisfies these conditions is

$$s_n(\sigma_1^2, \ldots, \sigma_G^2) = -\lambda \sum_{g=1}^{G} \left(\frac{\bar{\sigma}^2}{\sigma_g^2} + \log(\sigma_g^2) \right), \tag{5}$$

where $\bar{\sigma}^2$, the *target* variance, can be seen as our *prior* information on the scale structure and λ is the penalizing constant that is selected based on the data. Thus, the penalized log-likelihood can be written as

$$p\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + s_n(\sigma_1^2, \ldots, \sigma_G^2) \tag{6}$$

and the set of unknown parameters is found by ML with computation done by means of an EM algorithm that is available in closed form. Besides the constrained approach, the penalized approach is equivariant with respect to linear transformation in the response.

2.3 Selection of the Tuning Parameter

Both approaches require selection of the tuning parameter—c and λ, respectively, for the constrained and penalized estimators. The tuning constants can be pre-specified by the user if any prior knowledge on the scale structure of the cluster is available. If this is not the case, the tuning can be based on the data. We propose two alternative approaches to select the tuning constant that can be used for both constrained and penalized methods.

2.3.1 Cross-Validation

The first tuning approach is based on a cross-validation strategy that looks for a tuning parameter such that the cross-validated likelihood is maximized. For a given c or λ, this is done as follows:

1. Temporary estimates for the model parameters are obtained from the entire sample, and these are used as starting values to initialize the cross-validation procedure.
2. The data set is partitioned into training and test sets.
3. Parameters are estimated on the training set and the contribution to the log-likelihood of the test set is computed.
4. Steps 2–3 are repeated M times and the M contributions to the log-likelihood of the test set are summed for different values of c/λ.

2.3.2 k-Deleted Method

The second tuning approach is based on the modification of the k-deleted method [13, 14] that looks for a tuning parameter such that the (modified) k-deleted log-likelihood[1] is maximized.

For a given c or λ, this is done as follows:

1. Temporary estimates for the model parameters are obtained from the entire sample, and these are used as starting values to initialize the procedure.
2. For a given c/λ, the model parameters are estimated.
3. The (modified) k-deleted log-likelihood is computed.
4. Steps 2–3 are repeated for different values of c/λ.

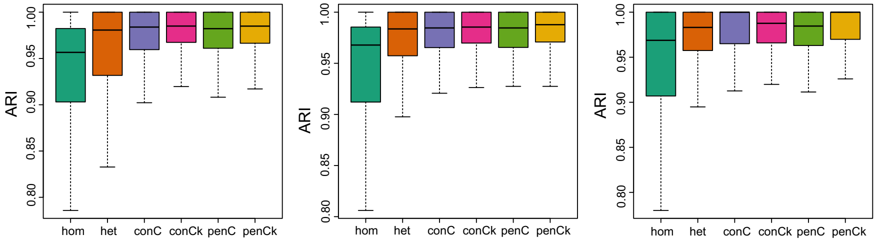
[1] For some estimates of the model parameters, this is computed by taking out the k units with the largest log-likelihood.

3 Simulation Study

A simulation study has been conducted to compare the quality of the parameter estimates and the ability to recover the clusters structure of the constrained and the penalized approaches. Both tuning strategies—cross-validation based and k-deleted method—were considered for the constrained and penalized approaches—respectively *conC*, *conCk*, *penC*, and *penCk*—and the unconstrained estimator with common (homoscedastic) component-scales (hom) and the unconstrained estimator with different (heretoscedastic) component-scales (het) were also included for comparison.

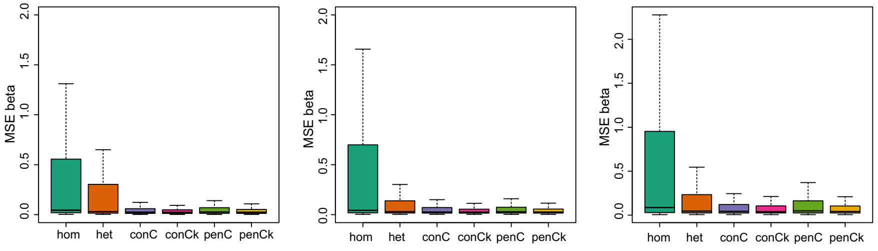
The target measures used for the comparisons were average Mean Squared Errors (MSE) of the regression coefficients (averaged across regressors and groups) and the adjusted Rand index [8].

We generated the data from a 3-group clusterwise linear regression model with 3 regressors and an intercept term. The group mixing weights were set equal to 0.1, 0.3, and 0.6. The regressors were generated from 3 independent standard normal distributions; regression coefficients were randomly generated from Uniform distributions U(−1.5, 1.5), and the group-specific intercepts were set equal to 4, 9, and 16.



(a) $\sigma^2 = (0.1, 0.8, 0.1)$ (b) $\sigma^2 = (0.2, 0.6, 0.2)$ (c) 0.5, 0.5, 0.5

Fig. 1 (average) MSE of the regression coefficients for all approaches, for the three scale scenarios and $n = 100$



(a) $\sigma^2 = (0.1, 0.8, 0.1)$ (b) $\sigma^2 = (0.2, 0.6, 0.2)$ (c) 0.5, 0.5, 0.5

Fig. 2 Adjusted Rand Index (ARI) for all approaches, for the three scale scenarios and $n = 100$

We considered 6 crossed simulation conditions of sample size—$n = 100, 200$—and scale scenarios—$\boldsymbol{\sigma}^2 = (0.1, 0.8, 0.1)'$ (*heteroscedasticity*), $\boldsymbol{\sigma}^2 = (0.2, 0.6, 0.2)'$ (*mild heteroscedasticity*), and $\boldsymbol{\sigma}^2 = (0.5, 0.5, 0.5)'$ (*homoscedasticity*)

For each simulation condition, we generated 250 samples and, for each approach, we selected the best solution (highest likelihood) out of 10 random starts. We report only the results for $n = 100$ as those for $n = 200$ were qualitatively the same (Figs. 1 and 2).

We observe that the penalized and constrained approaches overcome their unconstrained rivals (hom and het) both in terms of quality of regression parameter estimates and cluster recovery. It seems that while with a tuning based on the more time-consuming cross-validation strategy conC does slightly better than penC, with the more efficient k-deleted tuning the penalized approach penCk does better than conCk. Overall, penCk delivers the best performance.

4 Concluding Remarks

In this work, we have proposed a new penalized estimator for clusterwise linear regression models in which penalties are put on the component scales. This penalized estimator is equivariant under changes in the scale of the response. We have compared it with the constrained approach of [6] and illustrated two alternative tuning strategies for both methodologies. The constrained and penalized estimators perform uniformly better than unconstrained ones. Whenever the computing time of tuning strategies is not an issue, both approaches serve well the scope of fitting clusterwise linear regression models. For quicker—and perhaps less-refined selection strategies—like the k-deleted method, the penalized approach seems to be preferable.

References

1. Chen, J., Tan, X.: Inference for multivariate normal mixtures. J. Multivariate Anal. **100**(7), 1367–1383 (2009)
2. Chen, J., Tan, X., Zhang, R.: Inference for normal mixtures in mean and variance. Stat. Sinica, 443–465 (2008)
3. Ciuperca, G., Ridolfi, A., Idier, J.: Penalized maximum likelihood estimator for normal mixtures. Scandinavian J. Stat. **30**(1), 45–59 (2003)
4. Day, N.: Estimating the components of a mixture of normal distributions. Biometrika **56**(3), 463–474 (1969)
5. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977)
6. Di Mari, R., Rocci, R., Gattone, S.: Clusterwise linear regression modeling with soft scale constraints. Int. J. Approx. Reas. **91**, 160–178 (2017)
7. Hathaway, R.: A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann. Stat. **13**(2), 795–800 (1985)
8. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)

9. Ingrassia, S.: A likelihood-based constrained algorithm for multivariate normal mixture models. Stat. Methods Appl. **13**(2), 151–166 (2004)
10. Ingrassia, S., Rocci, R.: Constrained monotone em algorithms for finite mixture of multivariate gaussians. Comput. Stat. Data Anal. **51**(11), 5339–5351 (2007)
11. Kiefer, J., Wolfowitz, J.: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. 887–906 (1956)
12. Rocci, R., Gattone, S., Di Mari, R.: A data driven equivariant approach to constrained gaussian mixture modeling. Advanc. Data Anal. Classif. **12**(2), 235–260 (2018)
13. Seo, B., Kim, D.: Root selection in normal mixture models. Comput. Stat. Data Anal. **56**(8), 2454–2470 (2012)
14. Seo, B., Lindsay, B.G.: A computational strategy for doubly smoothed mle exemplified in the normal mixture model. Comput. Stat. Data Anal. **54**(8), 1930–1941 (2010)