International Journal of Approximate Reasoning 91 (2017) 160–178

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

# Clusterwise linear regression modeling with soft scale constraints [☆]

Roberto Di Mari [a,*], Roberto Rocci [b], Stefano Antonio Gattone [c]

[a] Department of Economics and Business, University of Catania, Italy
[b] Department of Economics and Finance, University of Rome Tor Vergata, Italy
[c] Department of Philosophical and Social Sciences, Economics and Quantitative Methods, University G. d'Annunzio, Chieti-Pescara, Italy

ARTICLE INFO

*Article history:*
Received 10 November 2016
Received in revised form 9 September 2017
Accepted 10 September 2017
Available online 18 September 2017

*Keywords:*
Clusterwise linear regression
Adaptive constraints
Regression equivariance
Plausible bounds
Soft estimators
Constrained EM algorithm

ABSTRACT

Constrained approaches to maximum likelihood estimation in the context of finite mixtures of normals have been presented in the literature. A fully data-dependent soft constrained method for maximum likelihood estimation of clusterwise linear regression is proposed, which extends previous work in equivariant data-driven estimation of finite mixtures of normals. The method imposes soft scale bounds based on the homoscedastic variance and a cross-validated tuning parameter $c$. In our simulation studies and real data examples we show that the selected $c$ will produce an output model with clusterwise linear regressions and clustering as a most-suited-to-the-data solution in between the homoscedastic and the heteroscedastic models.

## 1. Introduction

Let $\{(y_i, \mathbf{x}_i)\}_n = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ be a sample of $n$ independent units, where $y_i$ is the outcome variable and $\mathbf{x}_i$ are the $J$ covariates. A clusterwise linear regression model assumes that the density of $y_i|\mathbf{x}_i$ is given by

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

where $G$ is the number of clusters, $\boldsymbol{\psi} = \{(p_1, \dots, p_G; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G; \sigma_1^2, \dots, \sigma_G^2) \in \mathbb{R}^{G(J+2)} : p_1 + \cdots + p_G = 1, p_g \geq 0, \sigma_g^2 > 0, g = 1, \dots, G\}$ is the set of model parameters, and $p_g$, $\boldsymbol{\beta}_g$, and $\sigma_g^2$ are respectively the mixing proportions, the vector of $J$ regression coefficients, and the variance term for the $g$-th cluster. The model in Equation (1) is also known under the name of finite mixture of linear regression models, or switching regression model [21,22,15].

The parameters of finite mixtures of linear regression models are identified if some mild regularity conditions are met [10].

The clusterwise linear regression model of Equation (1) can naturally serve as a classification model. Based on the model, one computes the posterior membership probabilities for each observation as follows:

[☆] This paper is part of the Virtual special issue on Soft methods in probability and statistics, Edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani

[*] Corresponding author.
*E-mail addresses:* roberto.dimari@unict.it (R. Di Mari), roberto.rocci@uniroma2.it (R. Rocci), gattone@unich.it (S.A. Gattone).

http://dx.doi.org/10.1016/j.ijar.2017.09.006
0888-613X/© 2017 Elsevier Inc. All rights reserved.

$$\mathrm{p}(\mathrm{g}|\mathrm{y}_i) = \frac{p_g f_g(\mathrm{y}_i|\mathbf{x}_i; \sigma_g^2, \boldsymbol{\beta}_g)}{\sum_{h=1}^{G} p_h f_h(\mathrm{y}_i|\mathbf{x}_i; \sigma_h^2, \boldsymbol{\beta}_h)}, \tag{2}$$

and then classify each observation according, for instance, to fuzzy or crisp classification rules.

The problem of clustering sample points grouped around linear structures has been receiving a lot of attention in the statistical literature because of its important applications (see, for instance, [16], and references therein. For the robust literature, among the others, see [6,7]).

In order to estimate $\boldsymbol{\psi}$, one has to maximize the following sample likelihood function

$$\mathrm{L}(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi {\sigma_g}^2}} \exp\left[ - \frac{(\mathrm{y}_i - \mathbf{x}_i' \boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right] \right\}, \tag{3}$$

which can be done using iterative procedures like the EM algorithm [5], whose clustering can be interpreted as a fuzzy partition [9]. Unfortunately, maximum likelihood (ML) estimation of univariate unconditional or conditional normals suffers from the well-known issue of unboundedness of the likelihood function: whenever a sample point coincides with the group's centroid and the relative variance approaches zero, the likelihood function increases without bound ([14]; also the multivariate case suffers from the issue of unboundedness. See [4]). Hence a global maximum cannot be found.

Yet, ML estimation does not fail: Kiefer [15] showed that there is a sequence of consistent, asymptotically efficient and normally distributed estimators for switching regressions with different group-specific variances (heteroscedastic switching regressions). These estimators correspond, with probability approaching one, to local maxima in the interior of the parameter space. Nonetheless, although there is a local maximum which is also a consistent root, there is no tool for choosing it among the local maxima. Day [4] showed, for multivariate mixtures of normals, that potentially each sample point – or any pair of sample points being sufficiently close together, or co-planar [24] – can generate a singularity in the likelihood function of a mixture with heteroscedastic components. This gives rise, both in univariate and multivariate contexts, to a number of spurious maximizers [18].

The issue of unboundedness can be dealt with by imposing constraints on the component variances. This approach is based on the seminal work of Hathaway [8], who showed that imposing a lower bound, say $c$, to the ratios of the scale parameters of univariate mixtures of normals prevents the unboundedness of likelihood function. Although the resulting ML estimator is consistent and 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 – the proposed constraints are very difficult to apply within iterative procedures like the EM algorithm. Furthermore, the issue of how to choose $c$, which controls the strength of the constraints, remains open.

For multivariate mixtures of normals, Ingrassia [12] showed that it is sufficient, for Hathaway's constraints to hold, to impose bounds on the eigenvalues of the class conditional covariance matrices. This provides a constrained solution that 1) can be implemented at each iteration of the EM algorithm, and 2) still preserves the monotonicity of the resulting EM [13]. Recently, Rocci et al. [23, RGD] proposed a constrained estimation method for multivariate mixtures of normals, being characterized by 1) fully data-dependent constraints, 2) equivariance of the clustering algorithm under change of scale in the data, and 3) ease of implementation within standard routines [12,13].

The aim of this paper is to extend the RGD constrained estimation method to clusterwise linear regression models. We demonstrate that it works very well when a conditional distribution (linear regression) is specified for each mixture component.

A correct estimation of the regression coefficients is crucial in a regression context, where the focus is not only on the cluster recovery, but on the interpretation of the estimated associations. RGD (2017) looked at how good the method was at recovering clusters: in our simulation study we also bring into focus the regression parameters, and look at the quality of the estimators in terms of mean squared error – which embeds both the bias and the variance of the estimators. In this new perspective, we can now argue, based on the evidence of our simulation studies and empirical examples, that the RGD constrained estimation method yields a final model – in terms of clustering and estimated parameters – in between the fully constrained model and the unconstrained model. How close to which of the extremes is optimally determined by maximizing a suitable objective function.

Starting from Rousseeuw and Leroy [25]'s nomenclature, the equivariance property in linear models is of three types: regression, affine and scale. Regression equivariance holds if, by adding a linear combination of the covariates to the response variable through any column vector, the model parameters are shifted by that same vector. If instead an affine transformation is applied on the covariates, affine equivariance guarantees that the model parameters are transformed accordingly. That is, the linear predictor remains the same. The third type of equivariance refers to scale changes in the response variable, in that the model parameters are rescaled such that the linear predictor and the error's standard deviation are both on the new response scale. Either equivariance property types hold for the unconstrained clusterwise linear regression. Notice that neither affine transformations of the $\mathbf{x}$s nor shifts in the response proportional to the covariates affect the error's variance: therefore regression and affine equivariance still hold for the constrained model. Indeed, scale equivariance is no longer guaranteed in the constrained model, as constraints might prevent the error's variance to adapt to the new scale of the response variable.

Crucially in a clustering context, scale equivariance ensures that the clustering is the same for any (re)scaling of the response variable, hence solving the problem of finding the best scale for the response.

The remainder of the paper is organized as follows. In Section 2 we briefly review Hathaway's constraints and the sufficient condition in Ingrassia [12]. Section 3 is devoted to a description of the proposed methodology and of the estimation algorithm, which is evaluated through two simulation studies, presented in Section 4, and four real data examples, in Section 5. Section 6 concludes with some final discussion and some ideas for future research.

## 2. Constrained approaches for ML estimation

In the context of finite mixtures of univariate normals, Hathaway [8] proposed relative constraints on the variances of the kind

$$\min_{i\neq j}\frac{\sigma_i^2}{\sigma_j^2}\geq c \quad \text{with} \quad c\in(0,1]. \tag{4}$$

Hathaway's formulation of the maximum likelihood problem presents a strongly consistent global solution, no singularities, and a smaller number of spurious maxima.

Ingrassia [12] formulated a sufficient condition such that Hathaway's constraints hold, which is implementable directly within the EM algorithm, where the covariance matrices are iteratively updated. In a univariate setup, he shows that constraints in (4) are satisfied if

$$a\leq\sigma_g^2\leq b, \quad \text{with} \quad g=1,\dots,G, \tag{5}$$

where $a$ and $b$ are positive numbers such that $a/b\geq c$. Complementing the work of Ingrassia [12], Ingrassia and Rocci [13] formulated the conditions under which the constrained algorithm preserves the monotonicity of the unconstrained EM: this yields a non-decreasing sequence of the likelihood values if the initial guess $\sigma_g^{2(0)}$ satisfies the constraint.

## 3. The proposed methodology

The present Section gives details on the equivariance of the unconstrained ML estimation of the clusterwise linear regression model (Section 3.1). It then shows that equivariance is preserved by the proposed constraints (Section 3.2), which restrict the cluster-conditional variances to lie on a neighborhood of a given target variance, say $\xi^2$, the sole condition under which the equivariance hold being that the method used to estimate the target is also equivariant. Then we describe the constrained EM algorithm, which we use for estimating the model parameters (Section 3.3) and the cross-validation strategy used for selecting $c$ (Section 3.4). Finally, we discuss issues related to the choice of $\xi^2$ in Section 3.5.

### 3.1. Equivariance of the unconstrained model

We wish to deal with a clustering method which is not affected by the way the response variable is expressed – in terms of translations or changes in the unit of measurement. To see how this works in the clusterwise linear regression case, note that if we let $\mathrm{y}_i^*=\gamma\mathrm{y}_i$, where $\gamma$ is any real number different than zero, we have

$$f(\mathrm{y}_i^*|\mathbf{x}_i;\boldsymbol{\psi}^*)=\sum_{g=1}^{G}p_g f_g(\mathrm{y}_i^*|\mathbf{x}_i;\sigma_g^{*2},\boldsymbol{\beta}_g^*), \tag{6}$$

where $\sigma_g^{*2}=\gamma^2\sigma_g^2$, and $\boldsymbol{\beta}_g^*=\gamma\boldsymbol{\beta}_g$. This implies the following relation

$$f(\mathrm{y}_i|\mathbf{x}_i;\boldsymbol{\psi})=\gamma f(\mathrm{y}_i^*|\mathbf{x}_i;\boldsymbol{\psi}^*) \tag{7}$$

showing the equivariance of the clusterwise linear regression model under scale transformations of the response variable. The equivariance of the unconstrained (heteroscedastic) model implies the invariance of the classification which would not be affected by the scale of the response variable since the posterior probabilities are given by

$$\begin{aligned}\mathrm{p}(g|\mathrm{y}_i)&=\frac{p_g f_g(\mathrm{y}_i|\mathbf{x}_i;\sigma_g^2,\boldsymbol{\beta}_g)}{\sum_{h=1}^{G}p_h f_h(\mathrm{y}_i|\mathbf{x}_i;\sigma_h^2,\boldsymbol{\beta}_h)}\\&=\frac{p_g\gamma f_g(\mathrm{y}_i^*|\mathbf{x}_i;\sigma_g^{*2},\boldsymbol{\beta}_g^*)}{\sum_{h=1}^{G}p_h\gamma f_h(\mathrm{y}_i^*|\mathbf{x}_i;\sigma_h^{*2},\boldsymbol{\beta}_h^*)}=\mathrm{p}(g|\mathrm{y}_i^*).\end{aligned} \tag{8}$$

Being the transformed data likelihood

$$\mathrm{L}^*=\mathrm{L}\left(\boldsymbol{\psi}^*;\mathbf{y}^*\right)=\prod_{i=1}^{n}\sum_{g=1}^{G}p_g f_g(\mathrm{y}_i^*|\mathbf{x}_i;\sigma_g^{*2},\boldsymbol{\beta}_g^*), \tag{9}$$

we have that

$$\mathrm{L} = \mathrm{L}\left(\boldsymbol{\psi};\mathbf{y}\right) = \gamma^{n}\mathrm{L}\left(\boldsymbol{\psi}^{*};\mathbf{y}^{*}\right) = \gamma^{n}\mathrm{L}^{*}. \tag{10}$$

Then, the scale equivariance property of the maximum likelihood estimator (MLE) simply follows from ML theory [31]: if $\hat{\sigma}_g^2$ and $\hat{\boldsymbol{\beta}}_g$ are the MLE for the model of $y$ on $x$ then $\hat{\sigma}_g^{*2} = \gamma^2\hat{\sigma}_g^2$ and $\hat{\boldsymbol{\beta}}_g^{*} = \gamma\hat{\boldsymbol{\beta}}_g$ are the MLE for the model of $y^*$ on $x$. Notice that scale equivariance holds also for the homoscedastic model - the equations are the same as above but without the subscript $g$ on the variance term.

### 3.2. Scale equivariant constraints

Starting from the set of constraints of equation (5), let $\xi^2$ be the target variance. The set of constraints for a clusterwise linear regression context are formulated as follows

$$\sqrt{c} \leq \frac{\sigma_g^2}{\xi^2} \leq \frac{1}{\sqrt{c}},$$

or equivalently

$$\xi^2\sqrt{c} \leq \sigma_g^2 \leq \xi^2\frac{1}{\sqrt{c}}, \tag{11}$$

where $c \in (0, 1]$.

It is easy to show that (11) implies (4) while the converse is not necessarily true, since (11) is more stringent than (4). That is

$$\frac{\sigma_g^2}{\sigma_j^2} = \frac{\sigma_g^2/\xi^2}{\sigma_j^2/\xi^2} \geq \frac{\sqrt{c}}{1/\sqrt{c}} = c.$$

By considering the re-scaled $y$, i.e. $y^*$, we have that

$$\frac{\sigma_g^2}{\xi^2} = \frac{\sigma_g^2\gamma^2}{\xi^2\gamma^2} = \frac{\sigma_g^{*2}}{\xi^{*2}}. \tag{12}$$

Equation (12) shows that the constraints (11) are scale equivariant, provided that the method used to select $\xi^2$ is also scale equivariant. Note that, together with scale equivariance, regression and affine equivariance hold as well as in the unconstrained case, and would still hold also using Ingrassia [12]'s constraints of Equation (5), the reason being that neither transformation of the **x**'s nor shifts in the response proportional to the covariates affect the error's variance.

Such constraints have the effect of shrinking the component variances to $\xi^2$, and the level of shrinkage is given by the value of $c$: such a formulation makes it possible to reduce the number of tuning constants from two – $(a, b)$ as in Ingrassia [12]'s proposal – to one – $c$. Note that for $c = 1$, $\hat{\sigma}_g^2 = \xi^2$, whereas if $c \to 0$, the final solution will approach the unconstrained one. Intuitively, the constraints (11) provide with a way to obtain a model in between a too restrictive model, the common-variance model, and an ill-conditioned model, the heteroscedastic model. In other terms, high scale balance is generally an asset – as it means that there is some unknown transformation of the sample space that transfers the component not too far from the common variance setting – but it has to be traded with fit [24].

### 3.3. A constrained EM algorithm

For the sake of exposition, we will briefly outline the EM algorithm for the heteroscedastic clusterwise linear regression model, then describe the constrained EM, adapted from Ingrassia and Rocci [13] – in which it was used for mixtures of multivariate normals.

The expectation step (E-step) at the $(k+1)$-th iteration produces an update for the quantity

$$\mathrm{z}_{ig}^{(k+1)}(\mathrm{y}_i;\boldsymbol{\psi}^{(k)}) = \frac{p_g f_g(\mathrm{y}_i;\boldsymbol{\beta}_g^{(k)},\sigma_g^{2(k)})}{\sum_{h=1}^{G} p_h f_h(\mathrm{y}_i;\boldsymbol{\beta}_g^{(k)},\sigma_g^{2(k)})}, \tag{13}$$

where $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Using the computed quantities from step E, the maximization step (M-step) for the heteroscedastic model involves the following closed-form updates:

$$p_g^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{z}_{ig}^{(k+1)}, \tag{14}$$

$$\boldsymbol{\beta}_g^{(k+1)} = (\sum_{i=1}^{n}\mathrm{z}_{ig}^{(k+1)}\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_{i=1}^{n}\mathrm{z}_{ig}^{(k+1)}\mathbf{x}_i\mathrm{y}_i, \tag{15}$$

$$\sigma_g^{2(k+1)} = \frac{\sum_{i=1}^{n} z_{ig}^{(k+1)} (y_i - \mathbf{x}_i' \boldsymbol{\beta}_g^{(k+1)})^2}{\sum_{i=1}^{n} z_{ig}^{(k+1)}}. \tag{16}$$

Our constrained approach uses the same EM algorithm, in which the component variances are updated by using Equation (16) and then by applying the following rule

$$\sigma_{cg}^{2(k+1)} = \min\left(\xi^2 \frac{1}{\sqrt{c}}, \max\left(\xi^2\sqrt{c}, \sigma_g^{2(k+1)}\right)\right), \quad \text{for } g = 1, \ldots, G. \tag{17}$$

The resulting constrained EM is monotone as Equation (17) is a maximum of the likelihood of Equation (3) subject to the constraints (11). The same rule has been used, among others, in Ingrassia and Rocci [13], and Won et al. [29]. In Won et al. [29] for instance, in a context of maximum likelihood estimation of the covariance matrix under a set of constraints, Equation (17) is the optimization step for each eigenvalue of the covariance matrix.

It should be noted that the update in (17) for the constrained algorithm takes $\xi^2$ as an input parameter, which has been obtained outside of the constrained EM.

### 3.4. *Adaptive choice of c*

The goal is to choose a $c$, as we have seen above, which delivers the best compromise model between the common-variance model – too restrictive – and the totally unconstrained model – ill-conditioned. Phillips [20] showed the consistency, asymptotic normality and efficiency of the maximizer of the constrained likelihood function for a fixed $c$ in a switching regression context for Hathaway's constraints. Xu et al. [30], in the same context, extended the result of Phillips [20] for $c$ decreasing to zero as the sample size increases. Nevertheless, how to effectively choose $c$ in finite samples is an open issue.

Selecting $c$ jointly with the mixture parameters by maximizing the likelihood on the entire sample would trivially yield an overfitted scale balance approaching zero (RGD, 2017).

A practical way to select $c$, solving the issue of overfitting, would be to estimate, for a given $c$, the model parameters on a subset of the data – the training set. Then, select $c$ such that the log-likelihood of the remaining observations – test set – is maximized.

Smyth [26,27] advocates the use of the test set log-likelihood for selecting the number of mixture components. The rationale is that it can be shown to be an unbiased estimator (within a constant) of the Kullback–Leibler divergence between the *truth* and the model under consideration. As large test sets are hardly available, the cross-validated log-likelihood can be used to estimate the test set log-likelihood. In our case – like in Smyth's case [26] – given the model parameters, the cross-validated log-likelihood is a function of $c$ only, and maximizing it with respect to $c$, given the other model parameters, would handle the issue of overfitting as training and test sets are independent [1].

Let us consider $K$ random partitions of the data, where at each partition the data are split into a training set $(y, \mathbf{x})_S = \left\{(y_i, \mathbf{x}_i); i \in S\right\}_{n_S}$, and a test set $(y, \mathbf{x})_{\bar{S}} = \left\{(y_i, \mathbf{x}_i); i \in \bar{S}\right\}_{n_{\bar{S}}}$, and we indicate the entire data as $\left\{(y_i, \mathbf{x}_i); i \in N\right\}_n$, where $S \cup \bar{S} = N$, and $n_S + n_{\bar{S}} = n$. Let us denote by $\hat{\boldsymbol{\psi}}(c, S_k)$ the constrained ML estimate obtained on the training set at the $k$-th partition, and by $\ell_{\bar{S}_k}[\hat{\boldsymbol{\psi}}(c, S_k)]$ the log-likelihood evaluated at the test set. The cross-validated log-likelihood is defined as

$$\mathrm{CV}(c) = \sum_{k=1}^{K} \ell_{\bar{S}_k}[\hat{\boldsymbol{\psi}}(c, S_k)], \tag{18}$$

that is, as the sum of the contributions of all $K$ test sets to the log-likelihood.

The cross-validation strategy can be summarized according to the following steps (see also RGD, 2017).

1. Select a plausible value for $c \in (0, 1]$.
2. Obtain a temporary estimate $\hat{\boldsymbol{\psi}}$ for the model parameters using the entire sample, which is used as starting point for the cross-validation procedure.
3. Randomly partition the full data set into a training set and a test set.
4. Estimate the parameters on the training set using the starting point obtained in step 2. Compute the contribution to the log-likelihood of the test set.
5. Repeat $K$ times steps 3–4 and sum the contributions of the test sets to the log-likelihood (cross-validated log-likelihood).
6. Repeat steps 3–5 for different values of $c$.
7. Select $c$ which maximizes the cross-validated log-likelihood.

As from Equation (10), scale changes in the response variable simply add a constant to the log-likelihood and, with scale equivariant constraints, $c$ is left unaltered.

### 3.5. How to choose $\xi^2$

Imposing lower (very small) bounds to the component variances and no upper bound, as it is commonly done among practitioners, amounts to using our method with a very small $c$ and $\xi^2 = 1$. However, one can possibly obtain more accurate estimates by using larger values of $c$ – hence with tighter bounds. Yet, the bounds would be – perhaps unreasonably – still centered at 1.

In Subsection 3.2 we have seen that the method used to select $\xi^2$ has to be scale equivariant in order to ensure the scale equivariance of the constraints. The most natural scale equivariant candidate for the target is the homoscedastic normal variance. The reason is that, for $c = 1$ and the variance of a homoscedastic mixture of conditional normals as target, the method would simply estimate the homoscedastic mixture of conditional normals model. Nonetheless, if another scale equivariant method were used to estimate the target, the constrained method, for $c = 1$, would estimate a model with common variance equal to the target value, but all other parameters would be estimated at their – conditional normal – ML values.

The (target) homoscedastic normal variance is computed by using a similar EM algorithm as above. In the homoscedastic case, the only difference with respect to the heteroscedastic model is in the update for the component variances. Equation (16) is replaced, in the M-step for the estimation of the model parameters of the homoscedastic clusterwise linear regression, by the following update for the variance term

$$\sigma^{2(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}^{(k+1)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_g^{(k+1)})^2, \tag{19}$$

where we drop the index $g$ due to the common variance model assumption.

Other choices of target variance can also be considered, provided that the method used to estimate the target from the data is itself scale equivariant, hence automatically replaces $\xi^2$ with $\xi^{*2}$ if the data are transformed.

## 4. Numerical studies

### 4.1. Design

Two simulation studies were conducted in order to evaluate the quality of the parameter estimates of our constrained algorithm (ConC). The latter was compared with the unconstrained algorithm with common (homoscedastic) component-scales (HomN), and the unconstrained algorithm with different (heretoscedastic) component-scales (HetN). The target measures used for the comparisons were average Mean Squared Errors (MSE) of the regression coefficients (averaged across regressors and groups), and MSE of the component variances (averaged across groups). These measures will allow us to state the precision of the estimates. We measured how well the algorithms were able to classify sample units within clusters through the adjusted *Rand* index [11]. Average computation time in seconds is also reported for each method, and the CPU time reported for ConC includes cross-validation running time.

In the first simulation study, the data were generated from a clusterwise linear regression with 3 regressors and intercept, with 2 and 3 components and sample sizes of 100 and 200. The class proportions considered were, respectively, $(0.5, 0.5)'$ and $(0.2, 0.8)'$, and $(0.2, 0.4, 0.4)'$ and $(0.1, 0.3, 0.6)'$. Regressors have been drawn from 3 independent standard normals, whereas regression coefficients have been drawn from $\mathrm{U}(-1.5, 1.5)$ and intercepts are $(4, 9)'$ and $(4, 9, 16)'$ for the 2-component and 3-component models respectively. The component variances have been set equal to $(0.5, 0.5)'$, $(0.2, 0.6)'$, and $(0.1, 0.8)'$ in the two-component setups, and to $(0.5, 0.5, 0.5)'$, $(0.2, 0.6, 0.2)'$, and $(0.1, 0.8, 0.1)'$ in the three-component setups. This yielded variance ratios (smallest variance over the biggest) of respectively 0.125 (heteroscedastic components), 0.333 (mildly heteroscedastic components) and 1 (homoscedastic components).

For each of the 24 combinations sample size $\times$ class proportions $\times$ variance ratios, we generated 250 samples. For each sample, each algorithm – HomN, HetN, ConC (our proposal) – is initialized from the same 9 random fuzzy assignments and one rational assignment. We obtain random fuzzy partitions by drawing entries for the $n \times G$ matrix of the posterior probabilities from a $\mathrm{U}(0, 1)$ and then normalizing by row. The rational start is obtained by fitting an OLS regression of $y$ on $\mathbf{X}$, and then assigning the sample units to the clusters according to the percentile of the fitted residuals distribution. The solution selected – out of the 10 starts – is the one delivering the highest likelihood.[1]

In the second simulation study, we have increased the number of clusters, the number of regressors and the sample size, and assessed the method in a reduced set of scenarios, where we varied only the variance ratios – 0.125, 0.333 and 1 as above. We generated samples of 300 units from a clusterwise linear regression with 7 regressors and intercept, with 7 components, each respectively with proportions of 0.14, 0.14, 0.14, 0.14, 0.14, 0.14, and 0.16. Regressors were generated as above, whereas the component variances are set to $(0.2, 0.6, 0.2, 0.6, 0.2, 0.6, 0.2)'$, $(0.1, 0.8, 0.1, 0.8, 0.1, 0.8, 0.1)'$, and $(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)'$.

[1] Computer programs are available from the corresponding author upon request.

**Table 1**
Mixing proportions (0.5, 0.5), 250 samples, $n = 100$ and $n = 200$, scale ratios of 1, 0.333 and 0.125, 10 starts, 3 regressors and intercept. Values averaged across samples. Standard deviations in parentheses. Computation time in seconds.

| Algorithm | Mixing proportions $= (0.5, 0.5)'$ | | | | |
|---|---|---|---|---|---|
| | Avg MSE $\hat{\boldsymbol{\beta}}$ | Avg MSE $\hat{\boldsymbol{\sigma}}^2$ | Adj-Rand | Time | $c$ |
| $n = 100$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0120 (0.0081) | 0.0040 (0.0052) | 0.9608 (0.0492) | 0.0365 (0.0131) | – |
| HetN | 0.0121 (0.0081) | 0.0071 (0.0073) | 0.9565 (0.0539) | 0.0392 (0.0154) | – |
| ConC | 0.0120 (0.0081) | 0.0046 (0.0057) | 0.9599 (0.0510) | 1.1840 (0.1768) | 0.9364 (0.1053) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0099 (0.0071) | 0.0299 (0.0044) | 0.9564 (0.0512) | 0.0330 (0.0101) | – |
| HetN | 0.0097 (0.0070) | 0.0057 (0.0068) | 0.9748 (0.0384) | 0.0343 (0.0127) | – |
| ConC | 0.0097 (0.0069) | 0.0071 (0.0072) | 0.9744 (0.0381) | 1.1575 (0.1260) | 0.4369 (0.1660) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0120 (0.0096) | 0.0879 (0.0060) | 0.9399 (0.0576) | 0.0341 (0.0115) | – |
| HetN | 0.0108 (0.0084) | 0.0061 (0.0071) | 0.9799 (0.0327) | 0.0337 (0.0139) | – |
| ConC | 0.0109 (0.0085) | 0.0068 (0.0089) | 0.9791 (0.0345) | 1.1379 (0.1103) | 0.0968 (0.0731) |
| $n = 200$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0058 (0.0034) | 0.0015 (0.0021) | 0.9681 (0.0336) | 0.0859 (0.0447) | – |
| HetN | 0.0058 (0.0035) | 0.0030 (0.0028) | 0.9665 (0.0349) | 0.0971 (0.1353) | – |
| ConC | 0.0058 (0.0034) | 0.0018 (0.0022) | 0.9677 (0.0333) | 3.8566 (0.4982) | 0.9450 (0.0764) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0047 (0.0032) | 0.0281 (0.0018) | 0.9618 (0.0388) | 0.0923 (0.1911) | – |
| HetN | 0.0046 (0.0031) | 0.0022 (0.0025) | 0.9827 (0.0231) | 0.0877 (0.0881) | – |
| ConC | 0.0046 (0.0031) | 0.0026 (0.0029) | 0.9836 (0.0236) | 3.8728 (0.2437) | 0.3372 (0.1214) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0056 (0.0043) | 0.0873 (0.0045) | 0.9492 (0.0433) | 0.0831 (0.0489) | – |
| HetN | 0.0052 (0.0041) | 0.0024 (0.0032) | 0.9895 (0.0176) | 0.0814 (0.0585) | – |
| ConC | 0.0052 (0.0040) | 0.0025 (0.0032) | 0.9895 (0.0173) | 3.9150 (0.2414) | 0.0678 (0.0269) |

Concerning the choice of the number of random partitions $K$ and the size of the training-test set in the cross-validation strategy, Smyth [27] argues that $K$ between 20 and 50 is adequate for most applications, whereby any relative size of the test set between 0.1 and 0.5, as found by van der Laan et al. [28], works equally well. We choose $K = n/2$ and a training-set size $n_S = n - \frac{n}{5}$. A similar setting was also used in RGD (2017), where they tested also different settings and found that the final results were not sensitive to changes in the number of random partitions or in the relative size of training-test set – provided that all components are represented in the training set. The constrained approach was initialized with a non-informative value ($c = 1/10000$), in order to avoid starting off with a degenerate solution.

### 4.2. Results

Results for the 24 simulation conditions of the first simulation study are summarized in Tables 1, 2, 3 and 4. In the two-component condition, with components of equal sizes (Table 1) and $n = 100$, HomN is the most accurate in the case where the variance ratio equals 1. Our constrained estimator delivers statistically equal numbers. With smaller variance scales (0.333 and 0.125), HetN takes over, closely followed by ConC, again delivering statistically equal results compared to the most accurate of the two unconstrained approaches. Differences among the approaches tend to fade as $n = 200$. In Table 2 (class proportions of 0.2 and 0.8), we observe similar results in the condition with data generated from components with equal common scale. ConC, with $n = 100$, slightly improves over the unconstrained approaches when the variance ratio equals 0.333 or 0.125, and does as well as HomN in the variance ratio $= 1$ case. Again, differences almost vanish for $n = 200$. Overall, we observe from Tables 1 and 2 that, when either of the two unconstrained approaches is optimal, ConC keeps up closely. In this respect, also the average selected constant $c$ nicely displays how close to each of the two extremes the final constrained result is.

In the three-component scenarios (Tables 3 and 4), results for the unconstrained approaches are somewhat tarnished, especially with uneven mixing proportions $(0.1, 0.3, 0.6)'$. In the latter case, in which ConC is the most accurate estimator, it is from twice up to twenty times more accurate than the second-best approach. Compared to the two-component scenarios, also the average selected $c$ tends to be further away from the extremes (particularly with uneven mixing proportions), showing that there is actual room for significantly improving the fit of the unconstrained approaches in all considered scenarios.

**Table 2**
Mixing proportions $(0.2, 0.8)$, 250 samples, $n = 100$ and $n = 200$, variance ratios of 1, 0.333 and 0.125, 10 starts, 3 regressors and intercept. Values averaged across samples. Standard deviations in parentheses. Computation time in seconds.

| Algorithm | Mixing proportions $= (0.2, 0.8)'$ | | | | |
|---|---|---|---|---|---|
| | Avg MSE $\hat{\boldsymbol{\beta}}$ | Avg MSE $\hat{\boldsymbol{\sigma}}^2$ | Adj-Rand | Time | $c$ |
| $n = 100$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0280 (0.0247) | 0.0042 (0.0058) | 0.9723 (0.0394) | 0.0407 (0.0148) | – |
| HetN | 0.0328 (0.0426) | 0.0275 (0.1089) | 0.9694 (0.0438) | 0.0381 (0.0142) | – |
| ConC | 0.0286 (0.0259) | 0.0059 (0.0090) | 0.9718 (0.0401) | 1.1312 (0.1019) | 0.9026 (0.2047) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0162 (0.0161) | 0.0361 (0.0100) | 0.9648 (0.0450) | 0.0426 (0.0166) | – |
| HetN | 0.0178 (0.0299) | 0.0221 (0.1427) | 0.9679 (0.0448) | 0.0392 (0.0162) | – |
| ConC | 0.0149 (0.0130) | 0.0125 (0.0134) | 0.9688 (0.0445) | 1.1575 (0.1181) | 0.4365 (0.3202) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0159 (0.0183) | 0.1172 (0.0245) | 0.9422 (0.0581) | 0.0464 (0.0196) | – |
| HetN | 0.0160 (0.0365) | 0.0345 (0.2240) | 0.9596 (0.0534) | 0.0418 (0.0167) | – |
| ConC | 0.0114 (0.0096) | 0.0093 (0.0126) | 0.9647 (0.0467) | 1.1714 (0.1081) | 0.0812 (0.1112) |
| $n = 200$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0113 (0.0081) | 0.0015 (0.0021) | 0.9842 (0.0230) | 0.1028 (0.0425) | – |
| HetN | 0.0113 (0.0082) | 0.0057 (0.0063) | 0.9824 (0.0238) | 0.0892 (0.0252) | – |
| ConC | 0.0113 (0.0081) | 0.0023 (0.0034) | 0.9841 (0.0232) | 3.8962 (0.2141) | 0.9282 (0.1422) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0067 (0.0049) | 0.0369 (0.0073) | 0.9728 (0.0294) | 0.1020 (0.0431) | – |
| HetN | 0.0058 (0.0037) | 0.0029 (0.0030) | 0.9807 (0.0232) | 0.0882 (0.0216) | – |
| ConC | 0.0059 (0.0037) | 0.0033 (0.0038) | 0.9813 (0.0232) | 3.9141 (0.2447) | 0.2300 (0.1272) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0066 (0.0048) | 0.1203 (0.0174) | 0.9520 (0.0381) | 0.1106 (0.0597) | – |
| HetN | 0.0046 (0.0030) | 0.0023 (0.0024) | 0.9746 (0.0290) | 0.0936 (0.0221) | – |
| ConC | 0.0046 (0.0030) | 0.0025 (0.0027) | 0.9772 (0.0268) | 3.9638 (0.2126) | 0.0367 (0.0220) |

In the second simulation study, the above results are even more evident. Interestingly however, we observe that, where the variance ratio equals 1, although HomN accurately estimates the component common variance, ConC (with an average $c \approx 0.83$) improves in both MSE of the betas and cluster recovery (Table 5).

## 5. Four real data applications

The aim is to show, through the four real data applications we present in this section, that the method works well in terms of quality of model parameter estimates and classification. Whereas one can compare true with estimated parameter values in a simulation study (where the data generating process is known), this is not possible in real data applications, where possibly also the number of clusters in the sample is unknown. What we wish to find is a method being able to handle a number of groups larger than what is best for the data at hand – which is typical in exploratory stages of the analysis – without being unreasonably drawn by the spurious solutions such cases are likely to deliver. If this is a feature the method possesses, automatic model selection procedures can be used without any concern.

Having this in mind, we estimated a clusterwise linear regression, using the 3 methods under comparison, on the *CEO* data set (http://lib.stat.cmu.edu/DASL/DataArchive.html), with 2, 3 and 4 components, assessing the plausibility of the estimated regression lines. We carried out a similar exercise on the *Temperature* data set ([17]; available at http://rcarbonneau.com/ClusterwiseRegressionDatasets/data_USTemperatures.txt), where we analyzed and compared clusterwise linear regression models with 2, 3, 4 and 5 components. We compared the three methods also using the *Auto-Mpg* data set (https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data), fitting mixtures of 2, 3, 4 and 5 components.

Finally, we estimated a 3-component clusterwise linear regression model on Fisher's Iris data, using petal width as dependent variable and sepal width as explanatory variable, in order to assess the clusters recovery.

In all applications the estimated groups are ordered from the smallest to the largest in terms of cluster size.

The constrained approach was initialized with a non-informative value ($c = 1/10000$), in order to avoid starting off with a degenerate solution.

**Table 3**
Mixing proportions $(0.2, 0.4, 0.4)$, 250 samples, $n = 100$ and $n = 200$, variance ratios of 1, 0.333 and 0.125, 10 starts, 3 regressors and intercept. Values averaged across samples. Standard deviations in parentheses. Computation time in seconds.

| Algorithm | Mixing proportions $= (0.2, 0.4, 0.4)'$ | | | | |
|---|---|---|---|---|---|
| | Avg MSE $\hat{\boldsymbol{\beta}}$ | Avg MSE $\hat{\boldsymbol{\sigma}}^2$ | Adj-Rand | Time | $c$ |
| $n = 100$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0590 (0.3771) | 0.0085 (0.0375) | 0.9702 (0.0392) | 0.1335 (0.0916) | – |
| HetN | 0.0721 (0.3834) | 0.0369 (0.1592) | 0.9611 (0.0583) | 0.1259 (0.0658) | – |
| ConC | 0.0269 (0.0228) | 0.0062 (0.0069) | 0.9704 (0.0382) | 1.8136 (0.4145) | 0.9030 (0.1627) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0304 (0.1892) | 0.0285 (0.0291) | 0.9660 (0.0404) | 0.1218 (0.0166) | – |
| HetN | 0.0460 (0.2339) | 0.0396 (0.2107) | 0.9675 (0.0496) | 0.1101 (0.0162) | – |
| ConC | 0.0151 (0.0109) | 0.0084 (0.0081) | 0.9754 (0.0320) | 1.7223 (0.1181) | 0.4914 (0.1788) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0458 (0.2827) | 0.0850 (0.0402) | 0.9473 (0.0532) | 0.1311 (0.1279) | – |
| HetN | 0.0554 (0.2640) | 0.0580 (0.2837) | 0.9695 (0.0577) | 0.1204 (0.1074) | – |
| ConC | 0.0223 (0.1397) | 0.0094 (0.0250) | 0.9770 (0.0362) | 1.8079 (0.4384) | 0.1598 (0.0847) |
| $n = 200$ | | | | | |
| | Variance ratio $= 1$ | | | | |
| HomN | 0.0113 (0.0066) | 0.0017 (0.0023) | 0.9866 (0.0176) | 0.3049 (0.7659) | – |
| HetN | 0.0113 (0.0065) | 0.0054 (0.0049) | 0.9845 (0.0193) | 0.2527 (0.1127) | – |
| ConC | 0.0112 (0.0065) | 0.0024 (0.0031) | 0.9864 (0.0176) | 5.5820 (0.7345) | 0.9312 (0.1090) |
| | Variance ratio $= 0.333$ | | | | |
| HomN | 0.0067 (0.0039) | 0.0254 (0.0029) | 0.9803 (0.0192) | 0.2827 (0.5192) | – |
| HetN | 0.0349 (0.2976) | 0.0238 (0.1886) | 0.9816 (0.0409) | 0.2430 (0.1245) | – |
| ConC | 0.0065 (0.0037) | 0.0037 (0.0035) | 0.9861 (0.0170) | 5.5699 (0.4304) | 0.4205 (0.1053) |
| | Variance ratio $= 0.125$ | | | | |
| HomN | 0.0067 (0.0045) | 0.0813 (0.0086) | 0.9639 (0.0284) | 0.2707 (0.3522) | – |
| HetN | 0.0124 (0.1079) | 0.0091 (0.0960) | 0.9879 (0.0231) | 0.2348 (0.1047) | – |
| ConC | 0.0056 (0.0035) | 0.0030 (0.0035) | 0.9893 (0.0157) | 5.5371 (0.4026) | 0.1157 (0.0482) |

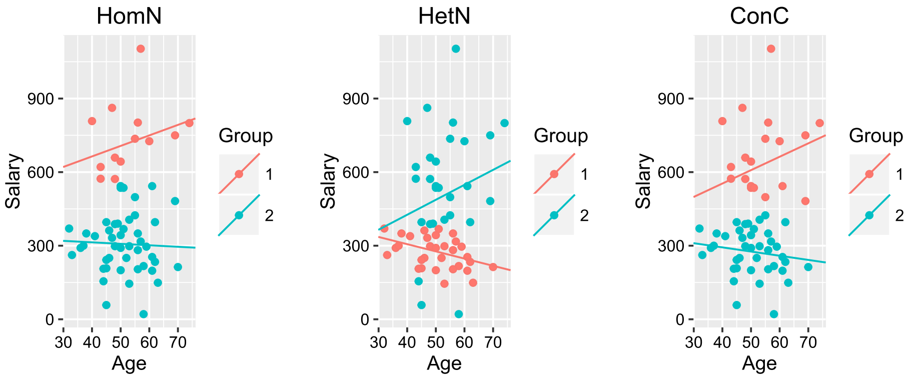

**Fig. 1.** *CEO* data. Best solutions out of 50 random starts, $G = 2$. $K = n/5$, and test set size $= n/10$.

### 5.1. CEO data

This data set has a well-known structure, although a *true* number of clusters is not available. It contains 59 records for some U.S. small-companies' CEO salaries (dependent variable), and CEO ages (independent variable).

Bagirov et al. [2] fitted a 2-component and a 4-component clusterwise linear regression, whereas Carbonneau et al. [3] focused on the perhaps most intuitive 2-component setup. We fitted respectively 2-component (Fig. 1), 3-component (Fig. 2), and 4-component (Fig. 3) clusterwise linear regressions, and graphically compared the regression lines and crisp classifications obtained.

The 2-class solution (Fig. 1) would be favored by HomN and ConC in terms of BIC, whereas HetN would select a 4-class model (Fig. 3) as the one minimizing the BIC. Results on BIC values are summarized in Table 6.

As shown in Fig. 1, HomN and HetN deliver different solutions in terms of both estimated regression lines and clustering. Interestingly, ConC, with a selected $c$ of about 0.3, yields a solution which is in between homoscedasticity and heteroscedas-

**Table 4**
Mixing proportions $(0.1, 0.3, 0.6)$, 250 samples, $n = 100$ and $n = 200$, variance ratios of 1, 0.333 and 0.125, 10 starts, 3 regressors and intercept. Values averaged across samples. Standard deviations in parentheses. Computation time in seconds.

| Algorithm | Mixing proportions $= (0.1, 0.3, 0.6)'$ | | | | |
|---|---|---|---|---|---|
| | Avg MSE $\hat{\boldsymbol{\beta}}$ | Avg MSE $\hat{\boldsymbol{\sigma}}^2$ | Adj-Rand | Time | $c$ |
| $n = 100$ | | | | | |
| | Variance ratio = 1 | | | | |
| HomN | 0.9209 (1.6722) | 0.1015 (0.2024) | 0.9463 (0.0596) | 0.1197 (0.1069) | – |
| HetN | 0.6292 (2.0556) | 0.1295 (0.2910) | 0.9608 (0.0595) | 0.1096 (0.0382) | – |
| ConC | 0.3024 (0.8933) | 0.0473 (0.1371) | 0.9698 (0.0656) | 2.0553 (0.5606) | 0.3545 (0.3927) |
| | Variance ratio = 0.333 | | | | |
| HomN | 0.7861 (1.6161) | 0.1317 (0.2349) | 0.9486 (0.0555) | 0.1189 (0.0787) | – |
| HetN | 0.5338 (1.5861) | 0.1812 (0.4002) | 0.9597 (0.0608) | 0.1090 (0.0388) | – |
| ConC | 0.2308 (0.6126) | 0.0801 (0.2358) | 0.9711 (0.0557) | 2.0744 (0.7441) | 0.2286 (0.2867) |
| | Variance ratio = 0.125 | | | | |
| HomN | 0.6558 (1.5973) | 0.1878 (0.2516) | 0.9387 (0.0550) | 0.1188 (0.0860) | – |
| HetN | 0.8355 (3.9024) | 0.3064 (0.5605) | 0.9497 (0.0678) | 0.1100 (0.0410) | – |
| ConC | 0.2901 (0.8464) | 0.1239 (0.3384) | 0.9624 (0.0733) | 2.0304 (0.6606) | 0.1005 (0.1207) |
| $n = 200$ | | | | | |
| | Variance ratio = 1 | | | | |
| HomN | 0.3547 (0.5893) | 0.2570 (0.3731) | 0.9421 (0.0717) | 0.2272 (0.1252) | – |
| HetN | 0.1332 (0.5473) | 0.0720 (0.2854) | 0.9847 (0.0317) | 0.3535 (0.1502) | – |
| ConC | 0.0300 (0.1176) | 0.0092 (0.0698) | 0.9920 (0.0122) | 7.8230 (3.0895) | 0.1859 (0.3218) |
| | Variance ratio = 0.333 | | | | |
| HomN | 0.3573 (0.7375) | 0.2949 (0.4153) | 0.9455 (0.0674) | 0.2158 (0.1366) | – |
| HetN | 0.1369 (0.5990) | 0.0850 (0.3425) | 0.9864 (0.0313) | 0.3165 (0.1536) | – |
| ConC | 0.0136 (0.0488) | 0.0066 (0.0384) | 0.9932 (0.0110) | 7.0303 (1.8301) | 0.0851 (0.1624) |
| | Variance ratio = 0.125 | | | | |
| HomN | 0.3359 (0.7195) | 0.3448 (0.4307) | 0.9397 (0.0635) | 0.2524 (0.4298) | – |
| HetN | 0.2677 (0.7267) | 0.2294 (0.5715) | 0.9724 (0.0526) | 0.2883 (0.1363) | – |
| ConC | 0.0138 (0.0553) | 0.0169 (0.1261) | 0.9914 (0.0219) | 6.7318 (2.1352) | 0.0335 (0.0638) |

**Table 5**
250 samples, $n = 300$, mixing proportions $(0.14, 0.14, 0.14, 0.14, 0.14, 0.14, 0.16)$, variance ratios of 1, 0.333 and 0.125, 10 starts, 7 regressors and intercept. Values averaged across samples. Standard deviations in parentheses. Computation time in seconds.

| Algorithm | Mixing proportions $= (0.14, 0.14, 0.14, 0.14, 0.14, 0.14, 0.16)'$ | | | | |
|---|---|---|---|---|---|
| | Avg MSE $\hat{\boldsymbol{\beta}}$ | Avg MSE $\hat{\boldsymbol{\sigma}}^2$ | Adj-Rand | Time | $c$ |
| | Variance ratio = 1 | | | | |
| HomN | 0.5051 (3.4355) | 0.0755 (0.3375) | 0.9448 (0.0829) | 5.8671 (1.2955) | – |
| HetN | 0.2433 (0.8563) | 0.2718 (0.7052) | 0.9373 (0.0737) | 5.3796 (1.1836) | – |
| ConC | 0.2270 (2.1530) | 0.1029 (0.5218) | 0.9557 (0.0621) | 61.1625 (6.3034) | 0.8353 (0.3083) |
| | Variance ratio = 0.333 | | | | |
| HomN | 0.1813 (0.5335) | 0.0952 (0.2286) | 0.9492 (0.0716) | 5.8159 (0.9937) | – |
| HetN | 0.3043 (1.4071) | 0.2427 (0.8051) | 0.9557 (0.0768) | 5.2950 (0.8622) | – |
| ConC | 0.0686 (0.2809) | 0.0638 (0.3226) | 0.9720 (0.0409) | 61.5495 (4.1174) | 0.4289 (0.1835) |
| | Variance ratio = 0.125 | | | | |
| HomN | 0.2475 (1.1008) | 0.1604 (0.2646) | 0.9451 (0.0757) | 5.8086 (1.1235) | – |
| HetN | 0.2311 (1.2021) | 0.2282 (0.7504) | 0.9613 (0.0693) | 5.2996 (0.8895) | – |
| ConC | 0.0879 (0.2883) | 0.0878 (0.3241) | 0.9750 (0.0455) | 60.8264 (4.1667) | 0.1530 (0.0751) |

**Table 6**
*CEO* data. BIC values for $G = 2$, $G = 3$, and $G = 4$, computed under HomN, HetN, and ConC. Best solutions out of 100 random starts.

| | $G = 2$ | $G = 3$ | $G = 4$ |
|---|---|---|---|
| $\mathrm{BIC}_{\mathrm{HomN}}$ | 706.30 | 712.14 | 719.41 |
| $\mathrm{BIC}_{\mathrm{HetN}}$ | 704.42 | 707.70 | 599.52 |
| $\mathrm{BIC}_{\mathrm{ConC}}$ | 706.86 | 721.13 | 740.92 |

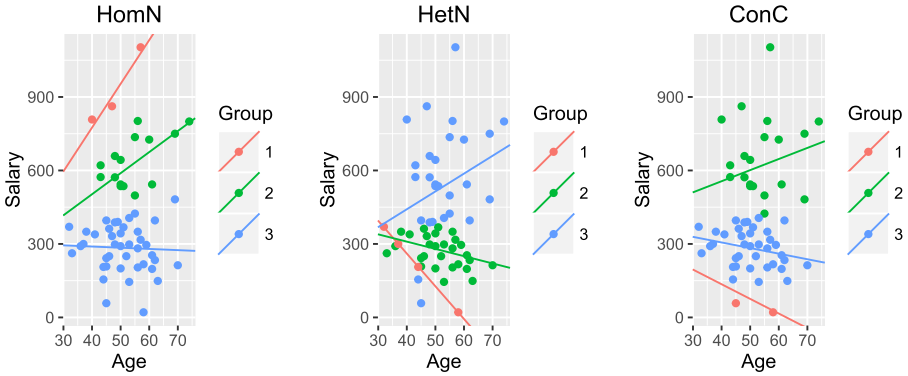

**Fig. 2.** *CEO* data. Best solutions out of 50 random starts, $G = 3$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

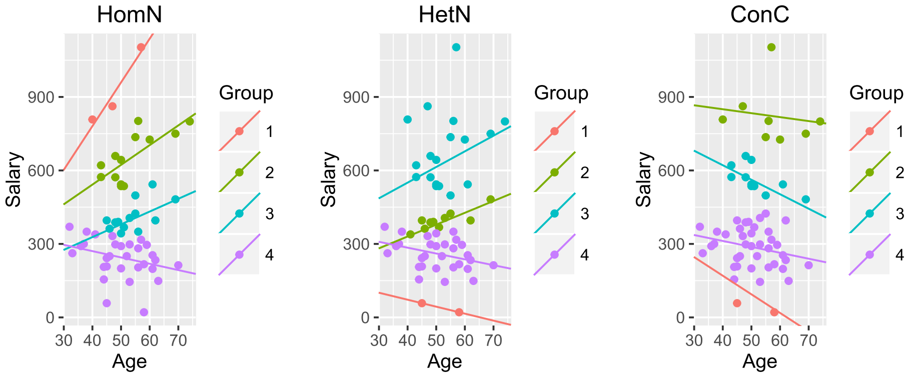

**Fig. 3.** *CEO* data. Best solutions out of 50 random starts, $G = 4$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

ticity: this is the case not only in terms of estimated component-scales, as expected, but also in terms of fitted regression lines and clustering. In other words, the proposed procedure has the very nice feature of letting the data decide the most appropriate model as the most suited-to-the-data compromise between homoscedasticity and heteroscedasticity.

A similar soft solution, in between homoscedasticity and heteroscedasticity, is delivered also with 3 (Fig. 2) and 4 mixture components (Fig. 3). In particular, the ConC's solution with 3 components is closer to HetN's than HomN's ($c \approx 0.15$), although we observe some departure, especially in the first (red) component: indeed HetN red group's regression line crosses the 4 units assigned to that component. This would signal the spurious nature of the solution delivered by HetN. On the other hand, HomN's solution, especially for the first component, seems to be driven by the unit with the highest salary.

As we turn to the 4 components case, ConC yields results closer to HomN ($c \approx 0.9$). In the latter case we observe a solution for HetN which is very likely to be spurious, as the first component's regression line (in red) is aligned with two data points, and the relative component variance is relatively very small (the scale ratio between first and second component is $< 10^{-11}$). As for the 3-component case, HomN first component's solution seems to be driven by the unit with the highest salary.

It is interesting to note that, for $G \geq 3$, also the solutions delivered by ConC seem spurious. The fact that this happens to be the case, combined with the evidence from Table 6 – where ConC selects two components – confirms that the proposed method can handle a larger than what is best for the data number of groups, but being still able to deliver the seemingly correct (2-component) solution.

### 5.2. Temperature *data*

This data set concerns average minimum temperatures in 56 US cities in January, including latitude and longitude of each city.[2] Among the others who already analyzed these data, Peixoto [19] fitted a polynomial regression of temperature on latitude and a cubic polynomial in longitude to this data set, whereas Carbonneau and co-authors [3] fitted a 2-component clusterwise linear regression of temperature on latitude and longitude. We fitted a clusterwise linear regression model

[2] The time span considered for taking the average is January 1930 – January 1960. For further information on how average minimum temperatures are obtained please refer to Peixoto [19].

**Table 7**
*Temperature* data. BIC values for $G = 2$, $G = 3$, $G = 4$, and $G = 5$, computed under HomN, HetN, and ConC. Best solutions out of 100 random starts.

| | $G = 2$ | $G = 3$ | $G = 4$ | $G = 5$ |
|---|---|---|---|---|
| $\text{BIC}_{\text{HomN}}$ | 257.98 | 263.33 | 261.81 | 247.06 |
| $\text{BIC}_{\text{HetN}}$ | 257.44 | 256.51 | 251.30 | 107.42 |
| $\text{BIC}_{\text{ConC}}$ | 257.52 | 264.61 | 262.80 | 261.12 |

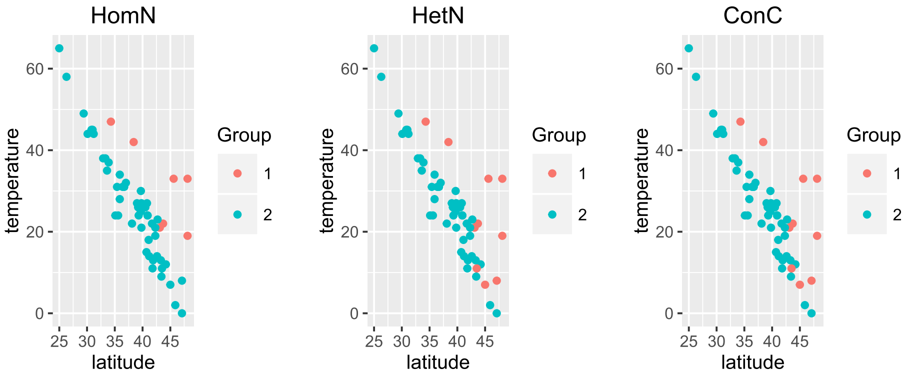

**Fig. 4.** *Temperature* data. Best solutions out of 100 random starts, $G = 2$. $K = n/5$, and test set size $= n/10$.

of temperature on latitude and longitude with respectively 2, 3, 4, and 5 components (all Tables with estimated model parameters can be found in the Appendix). For obvious reasons, latitude is more informative than longitude to determine the final clustering. This is why we will plot the clustering results focusing on temperatures and latitudes only.

The 2-component solution is the one chosen by minimizing $\text{BIC}_{\text{ConC}}$. We argue that such a solution be the most appropriate one, although BIC computed under both HetN and HomN seems to favor the 5-component model (Table 7). Nonetheless, due to the spurious nature of the final solution delivered by HetN, $\text{BIC}_{\text{HetN}}$ should not be trusted. This is seemingly also the case for $\text{BIC}_{\text{HomN}}$, as the related solution is characterized by a too small class proportion for one component compared to the others ($p_1 = 0.0881$), and a small common scale (approximately half of that estimated with $G = 4$).

The 2-class solution seems to be the most suitable in terms of non-overlapping classes and regression parameters' interpretation. In both classes, latitude has a negative effect on temperature, whereas longitude has a negative effect on temperature in the first (smaller) class, and a positive effect on the second (bigger) class. In the 3-class and the 4-class solutions (see Appendix A.1), the additional classes are mainly obtained from splits of the second (bigger) class: the resulting clusters are characterized by a negative sign for the longitude coefficient. In the 5-class solution, also the first (smaller) class is split into 2 sub-classes, both having the feature of positive sign for the longitude coefficient. In addition, we observe that HetN converged to a spurious solution, which consists in one component having a variance very close to zero. ConC, in all scenarios, estimates a model which is in between HetN and HomN: while being closer to HetN with $G = 2$ and $G = 4$, it gets relatively closer to the common scale with $G = 3$ and $G = 5$ (Fig. 4).

### 5.3. AutoMpg *data*

The data concern city-cycle fuel consumption in miles per gallon of a sample of 398 vehicles, to be predicted in terms of 7 covariates: number of cylinders, model year, and origin, which are discrete valued; displacement, horsepower, weight, and acceleration, which are instead continuous valued. Although the available data set has six missing values for horsepower, given that we have information on the car models with most relevant characteristics, we were able to back out the values and include them in the data set.

We fitted clusterwise linear regressions with intercept of miles per gallon on acceleration, cylinders, displacement, horsepower, model year, weight, and origin, with 2, 3, 4, and 5 components, comparing results for HomN, HetN, and ConC (see Appendix A.2). The most preferred solution for ConC is the 2-component mixture, whereas HomN and HetN select respectively 3 and 5 components (Table 8).

Spurious solutions are typically characterized by a small number of points in at least one cluster, which has a relatively small variance [18, p. 103]. The HetN solution with $G = 5$ appears to be spurious, with two relatively small groups – mixing proportions of 0.05 and 0.06 – with relatively small cluster variances – 0.02 and 0.05. Hence, also in this case, we cannot trust the model selection results of HetN. The 3-component solution selected by HomN has one very small component - mixing proportion of 0.06 – compared to the other two – 0.32 and 0.61 respectively (see Table A.15 in Appendix A.2). By looking at the 2-component solution (Table 9), which is indeed the one selected by minimizing $\text{BIC}_{\text{ConC}}$, we observe that the data seem to support a heteroscedastic structure, as ConC and HetN deliver two almost identical solutions. This would

**Table 8**
*Auto-Mpg* data. BIC values for $G = 2$, $G = 3$, $G = 4$, and $G = 5$, computed under HomN, HetN and ConC. Best solutions out of 100 random starts.

| | $G = 2$ | $G = 3$ | $G = 4$ | $G = 5$ |
|---|---|---|---|---|
| $\text{BIC}_{\text{HomN}}$ | 1365.89 | 1363.22 | 1381.12 | 1398.56 |
| $\text{BIC}_{\text{HetN}}$ | 1329.35 | 1340.55 | 1320.86 | 1319.93 |
| $\text{BIC}_{\text{ConC}}$ | 1329.74 | 1340.95 | 1364.76 | 1371.26 |

**Table 9**
*Auto-Mpg* data. Covariates are acceleration ($\mathbf{x}_1$), cylinders ($\mathbf{x}_2$), displacement ($\mathbf{x}_3$), horsepower ($\mathbf{x}_4$), model year ($\mathbf{x}_5$), weight ($\mathbf{x}_6$), and origin ($\mathbf{x}_7$). Best solutions out of 100 random starts, $G = 2$. $K = n/5$, and test set size $= n/10$.

| | HomN | | HetN | | ConC | |
|---|---|---|---|---|---|---|
| $p_g$ | 0.2215 | 0.7785 | 0.4473 | 0.5527 | 0.4353 | 0.5647 |
| Intercept | −35.0716 | −3.2278 | −23.3485 | 3.7071 | −23.5883 | 3.5861 |
| $\beta_{1g}$ | 0.1819 | −0.2530 | 0.1354 | −0.4212 | 0.1383 | −0.4177 |
| $\beta_{2g}$ | 1.1272 | −0.7172 | 0.1853 | −0.9055 | 0.1767 | −0.8938 |
| $\beta_{3g}$ | 0.0170 | 0.0004 | 0.0362 | −0.0116 | 0.0367 | −0.0116 |
| $\beta_{4g}$ | −0.2113 | −0.0077 | −0.1188 | −0.0035 | −0.1211 | −0.0031 |
| $\beta_{5g}$ | 1.1328 | 0.5862 | 0.9546 | 0.4699 | 0.9592 | 0.4730 |
| $\beta_{6g}$ | −0.0070 | −0.0042 | −0.0084 | −0.0022 | −0.0084 | −0.0022 |
| $\beta_{7g}$ | 0.6887 | 1.7958 | 0.7283 | 2.6872 | 0.7221 | 2.6430 |
| $\sigma_g^2$ | 2.3770 | 2.3770 | 3.1592 | 1.4190 | 3.1576 | 1.4906 |
| $c$ | – | – | – | – | 0.1547 | 0.1547 |

**Table 10**
*Iris* data. Adjusted Rand index, CPU time in seconds and selected $c$ for a 3-component clusterwise linear regression of petal width on sepal width. Best solution out of 500 random starts. $K = n/5$, and test set size $= n/10$.

| Algorithm | Adj-Rand | Time | $c$ |
|---|---|---|---|
| HomN | 0.5532 | 48.5605 | – |
| HetN | 0.4414 | 35.3266 | – |
| ConC | 0.8180 | 495.5422 | 0.0222 |

explain the 3-component result of HomN, which adapted to a seemingly non-spherical two-component structure by adding an extra component.

By looking at Table 9 we observe that acceleration ($\mathbf{x}_1$), cylinders ($\mathbf{x}_2$), and displacement ($\mathbf{x}_3$) have all positive effect on miles per gallon in the first (smaller) component, and negative effect in the second (larger) component. Cars with more horsepower, not surprisingly, tend to drive less miles per gallon – although the effect is relatively milder for the second (larger) component – whereby a more recent model year ($\mathbf{x}_5$), all else equal, is positively associated with miles per gallon in both components – again, with a relatively milder effect on the second component.

### 5.4. Iris *data*

We consider a subset of the Iris data, available at the link https://archive.ics.uci.edu/ml/index.html. Although this is a data set typically used for multivariate analysis, here we use it for illustrative purposes. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The clusters recovery obtained by the three methods is assessed in terms of Adj-Rand index. We also report computation time and the estimated $c$.

By looking at the true classification, the two upper groups (blue and green) seem to cluster around two parallel lines, whereas in the red group there seem to be no significant relation between Petal width and Sepal width. Only ConC – among the ones compared – is able to detect such a structure.

ConC yields a clustering which is the closest to the true classification compared to HomN and HetN (Fig. 5). The Adj-Rand obtained by ConC, as we observe from Table 10, is 0.82, whereas HomN and HetN obtain much lower values – 0.55 and 0.44 respectively. On the other hand, the computation time it takes for ConC to run, multiple starting value strategy included, is more than 10 times longer than HetN and HomN.

## 6. Conclusions

In the present paper, a scale equivariant soft constrained approach to maximum likelihood estimation of clusterwise linear regression model is formulated. This extends the approach proposed in RGD (2017) for multivariate mixtures of normals to the clusterwise linear regression context. Through the extensive simulation studies and the four empirical applications, we are able to show that the method does not only solve the issue of unboundedness, but it is also able to improve upon the unconstrained approaches it was compared with. Whenever either of the unconstrained approaches is instead optimal,

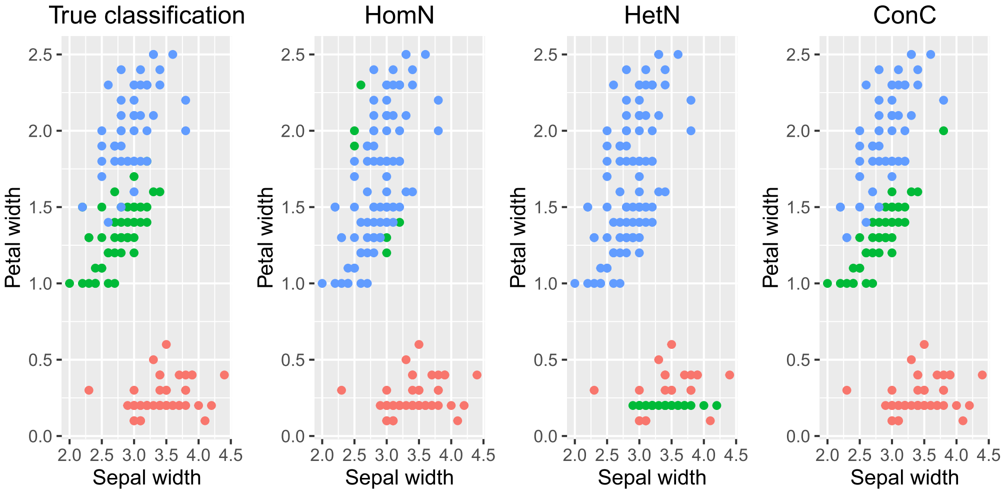

**Fig. 5.** *Iris* data. Best solutions out of 500 random starts, $G = 3$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

our constrained estimator keeps up closely. In addition, as pointed out by the empirical examples, the BIC based on the constrained solution is able to provide sensible choices for the number of clusters whereby the two unconstrained competitors cannot.

Whereas RGD (2017) showed that the method has merit in both fuzzy and crisp classification, the additional step ahead we take is twofold: 1) we show that the method works well also when conditional distributions (linear regressions) are specified for the mixture components, by looking at cluster recovery and by 2) bringing into focus, in our simulation study, bias and variance of the model parameter estimators, which are both entailed in the mean squared error.

Previous work on constrained estimation of switching regressions [20,30] had shown consistency of the estimator for fixed $c$, and for $c$ approaching zero with sample size going to infinity. We solve the finite sample problem by using a soft approach, which imposes *imprecise* bounds based on a cross-validated choice of $c$. The selected $c$ will produce an output model with clusterwise linear regressions and clustering as a most-suited-to-the-data solution in between the homoscedastic and the heteroscedastic models.

Our method shares common ground with the plain constrained maximum likelihood approach in that parameter updates are the same as in a constrained EM [12,13]. Nevertheless, the final solution we obtain maximizes the cross-validated log-likelihood, and the constraints are tuned on the data. This eliminates all unreasonable boundary solutions the standard constrained algorithm might converge to due to the arbitrary way constraints could be positioned.

The scale equivariance property in a clusterwise linear regression framework is related to linear transformations of the dependent variable only. Yet, it is as crucial as in the multivariate mixture of normals case, as it does not uniquely imply that the final clustering remains unaltered as one acts affine transformation on the variable of interest. More broadly, no matter how the data come in, scale equivariance means that there is no data transformation ensuring better results, since the method is unaffected by changes of scale in the response variable.

As it was noticed by Ritter [24], common scale is highly valuable, but it can be a too restrictive assumption for the clusters' scales. In this respect, our approach does not suffer the inappropriateness of the homoscedastic model, as the constant $c$ controls how close to (or how far from) it the final model will be. In both the numerical studies and the empirical applications, we observed that the method is able to detect departures from homoscedasticity in terms of selected $c$.

Other targets can be considered, perhaps more specific to the data at hand, provided that the method used to select $\xi^2$ from the data is also scale equivariant – hence replacing $\xi^2$ with $\xi^{*2}$ if the data are transformed. Indeed, this can be an interesting topic for future work.

The simulation study and the empirical applications have highlighted two related issues linked to the computation burden of the proposed method. ConC is indeed computationally intensive compared to the unconstrained approaches it was compared with. Building up a computationally more efficient procedure to select the constant $c$ from the data, and how to use it for model selection, can be both interesting topics for future research.

## Acknowledgements

The authors are grateful to the associate editor and the two anonymous referees, whose comments have lead to a substantial improvement of the paper.

## Appendix A. Additional tables and figures

### *A.1. Tables A.11–A.17 and Figs. A.6–A.8 for the* temperature *data*

**Table A.11**
*Temperature* data. Best solutions out of 100 random starts, $G = 2$. $K = n/5$, and test set size $= n/10$.

| | HomN | | HetN | | ConC | |
|---|---|---|---|---|---|---|
| $p_g$ | 0.2371 | 0.7629 | 0.2626 | 0.7374 | 0.2647 | 0.7353 |
| Intercept | 66.4905 | 142.2348 | 74.5325 | 150.6175 | 74.8657 | 150.5268 |
| $\beta_{1g}$ | −1.7204 | −2.5182 | −1.9829 | −2.5806 | −1.9945 | −2.5791 |
| $\beta_{2g}$ | 0.3353 | −0.2249 | 0.3434 | −0.2945 | 0.3460 | −0.2939 |
| $\sigma_g^2$ | 3.5602 | 3.5602 | 6.0948 | 2.7499 | 5.6950 | 2.7568 |
| $c$ | – | – | – | – | 0.1527 | 0.1527 |

**Table A.12**
*Temperature* data. Best solutions out of 100 random starts, $G = 3$. $K = n/5$, and test set size $= n/10$.

| | | HomN | |
|---|---|---|---|
| $p_g$ | 0.1091 | 0.1933 | 0.6975 |
| Intercept | 52.0354 | 96.7770 | 150.5393 |
| $\beta_{1g}$ | −1.2531 | −2.4057 | −2.5779 |
| $\beta_{2g}$ | 0.319 | 0.2735 | −0.2940 |
| $\sigma_g^2$ | 2.8466 | 2.8466 | 2.8466 |
| $c$ | – | – | – |
| | HetN | | |
| $p_g$ | 0.1841 | 0.2727 | 0.5432 |
| Intercept | 148.7525 | 75.1858 | 155.3267 |
| $\beta_{1g}$ | −2.9327 | −1.9155 | −2.3968 |
| $\beta_{2g}$ | −0.1350 | 0.3031 | −0.4282 |
| $\sigma_g^2$ | 0.4023 | 6.7563 | 1.7041 |
| $c$ | – | – | – |
| | ConC | | |
| $p_g$ | 0.1806 | 0.2635 | 0.5559 |
| Intercept | 118.0823 | 63.4605 | 159.4095 |
| $\beta_{1g}$ | −2.0272 | −1.8391 | −2.5445 |
| $\beta_{2g}$ | −0.1385 | 0.4038 | −0.4119 |
| $\sigma_g^2$ | 2.1395 | 3.7875 | 2.1395 |
| $c$ | 0.3191 | 0.3191 | 0.3191 |

**Table A.13**
*Temperature* data. Best solutions out of 100 random starts, $G = 4$. $K = n/5$, and test set size $= n/10$.

| | HomN | | | |
|---|---|---|---|---|
| $p_g$ | 0.0856 | 0.1881 | 0.2057 | 0.5206 |
| Intercept | 51.8842 | 117.1396 | 83.4083 | 159.6096 |
| $\beta_{1g}$ | −1.1266 | −1.9904 | −2.1729 | −2.5503 |
| $\beta_{2g}$ | 0.2758 | −0.1430 | 0.3186 | −0.4123 |
| $\sigma_g^2$ | 1.8003 | 1.8003 | 1.8003 | 1.8003 |
| $c$ | – | – | – | – |
| | HetN | | | |
| $p_g$ | 0.1459 | 0.2074 | 0.2148 | 0.4319 |
| Intercept | 112.1847 | 62.8423 | 143.6033 | 162.6549 |
| $\beta_{1g}$ | −1.7960 | −1.6773 | −2.8471 | −2.5137 |
| $\beta_{2g}$ | −0.1775 | 0.3498 | −0.1177 | −0.4610 |
| $\sigma_g^2$ | 0.3626 | 4.4269 | 0.3000 | 1.8447 |
| $c$ | – | – | – | – |
| | ConC | | | |
| $p_g$ | 0.1922 | 0.2387 | 0.2562 | 0.3129 |
| Intercept | 65.9919 | 179.3683 | 114.8658 | 140.3715 |
| $\beta_{2g}$ | −1.6827 | −2.7064 | −1.9692 | −2.7506 |
| $\beta_{1g}$ | 0.3263 | −0.5678 | −0.1328 | −0.1229 |
| $\sigma_g^2$ | 3.5915 | 1.1907 | 1.1907 | 1.1907 |
| $c$ | 0.1099 | 0.1099 | 0.1099 | 0.1099 |

**Table A.14**
*Temperature* data. Best solutions out of 100 random starts, $G = 5$. $K = n/5$, and test set size $= n/10$.

| | HomN | | | | |
|---|---|---|---|---|---|
| $p_g$ | 0.0881 | 0.1275 | 0.2117 | 0.2444 | 0.3284 |
| Intercept | 53.2996 | 53.1467 | 179.4572 | 114.5672 | 140.4554 |
| $\beta_{1g}$ | −0.8048 | -1.1204 | −2.7236 | −1.9806 | −2.7523 |
| $\beta_{2g}$ | 0.0347 | 0.2640 | −0.5639 | −0.1257 | −0.1224 |
| $\sigma_g^2$ | 0.9200 | 0.9200 | 0.9200 | 0.9200 | 0.9200 |
| $c$ | – | – | – | – | – |
| | HetN | | | | |
| $p_g$ | 0.0536 | 0.1180 | 0.1515 | 0.2168 | 0.4600 |
| Intercept | 130.2868 | 52.4339 | 109.4192 | 143.6856 | 155.3994 |
| $\beta_{1g}$ | −2.5978 | −1.1145 | −2.4808 | −2.8487 | −2.3851 |
| $\beta_{2g}$ | −0.0491 | 0.2673 | 0.1908 | −0.1179 | −0.4358 |
| $\sigma_g^2$ | $10^{-10}$ | 1.1011 | 3.8704 | 0.2956 | 1.7691 |
| $c$ | – | – | – | – | – |
| | ConC | | | | |
| $p_g$ | 0.0746 | 0.1309 | 0.2403 | 0.2639 | 0.2903 |
| Intercept | 26.6487 | 52.9665 | 174.4789 | 111.1235 | 139.2966 |
| $\beta_{1g}$ | −0.6535 | −1.1149 | −2.7457 | −1.7887 | −2.7735 |
| $\beta_{2g}$ | 0.2028 | 0.2635 | −0.5017 | −0.1729 | −0.1010 |
| $\sigma_g^2$ | 0.8116 | 1.0558 | 1.1956 | 0.8939 | 0.8116 |
| $c$ | 0.4608 | 0.4608 | 0.4608 | 0.4608 | 0.4608 |

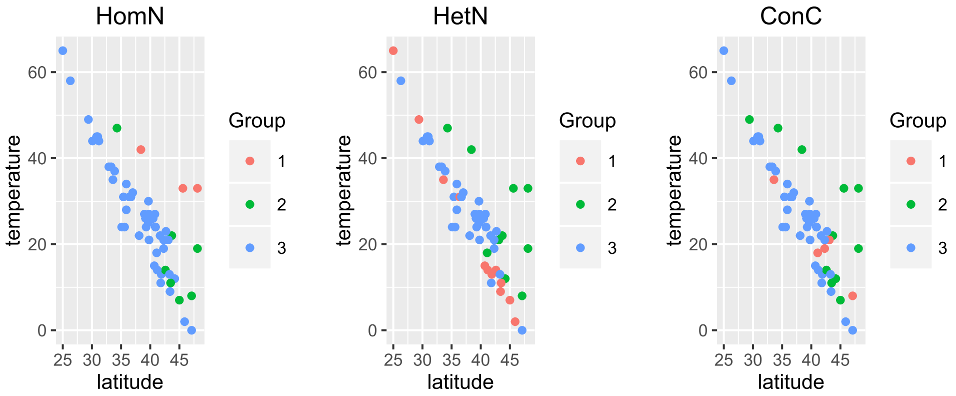

**Fig. A.6.** *Temperature* data. Best solutions out of 100 random starts, $G = 3$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

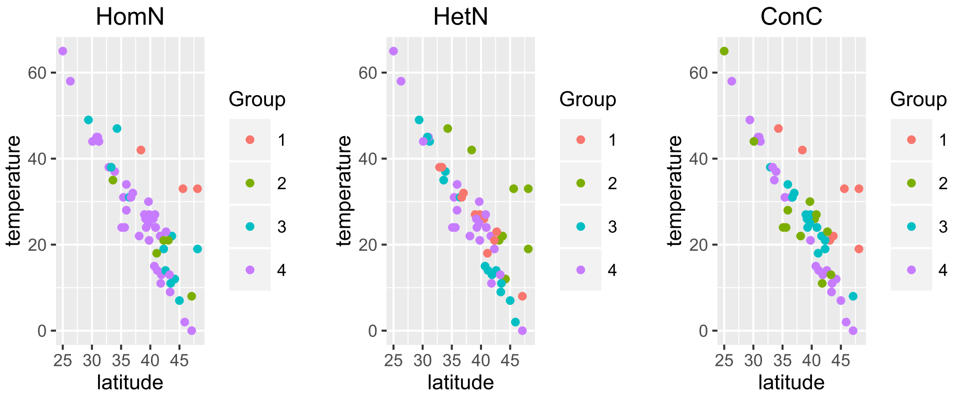

**Fig. A.7.** *Temperature* data. Best solutions out of 100 random starts, $G = 4$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

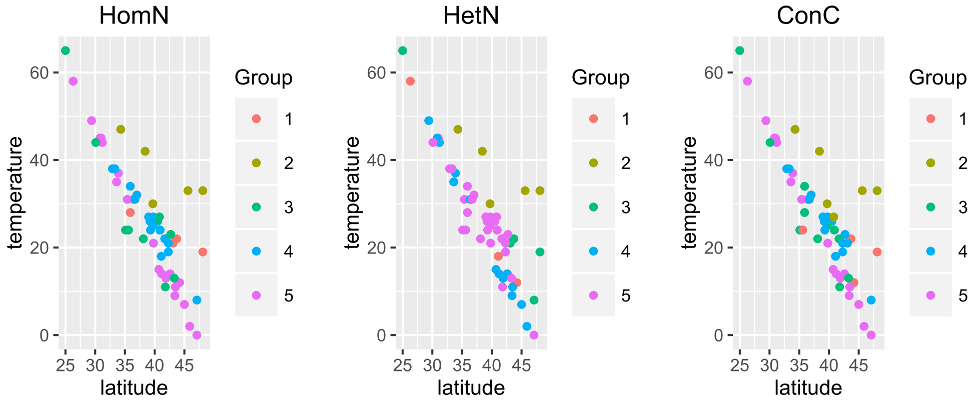

**Fig. A.8.** *Temperature* data. Best solutions out of 100 random starts, $G = 5$. $K = n/5$, and test set size $= n/10$. (For interpretation of the colors in this figure, the reader is referred to the web version of this article.)

### *A.2. Additional tables for* Auto-Mpg *data*

**Table A.15**
*Auto-Mpg* data. Covariates are acceleration ($\mathbf{x}_1$), cylinders ($\mathbf{x}_2$), displacement ($\mathbf{x}_3$), horsepower ($\mathbf{x}_4$), model year ($\mathbf{x}_5$), weight ($\mathbf{x}_6$), and origin ($\mathbf{x}_7$). Best solutions out of 100 random starts, $G = 3$. $K = n/5$, and test set size $= n/10$.

| | HomN | | |
|---|---|---|---|
| $p_g$ | 0.0632 | 0.3222 | 0.6146 |
| Intercept | −53.1186 | −19.8411 | 3.0902 |
| $\beta_{1g}$ | −0.4322 | −0.1873 | −0.3814 |
| $\beta_{2g}$ | 9.6131 | −0.9644 | −0.8957 |
| $\beta_{3g}$ | −0.1643 | 0.0551 | −0.0100 |
| $\beta_{4g}$ | −0.2817 | −0.1284 | −0.0006 |
| $\beta_{5g}$ | 1.4806 | 1.0009 | 0.4821 |
| $\beta_{6g}$ | −0.0094 | −0.0077 | −0.0026 |
| $\beta_{7g}$ | −0.4209 | 0.3868 | 2.5198 |
| $\sigma_g^2$ | 1.7582 | 1.7582 | 1.7582 |
| $c$ | – | – | – |
| | HetN | | |
| $p_g$ | 0.2412 | 0.3718 | 0.3871 |
| Intercept | 14.4577 | −34.1626 | 4.9302 |
| $\beta_{1g}$ | −0.2244 | 0.2365 | −0.2740 |
| $\beta_{2g}$ | −0.8038 | 0.3254 | −0.8088 |
| $\beta_{3g}$ | 0.0103 | 0.0343 | −0.0246 |
| $\beta_{4g}$ | 0.0049 | −0.1122 | 0.0229 |
| $\beta_{5g}$ | 0.4578 | 1.0739 | 0.3715 |
| $\beta_{6g}$ | −0.0067 | −0.0089 | −0.0015 |
| $\beta_{7g}$ | −0.6552 | 0.7969 | 3.5060 |
| $\sigma_g^2$ | 0.8505 | 3.2226 | 1.0832 |
| $c$ | – | – | – |
| | ConC | | |
| $p_g$ | 0.2601 | 0.3695 | 0.3704 |
| Intercept | 13.7311 | −33.2343 | 4.8070 |
| $\beta_{1g}$ | −0.2519 | 0.2324 | −0.2699 |
| $\beta_{2g}$ | −0.8430 | 0.3259 | −0.8162 |
| $\beta_{3g}$ | 0.0093 | 0.0344 | −0.0247 |
| $\beta_{4g}$ | 0.0018 | −0.1122 | 0.0236 |
| $\beta_{5g}$ | 0.4636 | 1.0642 | 0.3696 |
| $\beta_{6g}$ | −0.0062 | −0.0089 | −0.0014 |
| $\beta_{7g}$ | −0.5811 | 0.7944 | 3.5566 |
| $\sigma_g^2$ | 0.9572 | 3.2366 | 1.0659 |
| $c$ | 0.0724 | 0.0724 | 0.0724 |

**Table A.16**
*Auto-Mpg* data. Covariates are acceleration ($\mathbf{x}_1$), cylinders ($\mathbf{x}_2$), displacement ($\mathbf{x}_3$), horsepower ($\mathbf{x}_4$), model year ($\mathbf{x}_5$), weight ($\mathbf{x}_6$), and origin ($\mathbf{x}_7$). Best solutions out of 100 random starts, $G = 4$. $K = n/5$, and test set size $= n/10$.

| | HomN | | | |
|---|---|---|---|---|
| $p_g$ | 0.0889 | 0.2079 | 0.2772 | 0.4261 |
| Intercept | −28.3097 | −9.5904 | −19.4238 | 6.6500 |
| $\beta_{1g}$ | −0.4422 | 0.2457 | −0.4974 | −0.3867 |
| $\beta_{2g}$ | 2.1330 | −0.4710 | 3.9931 | −1.5974 |
| $\beta_{3g}$ | 0.0244 | 0.0373 | −0.0786 | −0.0116 |
| $\beta_{4g}$ | −0.2783 | −0.0260 | −0.0540 | 0.0036 |
| $\beta_{5g}$ | 1.2226 | 0.5874 | 0.8509 | 0.4364 |
| $\beta_{6g}$ | −0.0082 | −0.0065 | −0.0055 | −0.0014 |
| $\beta_{7g}$ | 2.1787 | 0.1497 | 0.4964 | 2.6528 |
| $\sigma_g^2$ | 1.4745 | 1.4745 | 1.4745 | 1.4745 |
| $c$ | – | – | – | – |
| | HetN | | | |
| $p_g$ | 0.0359 | 0.1801 | 0.3457 | 0.4383 |
| Intercept | 15.0766 | 20.1896 | −36.9270 | 5.0291 |
| $\beta_{1g}$ | 0.7837 | −0.1326 | 0.1547 | −0.3891 |
| $\beta_{2g}$ | −0.8292 | −0.8417 | 1.0346 | −0.8139 |
| $\beta_{3g}$ | 0.0983 | 0.0214 | 0.0174 | −0.0172 |
| $\beta_{4g}$ | 0.0389 | 0.0096 | −0.1523 | −0.0015 |
| $\beta_{5g}$ | 0.4991 | 0.4083 | 1.0995 | 0.4125 |
| $\beta_{6g}$ | −0.0205 | −0.0085 | −0.0071 | −0.0015 |
| $\beta_{7g}$ | −0.1739 | −0.7710 | 0.8632 | 3.4170 |
| $\sigma_g^2$ | 0.0060 | 0.5670 | 3.0554 | 1.1632 |
| $c$ | – | – | – | – |
| | | | ConC | |
| $p_g$ | 0.1047 | 0.1464 | 0.3453 | 0.4036 |
| Intercept | −8.0065 | −3.2482 | −21.0692 | 6.9983 |
| $\beta_{1g}$ | 0.6786 | −0.8861 | 0.1542 | −0.4218 |
| $\beta_{2g}$ | −0.8268 | −0.7765 | 0.2859 | −0.8561 |
| $\beta_{3g}$ | 0.0583 | −0.0109 | 0.0321 | −0.0263 |
| $\beta_{4g}$ | −0.0110 | −0.0302 | −0.1392 | 0.0163 |
| $\beta_{5g}$ | 0.4805 | 0.6542 | 0.9416 | 0.4316 |
| $\beta_{6g}$ | −0.0084 | −0.0011 | −0.0082 | −0.0020 |
| $\beta_{7g}$ | 1.6221 | 1.3858 | 0.7191 | 2.7470 |
| $\sigma_g^2$ | 0.2506 | 0.4085 | 3.1986 | 1.3927 |
| $c$ | 0.0008 | 0.0008 | 0.0008 | 0.0008 |

**Table A.17**
*Auto-Mpg* data. Covariates are acceleration ($\mathbf{x}_1$), cylinders ($\mathbf{x}_2$), displacement ($\mathbf{x}_3$), horsepower ($\mathbf{x}_4$), model year ($\mathbf{x}_5$), weight ($\mathbf{x}_6$), and origin ($\mathbf{x}_7$). Best solutions out of 100 random starts, $G = 5$. $K = n/5$, and test set size $= n/10$.

| | HomN | | | | |
|---|---|---|---|---|---|
| $p_g$ | 0.0616 | 0.1243 | 0.2082 | 0.2425 | 0.3634 |
| Intercept | −88.5380 | −6.4162 | −6.6549 | −23.9388 | 3.0745 |
| $\beta_{1g}$ | −0.8368 | −0.2127 | 0.4036 | −0.2718 | −0.3781 |
| $\beta_{2g}$ | 4.3099 | −2.5067 | −0.4849 | 4.2948 | −1.8472 |
| $\beta_{3g}$ | 0.0730 | 0.1544 | 0.0373 | −0.0922 | −0.0083 |
| $\beta_{4g}$ | −0.3422 | −0.0080 | −0.0191 | 0.0182 | 0.0007 |
| $\beta_{5g}$ | 2.0484 | 1.0615 | 0.5099 | 0.8404 | 0.4694 |
| $\beta_{6g}$ | −0.0114 | −0.0222 | −0.0067 | −0.0068 | −0.0009 |
| $\beta_{7g}$ | 2.3141 | 1.6666 | 0.6249 | 0.1703 | 2.8986 |
| $\sigma_g^2$ | 1.1825 | 1.1825 | 1.1825 | 1.1825 | 1.1825 |
| $c$ | – | – | – | – | – |

(*continued on next page*)

**Table A.17** (*continued*)

| | HetN | | | | |
|---|---|---|---|---|---|
| $p_g$ | 0.0480 | 0.0589 | 0.2000 | 0.2416 | 0.4516 |
| Intercept | −9.4586 | −25.5499 | 19.0180 | −42.4670 | 5.0845 |
| $\beta_{1g}$ | −0.7787 | -1.0379 | −0.1519 | 0.4130 | −0.3671 |
| $\beta_{2g}$ | 4.6792 | -4.0897 | −0.7741 | 2.2460 | −0.9636 |
| $\beta_{3g}$ | −0.0755 | 0.1232 | 0.0163 | −0.0114 | −0.0132 |
| $\beta_{4g}$ | −0.0792 | −0.2290 | 0.0086 | −0.1404 | −0.0033 |
| $\beta_{5g}$ | 0.8535 | 1.2628 | 0.4171 | 1.1048 | 0.4094 |
| $\beta_{6g}$ | −0.0082 | −0.0043 | −0.0080 | −0.0076 | −0.0016 |
| $\beta_{7g}$ | 0.4709 | 2.4836 | −0.7536 | 0.7408 | 3.5079 |
| $\sigma_g^2$ | 0.0187 | 0.0540 | 0.5931 | 2.9375 | 1.2773 |
| $c$ | – | – | – | – | – |
| | ConC | | | | |
| $p_g$ | 0.0973 | 0.1158 | 0.1677 | 0.3056 | 0.3135 |
| Intercept | −28.3372 | −7.2900 | 15.7667 | 9.4782 | −33.8232 |
| $\beta_{1g}$ | −0.0430 | −0.4034 | −0.1149 | −0.2713 | 0.3658 |
| $\beta_{2g}$ | −3.5998 | −2.3618 | −0.0925 | −0.1284 | 0.4192 |
| $\beta_{3g}$ | 0.1208 | 0.0105 | 0.0025 | −0.0373 | 0.0278 |
| $\beta_{4g}$ | −0.1160 | −0.0165 | 0.0032 | 0.0393 | −0.1106 |
| $\beta_{5g}$ | 1.1163 | 0.5974 | 0.4200 | 0.3145 | 1.0535 |
| $\beta_{6g}$ | −0.0079 | −0.0005 | −0.0073 | −0.0023 | −0.0088 |
| $\beta_{7g}$ | −0.4597 | 2.9456 | −0.8094 | 3.1784 | 0.6004 |
| $\sigma_g^2$ | 0.2432 | 0.2432 | 0.4962 | 0.8777 | 3.1862 |
| $c$ | 0.0013 | 0.0013 | 0.0013 | 0.0013 | 0.0013 |

## References

[1] S. Arlot, A. Celisse, Cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79.

[2] A.M. Bagirov, J. Ugon, H. Mirzayeva, Nonsmooth nonconvex optimization approach to clusterwise linear regression problems, Eur. J. Oper. Res. 229 (1) (2013) 132–142.

[3] R.A. Carbonneau, G. Caporossi, P. Hansen, Globally optimal clusterwise regression by mixed logical-quadratic programming, Eur. J. Oper. Res. 212 (1) (2011) 213–222.

[4] N.E. Day, Estimating the components of a mixture of two normal distributions, Biometrika 56 (1969) 463–474.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc., Ser. B, Stat. Methodol. 39 (1977) 1–38.

[6] L.A. García-Escudero, A. Gordaliza, A. Mayo-Iscar, R. San Martín, Robust clusterwise linear regression through trimming, Comput. Stat. Data Anal. 54 (12) (2010) 3057–3069.

[7] L.A. García-Escudero, A. Gordaliza, R. San Martín, S. Van Aelst, R. Zamar, Robust linear clustering, J. R. Stat. Soc., Ser. B, Stat. Methodol. 71 (1) (2009) 301–318.

[8] R.J. Hathaway, A constrained formulation of maximum-likelihood estimation for normal mixture distributions, Ann. Stat. 13 (1985) 795–800.

[9] R.J. Hathaway, Another interpretation of the EM algorithm for mixture distributions, Stat. Probab. Lett. 4 (2) (1986) 53–56.

[10] C. Hennig, Identifiablity of models for clusterwise linear regression, J. Classif. 17 (2) (2000) 273–296.

[11] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218.

[12] S. Ingrassia, A likelihood-based constrained algorithm for multivariate normal mixture models, Stat. Methods Appl. 13 (2004) 151–166.

[13] S. Ingrassia, R. Rocci, A constrained monotone EM algorithm for finite mixture of multivariate Gaussians, Comput. Stat. Data Anal. 51 (2007) 5339–5351.

[14] J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann. Math. Stat. 27 (1956) 886–906.

[15] N.M. Kiefer, Discrete parameter variation: efficient estimation of a switching regression model, Econometrica 46 (1978) 427–434.

[16] A.E. Lamont, J.K. Vermunt, L.M. Van Horn, Regression mixture models: does modeling the covariance between independent variables and latent classes improve the results? Multivar. Behav. Res. 51 (1) (2016) 35–52.

[17] L.H. Long (Ed.), The 1972 World Almanac and Book of Facts, Newspaper Enterprise Association, New York, 1972.

[18] G.J. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, New York, 2000.

[19] J.L. Peixoto, A property of well-formulated polynomial regression models, Am. Stat. 44 (1) (1990) 26–30.

[20] R.F. Phillips, A constrained maximum-likelihood approach to estimating switching regressions, J. Econom. 48 (1–2) (1991) 241–262.

[21] R.E. Quandt, A new approach to estimating switching regressions, J. Am. Stat. Assoc. 67 (338) (1972) 306–310.

[22] R.E. Quandt, J.B. Ramsey, Estimating mixtures of normal distributions and switching regressions, J. Am. Stat. Assoc. 73 (364) (1978) 730–738.

[23] R. Rocci, S.A. Gattone, R. Di Mari, A data driven equivariant approach to constrained Gaussian mixture modeling, Adv. Data Anal. Classif. (2017), http://dx.doi.org/10.1007/s11634-016-0279-1.

[24] G. Ritter, Robust Cluster Analysis and Variable Selection, Monographs on Statistics and Applied Probability, vol. 137, CRC Press, 2014.

[25] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.

[26] P. Smyth, Clustering using Monte–Carlo cross validation, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 126–133.

[27] P. Smyth, Model selection for probabilistic clustering using cross-validated likelihood, Stat. Comput. 10 (1) (2000) 63–72.

[28] M.J. van der Laan, S. Dudoit, S. Keles, Asymptotic optimality of likelihood-based cross-validation, Stat. Appl. Genet. Mol. Biol. 3 (1) (2004) 1–23.

[29] J.H. Won, J. Lim, S.J. Kim, B. Rajaratnam, Condition-number-regularized covariance estimation, J. R. Stat. Soc., Ser. B, Stat. Methodol. 75 (3) (2013) 427–450.

[30] J. Xu, X. Tan, R. Zhang, A note on Phillips (1991): "A constrained maximum likelihood approach to estimating switching regressions", J. Econom. 154 (1) (2010) 35–41.

[31] P.W. Zehna, Invariance of maximum likelihood estimators, Ann. Math. Stat. 37 (3) (1966) 744.