




# Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition

Salvatore Ingrassia<sup>1</sup> · Antonio Punzo<sup>1</sup> 

Published online: 16 July 2019  
© The Classification Society 2019

## Abstract

One of the challenges in cluster analysis is the evaluation of the obtained clustering results without using auxiliary information. To this end, a common approach is to use internal validity criteria. For mixtures of linear regressions whose parameters are estimated by maximum likelihood, we propose a three-term decomposition of the total sum of squares as a starting point to define some internal validity criteria. In particular, three types of mixtures of regressions are considered: with fixed covariates, with concomitant variables, and with random covariates. A ternary diagram is also suggested for easier joint interpretation of the three terms of the proposed decomposition. Furthermore, local and overall coefficients of determination are respectively defined to judge how well the model fits the data group-by-group but also taken as a whole. Artificial data are considered to find out more about the proposed decomposition, including violations of the model assumptions. Finally, an application to real data illustrates the use and the usefulness of these proposals.

**Keywords** Cluster validation · EM algorithm · Maximum likelihood · Mixtures of regressions · Model-based clustering · Ternary diagram

## 1 Introduction

The decomposition of the total sum of squares (total variation) in the explained sum of squares (explained variation) plus the sum of squared residuals (unexplained variation) is a peculiarity of the linear regression model whose parameters are estimated by least squares (see, e.g., Davidson and MacKinnon 2004, pp. 117–118). The famous coefficient of determination, universally referred to as  $R^2$ , which is defined as a measure of explained variation, arises from this decomposition and is used to evaluate the goodness-of-fit for the linear regression model. Surprisingly, as also emphasized by Cameron and Windmeijer (1996), the

---

✉ Antonio Punzo  
antonio.punzo@unict.it

Salvatore Ingrassia  
s.ingrassia@unict.it

<sup>1</sup> Department of Economics and Business, University of Catania, Corso Italia 55, 95129, Catania, Italy

extension to other models is rare, with the notable exceptions of models with heteroscedastic errors with known variance (Buse 1973), logit and probit models (see Maddala 1986, Windmeijer 1995, and the references therein), tobit models (surveyed by Veall and Zimmermann 1996), regression models for count data (Cameron and Windmeijer 1996), and some common nonlinear regression models (Cameron and Windmeijer 1997).

We focus on mixtures of (linear) regressions, also known in literature as switching regression or clusterwise regression models (see, e.g., Wedel 1990; Wedel and Kamakura 2000, Chapter 7; Frühwirth-Schnatter 2006, Chapter 8). These models represent a classical alternative/generalization of a single (linear) regression to be used when there is some latent or unobserved feature splitting the data into groups (or clusters) having a regression relationship.

Three eminent members of the class of mixtures of regressions—whose peculiarities and differences are detailed in Ingrassia et al. (2012) and Ingrassia and Punzo (2016)—are mixtures of regressions with fixed covariates (De Sarbo et al. 1988; see also Quandt 1972, Hosmer 1974, and Quandt and Ramsey 1978, for the special case of two mixture components), mixtures of regressions with concomitant variables (Dayton and Macready 1988), and mixtures of regressions with random covariates (Hennig 2000). For these three classes of mixtures of regressions, we propose a finer three-term decomposition of the total sum of squares when the parameters are estimated with the expectation-maximization (EM) algorithm (Dempster et al. 1977), within a maximum likelihood framework, under normally distributed errors in each mixture component. The terms of this decomposition allow the user to investigate the main aspects of the fitted model via normalized measures. These aspects are the association between the response variable and the latent groups, the goodness-of-fit of the model, and the proportion of the total variation in the dependent variable which remains unexplained by the fitted model. Furthermore, local and overall coefficients of determination are respectively introduced to evaluate how well the model fits the data group-by-group but also taken as a whole.

The proposed decomposition and measures can be seen also as cluster validity methods (see, e.g., Halkidi et al. 2001; Theodoridis and Koutroumbas 2008, Chapter 16) for mixtures of regressions, i.e., as methods aiming at the quantitative evaluation of the clusters from the fitted models; this is a step of fundamental importance in most applications (Rezaee et al. 1998; Steinley et al. 2015). According to the usual classification of cluster validity criteria as internal, external, and relative (Arbelaitz et al. 2013), our measures can be categorized as internal (Milligan and Cheng 1996), i.e., as criteria which measure the goodness of the estimated clusters without reference to external information.

The paper is organized as follows. Section 2 summarizes basic concepts about the mixtures of linear regressions we consider. Section 3 details the part of the EM algorithm devoted to the update of the local regression coefficients. The proposed three-term decomposition is presented in Section 4. The other proposals are presented in Section 5: normalized measures based on the proposed decomposition are given in Section 5.1, the use of the ternary diagram to display the normalized terms of the decomposition is suggested in Section 5.2, a normalized measure of explained response variation is defined in Section 5.3, and local and overall coefficients of determination are introduced in Section 5.4. Sections 6 and 7 illustrate applications to artificial and real data, respectively. Section 8 summarizes and concludes.

## 2 Mixtures of Regressions

Let  $X$  be a vector of covariates with values in  $\mathbb{R}^d$ , and let  $Y$  be a dependent (or response) variable taking values in  $\mathbb{R}$ . Suppose that the regression of  $Y$  on  $X$  varies across the  $k$  levels (groups or clusters) of a categorical latent variable  $G$ .

Mixtures of regressions with fixed covariates (MRFC; DeSarbo and Cron 1988) are characterized by the following conditional density function:

$$p(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^k \pi_g f(y|\mathbf{x}; \boldsymbol{\theta}_g), \tag{1}$$

where  $\pi_g = P(G = g)$  is the mixing weight, with  $\pi_g > 0$  and  $\sum_{g=1}^k \pi_g = 1$ , while  $f(y|\mathbf{x}; \boldsymbol{\theta}_g)$  is the conditional density of  $Y|X = \mathbf{x}, G = g$  depending on the parameter vector  $\boldsymbol{\theta}_g$ . In (1),  $\boldsymbol{\psi} = (\pi_1, \dots, \pi_{k-1}, \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_k)'$  denotes the set of all parameters of the model (see also Mazza and Punzo 2018).

Mixtures of regressions with concomitant variables (MRCV; Dayton and Macready 1988), when covariates and concomitant variables coincide, are characterized by the following density:

$$p(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^k p(G = g|\mathbf{x}; \boldsymbol{\alpha}) f(y|\mathbf{x}; \boldsymbol{\theta}_g), \tag{2}$$

where the mixing weight  $p(G = g|\mathbf{x}; \boldsymbol{\alpha})$  is now a function depending on  $\mathbf{x}$  through a parameter vector  $\boldsymbol{\alpha}$ , and  $\boldsymbol{\psi} = (\boldsymbol{\alpha}', \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_k)'$  denotes the set of all parameters of the model. The probability  $p(G = g|\mathbf{x}; \boldsymbol{\alpha})$  is usually modeled by the multinomial logistic model:

$$p(G = g|\mathbf{x}; \boldsymbol{\alpha}) = \frac{\exp(\alpha_{g0} + \boldsymbol{\alpha}'_{g1}\mathbf{x})}{\sum_{j=1}^k \exp(\alpha_{j0} + \boldsymbol{\alpha}'_{j1}\mathbf{x})},$$

where  $\boldsymbol{\alpha}_g = (\alpha_{g0}, \boldsymbol{\alpha}'_{g1})' \in \mathbb{R}^{d+1}$  and  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \dots, \boldsymbol{\alpha}'_k)'$ , with  $\boldsymbol{\alpha}_1 \equiv \mathbf{0}$  (see, e.g., Grün and Leisch 2008 and Mazza et al. 2019).

Mixtures of regressions with random covariates (MRRC) have been first discussed in Gershensfeld (1997), Hennig (2000), and Wedel (2002), and have been referred to as cluster-weighted models (CWM), clusterwise regression with random covariates, and saturated mixture regression models, respectively. Recent work on MRRC can be found in Ingrassia et al. (2012), Ingrassia et al. (2014), Ingrassia et al. (2015); Punzo (2014); Punzo and Ingrassia (2015); Subedi et al. (2013), Subedi et al. (2015); Berta et al. (2016); McNicholas (2016); Punzo and McNicholas (2017); Punzo et al. (2018); Zarei et al. (2018). Differently from MRFC and MRCV, which model the conditional density of  $Y|X = \mathbf{x}$ , MRRC models the joint distribution of  $(X', Y)'$  as:

$$p(\mathbf{x}, y; \boldsymbol{\psi}) = \sum_{g=1}^k \pi_g f(y|\mathbf{x}; \boldsymbol{\theta}_g) p(\mathbf{x}; \boldsymbol{\xi}_g), \tag{3}$$

where  $p(\mathbf{x}; \boldsymbol{\xi}_g)$  is the density of  $X|G = g$ , depending on the parameter vector  $\boldsymbol{\xi}_g$ , and  $\boldsymbol{\psi} = (\pi_1, \dots, \pi_{k-1}, \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_k, \boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_k)'$ .

### 3 Maximum Likelihood Estimation: the EM Algorithm

In models (1)–(3), assume a normal distribution for  $Y|X = \mathbf{x}, G = g$ . Denoting with  $\phi(y; \mu, \sigma^2)$ , the univariate normal density with mean  $\mu$ , and variance  $\sigma^2$ , these models specialize respectively as:

$$\text{MRFC} : p(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^k \pi_g \phi \left[ y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_g^2 \right], \tag{4}$$

$$\text{MRCV} : p(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^k p(G = g|\mathbf{x}; \boldsymbol{\alpha}) \phi \left[ y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_g^2 \right], \tag{5}$$

$$\text{MRRC} : p(\mathbf{x}, y; \boldsymbol{\psi}) = \sum_{g=1}^k \pi_g p(\mathbf{x}; \boldsymbol{\xi}_g) \phi \left[ y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_g^2 \right], \tag{6}$$

where the local conditional densities are based on the linear function  $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = \beta_{g0} + \boldsymbol{\beta}'_{g1}\mathbf{x}$ , with  $\boldsymbol{\beta}_g = (\beta_{g0}, \boldsymbol{\beta}'_{g1})'$ ,  $\beta_{g0} \in \mathbb{R}$ , and  $\boldsymbol{\beta}_{g1} \in \mathbb{R}^d$ .

Maximum likelihood (ML) parameter estimates for models (4)–(6) are usually obtained via the expectation-maximization (EM) algorithm (Dempster et al. 1977). Given a random sample  $(\mathbf{x}'_1, y_1)', \dots, (\mathbf{x}'_n, y_n)'$  of  $(\mathbf{X}', Y)'$ , for a fixed number  $k$  of groups, for models (4)–(6) the algorithm basically takes into account the complete-data log-likelihood:

$$\text{MRFC} : l_c(\boldsymbol{\psi}) = \sum_{g=1}^k \sum_{i=1}^n z_{ig} \ln(\pi_g) + l_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2), \tag{7}$$

$$\text{MRCV} : l_c(\boldsymbol{\psi}) = \sum_{g=1}^k \sum_{i=1}^n z_{ig} \ln[p(G = g|\mathbf{x}_i; \boldsymbol{\alpha})] + l_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2), \tag{8}$$

$$\text{MRRC} : l_c(\boldsymbol{\psi}) = \sum_{g=1}^k \sum_{i=1}^n z_{ig} \ln(\pi_g) + \sum_{g=1}^k \sum_{i=1}^n z_{ig} \ln[p(\mathbf{x}_i; \boldsymbol{\xi}_g)] + l_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2), \tag{9}$$

respectively, where  $z_{ig} = 1$  if  $(\mathbf{x}'_i, y_i)'$  comes from component  $g$  and  $z_{ig} = 0$  otherwise, and

$$l_{\text{reg}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \sum_{g=1}^k \sum_{i=1}^n z_{ig} \ln \left\{ \phi \left[ y_i; \mu(\mathbf{x}_i; \boldsymbol{\beta}_g), \sigma_g^2 \right] \right\}, \tag{10}$$

where  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_k)'$  and  $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_k^2)'$ . It is well known that the EM algorithm iterates between two steps, the E-step and the M-step, until convergence; their schematization, only with respect to the estimation of  $\boldsymbol{\beta}$  and  $\boldsymbol{\sigma}^2_g$  from  $l_{\text{reg}}$ , is given below.

E-step: Given the current parameter estimates  $\psi^{(r)}$  on the  $r$ th iteration, simply replace each  $z_{ig}$  by the estimated posterior probability:

$$\text{MRFC : } z_{ig}^{(r)} = \frac{\pi_g^{(r)} \phi(y_i | \mathbf{x}_i; \boldsymbol{\beta}_g^{(r)}, \sigma_g^{2,(r)})}{p(y_i | \mathbf{x}_i; \boldsymbol{\psi}^{(r)})}, \tag{11}$$

$$\text{MRCV : } z_{ig}^{(r)} = \frac{p(G = g | \mathbf{x}_i; \boldsymbol{\alpha}^{(r)}) \phi(y_i | \mathbf{x}_i; \boldsymbol{\beta}_g^{(r)}, \sigma_g^{2,(r)})}{p(y_i | \mathbf{x}_i; \boldsymbol{\psi}^{(r)})}, \tag{12}$$

$$\text{MRRC : } z_{ig}^{(r)} = \frac{\pi_g^{(r)} p(\mathbf{x}_i; \boldsymbol{\xi}_g^{(r)}) \phi(y_i | \mathbf{x}_i; \boldsymbol{\beta}_g^{(r)}, \sigma_g^{2,(r)})}{p(\mathbf{x}_i, y_i; \boldsymbol{\psi}^{(r)})}. \tag{13}$$

M-step (regression parameters only): the values  $z_{ig}^{(r)}$  are substituted to  $z_{ig}$  in (7)–(9) yielding the expected complete-data log-likelihood whose terms can be maximized separately. In particular, the expectation of  $l_{\text{reg}}$  in (10) yields:

$$Q_{\text{reg}}(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^n \sum_{g=1}^k z_{ig}^{(r)} \ln [\phi(y_i | \mathbf{x}_i; \boldsymbol{\beta}_g, \sigma_g^2)].$$

The maximization of  $Q_{\text{reg}}$  with respect to  $\boldsymbol{\beta}$  and  $\sigma^2$  is equivalent to independently maximize each of the  $k$  expressions:

$$Q_{\text{reg},g}(\boldsymbol{\beta}_g, \sigma_g^2) = \frac{1}{2} \sum_{i=1}^n z_{ig}^{(r)} \left[ -\ln(2\pi) - \ln(\sigma_g^2) - \frac{(y_i - \boldsymbol{\beta}'_g \mathbf{x}_i)^2}{\sigma_g^2} \right] \tag{14}$$

with respect to  $\boldsymbol{\beta}_g$  and  $\sigma_g^2$ ,  $g = 1, \dots, k$ . The maximization of (14) is equivalent to the maximization problem of the linear regression model (for the complete data), except that each observation  $(\mathbf{x}'_i, y_i)'$  contributes to the log-likelihood with a known weight  $z_{ig}^{(r)}$ .

**Update for  $\boldsymbol{\beta}_{g1}$**  Equating to zero the differentiation of (14) with respect to  $\boldsymbol{\beta}_{g1}$  yields

$$\sum_{i=1}^n z_{ig}^{(r)} (y_i - \beta_{g0} - \boldsymbol{\beta}'_{g1} \mathbf{x}_i) \mathbf{x}_i = \mathbf{0} \tag{15}$$

$$\sum_{i=1}^n z_{ig}^{(r)} (y_i - \beta_{g0} - \boldsymbol{\beta}'_{g1} \mathbf{x}_i) \mathbf{x}_i - \sum_{i=1}^n z_{ig}^{(r)} (y_i - \beta_{g0} - \boldsymbol{\beta}'_{g1} \mathbf{x}_i) \bar{\mathbf{x}}_g = \mathbf{0}$$

$$\sum_{i=1}^n z_{ig}^{(r)} (y_i - \beta_{g0} - \boldsymbol{\beta}'_{g1} \mathbf{x}_i) (\mathbf{x}_i - \bar{\mathbf{x}}_g) = \mathbf{0} \tag{16}$$

$$\sum_{i=1}^n z_{ig}^{(r)} [(y_i - \bar{y}_g) - \boldsymbol{\beta}'_{g1} (\mathbf{x}_i - \bar{\mathbf{x}}_g)] (\mathbf{x}_i - \bar{\mathbf{x}}_g) = \mathbf{0}$$

$$\sum_{i=1}^n z_{ig}^{(r)} (y_i - \bar{y}_g) (\mathbf{x}_i - \bar{\mathbf{x}}_g) - \left[ \sum_{i=1}^n z_{ig}^{(r)} (\mathbf{x}_i - \bar{\mathbf{x}}_g) (\mathbf{x}_i - \bar{\mathbf{x}}_g)' \right] \boldsymbol{\beta}_{g1} = \mathbf{0}, \tag{17}$$

where

$$\bar{y}_g = \frac{1}{n_g^{(r)}} \sum_{i=1}^n z_{ig}^{(r)} y_i \quad \text{and} \quad \bar{\mathbf{x}}_g = \frac{1}{n_g^{(r)}} \sum_{i=1}^n z_{ig}^{(r)} \mathbf{x}_i, \tag{18}$$

with  $n_g^{(r)} = \sum_{i=1}^n z_{ig}^{(r)}$  being the expected *a posteriori* size of the  $g$ th group. Solving (17) with respect to  $\beta_{g1}^{(r)}$  yields:

$$\beta_{g1}^{(r+1)} = \left[ \sum_{i=1}^n z_{ig}^{(r)} (\mathbf{x}_i - \bar{\mathbf{x}}_g) (\mathbf{x}_i - \bar{\mathbf{x}}_g)' \right]^{-1} \sum_{i=1}^n z_{ig}^{(r)} (y_i - \bar{y}_g) (\mathbf{x}_i - \bar{\mathbf{x}}_g). \tag{19}$$

**Update for  $\beta_{g0}$**  Equating to zero the differentiation of (14) with respect to  $\beta_{g0}$ , with  $\beta_{g1}$  substituted by  $\beta_{g1}^{(r+1)}$  in (19), yields:

$$\sum_{i=1}^n z_{ig}^{(r)} (y_i - \beta_{g0} - \beta_{g1}^{(r+1)' } \mathbf{x}_i) = 0$$

$$n_g^{(r)} \beta_{g0} = \sum_{i=1}^n z_{ig}^{(r)} y_i - \sum_{i=1}^n z_{ig}^{(r)} \beta_{g1}^{(r+1)' } \mathbf{x}_i. \tag{20}$$

Solving (20) with respect to  $\beta_{g0}$  yields:

$$\beta_{g0}^{(r+1)} = \bar{y}_g - \beta_{g1}^{(r+1)' } \bar{\mathbf{x}}_g. \tag{21}$$

Note that the local regression coefficients  $\beta_{g0}^{(r+1)}$  in (21) and  $\beta_{g1}^{(r+1)}$  in (19) are weighted least squares estimates of  $\beta_{g0}$  and  $\beta_{g1}$  (see, e.g., Chat et al. 2006, Chapter 7).

**Update for  $\sigma_g^2$**  The maximization of (14) with respect to  $\sigma_g^2$ , with  $\beta_{g0}$  substituted with  $\beta_{g0}^{(r+1)}$  and  $\beta_{g1}$  with  $\beta_{g1}^{(r+1)}$ , yields:

$$\sigma_g^{2,(r+1)} = \frac{1}{n_g^{(r)}} \sum_{i=1}^n z_{ig}^{(r)} \left( y_i - \beta_{g0}^{(r+1)} - \beta_{g1}^{(r+1)' } \mathbf{x}_i \right)^2.$$

A complete description of the M-step can be found in Wedel and De Sarbo (1995) and Wedel and Kamakura (2000), pp. 120–124, for the MRFC, in Leisch (2004) and Grün and Leisch (2008) for the MRCV, and in Mazza et al. (2018) for the MRRC.

Once the model is fitted, each observation is classified into one of the  $k$  categories of the latent variable  $G$  according to the maximum *a posteriori* probability (MAP) estimate:  $\text{MAP}(\hat{z}_{ig}) = 1$  if  $\max_h \{\hat{z}_{ih}\}$  occurs in cluster  $g$ , and 0 otherwise, where  $\hat{z}_{ig}$  denotes the value of  $z_{ig}$  at convergence of the EM algorithm.

### 4 Three-Term Decomposition of the Total Sum of Squares

The total sum of squares (total variability) on  $Y$ , i.e.:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2,$$

can be written, because  $\sum_{g=1}^k \hat{z}_{ig} = 1, i = 1, \dots, n$ , as:

$$\begin{aligned} \text{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \sum_{g=1}^k \hat{z}_{ig} = \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y})^2 = \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g + \bar{y}_g - \bar{y})^2 \\ &= \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g)^2 + \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (\bar{y}_g - \bar{y})^2 + 2 \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g) (\bar{y}_g - \bar{y}) \\ &= \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g)^2 + \sum_{g=1}^k \hat{n}_g (\bar{y}_g - \bar{y})^2 + 2 \sum_{g=1}^k (\bar{y}_g - \bar{y}) \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g), \end{aligned} \tag{22}$$

where  $\hat{n}_g = \sum_{i=1}^n \hat{z}_{ig}$  denotes the expected (soft) size of the  $g$ th group according to the fitted model. Based on (18),  $\sum_{i=1}^n \hat{z}_{ig} y_i = \bar{y}_g \sum_{i=1}^n \hat{z}_{ig}$  and then the last term on the right-hand side of (22) is null. Thus,

$$\text{TSS} = \sum_{g=1}^k \text{SS}_g + \sum_{g=1}^k \hat{n}_g (\bar{y}_g - \bar{y})^2 = \text{WSS} + \text{BSS}, \tag{23}$$

where

$$\text{SS}_g = \sum_{i=1}^n \hat{z}_{ig} (y_i - \bar{y}_g)^2$$

is the (soft) sum of squares in the  $g$ th group,

$$\text{WSS} = \sum_{g=1}^k \text{SS}_g \tag{24}$$

is the (soft) within-group sum of squares, and

$$\text{BSS} = \sum_{g=1}^k \hat{n}_g (\bar{y}_g - \bar{y})^2 \tag{25}$$

is the (soft) between-group sum of squares. The wording ‘‘soft’’ is used because the group memberships  $\hat{z}_{ig}, i = 1, \dots, n$  and  $g = 1, \dots, k$ , are a posteriori probabilities and not ‘‘hard’’ 0/1 values. Denoting with  $\hat{\beta}_g = (\hat{\beta}_{g0}, \hat{\beta}'_{g1})'$  the ML estimate of  $\beta_g = (\beta_{g0}, \beta'_{g1})'$  at convergence of the EM algorithm, the WSS term can be further decomposed as:

$$\begin{aligned} \text{WSS} &= \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} \left[ y_i - \mu(\mathbf{x}_i; \hat{\beta}_g) + \mu(\mathbf{x}_i; \hat{\beta}_g) - \bar{y}_g \right]^2 \\ &= \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} \left[ y_i - \mu(\mathbf{x}_i; \hat{\beta}_g) \right]^2 + \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} \left[ \mu(\mathbf{x}_i; \hat{\beta}_g) - \bar{y}_g \right]^2 \\ &\quad + 2 \sum_{g=1}^k \sum_{i=1}^n \hat{z}_{ig} \left[ y_i - \mu(\mathbf{x}_i; \hat{\beta}_g) \right] \left[ \mu(\mathbf{x}_i; \hat{\beta}_g) - \bar{y}_g \right]. \end{aligned} \tag{26}$$

The use of (16) and (21) in (26) yields:

$$\begin{aligned} & \sum_{i=1}^n \hat{z}_{ig} \left[ y_i - \mu \left( \mathbf{x}_i; \hat{\boldsymbol{\beta}}_g \right) \right] \left[ \mu \left( \mathbf{x}_i; \hat{\boldsymbol{\beta}}_g \right) - \bar{y}_g \right] \\ &= \sum_{i=1}^n \hat{z}_{ig} \left( y_i - \hat{\beta}_{g0} - \hat{\boldsymbol{\beta}}'_{g1} \mathbf{x}_i \right) \left( \hat{\beta}_{g0} + \hat{\boldsymbol{\beta}}'_{g1} \mathbf{x}_i - \hat{\beta}_{g0} - \hat{\boldsymbol{\beta}}'_{g1} \bar{\mathbf{x}}_g \right) \\ &= \hat{\boldsymbol{\beta}}'_{g1} \sum_{i=1}^n \hat{z}_{ig} \left( y_i - \hat{\beta}_{g0} - \hat{\boldsymbol{\beta}}'_{g1} \mathbf{x}_i \right) \left( \mathbf{x}_i - \bar{\mathbf{x}}_g \right) = 0, \end{aligned}$$

so that the third term on the right-hand side of (26) vanishes. Thus, the WSS term in (26) simplifies as:

$$\text{WSS} = \text{EWSS} + \text{RWSS}, \tag{27}$$

with

$$\text{EWSS} = \sum_{g=1}^k \text{ESS}_g, \tag{28}$$

$$\text{RWSS} = \sum_{g=1}^k \text{RSS}_g, \tag{29}$$

where, for each group  $g$ ,

$$\text{ESS}_g = \sum_{i=1}^n \hat{z}_{ig} \left[ \mu \left( \mathbf{x}_i; \hat{\boldsymbol{\beta}}_g \right) - \bar{y}_g \right]^2 \tag{30}$$

is the (soft) explained sum of squares and

$$\text{RSS}_g = \sum_{i=1}^n \hat{z}_{ig} \left[ y_i - \mu \left( \mathbf{x}_i; \hat{\boldsymbol{\beta}}_g \right) \right]^2 \tag{31}$$

is the (soft) residual sum of squares. Finally, substituting (27) in (23) yields

$$\text{TSS} = \text{BSS} + \text{RWSS} + \text{EWSS}. \tag{32}$$

Thus, by considering the classical nomenclature from the (one-factor) analysis of covariance (ANCOVA; see, e.g., Huitema 2011, Chapter 6), the total sum of squares TSS can be broken into three parts: the (soft) between-group sum of squares (i.e., the variability of  $Y$  explained by the latent group variable  $G$ ), or BSS, the (soft) within-group sum of squares explained by the model (thanks to the covariates), or EWSS, and the (soft) residual within-group sum of squares, or RWSS. This means that the (soft) within-group sum of squares (WSS) is decomposed in the WSS predictable from the covariates  $\mathbf{X}$  via the chosen model (EWSS) and the WSS not predictable from  $\mathbf{X}$  via the chosen model (RWSS). Finally note that, when  $k = 1$ , the BSS term in (32) vanishes and  $\text{TSS} = \text{EWSS} + \text{RWSS}$ , which is the classical decomposition of the total sum of squares for the standard linear regression model whose parameters are estimated by least squares.

In terms of clustering validation, BSS can be seen as a separation measure (see, e.g., Cerdeira et al. 2012), i.e., as a measure of how well separated clusters are along the  $y$ -axis (the greater the value of BSS, the more “separated” the clusters are on  $Y$ ), while WSS can be seen as a compactness measure (see, e.g., Panagiotakis 2015), i.e., as a measure of how close observations in a cluster are with respect to the regression line of that cluster (the smaller the value of WSS, the more “compact” the clusters are around their regression line).



## 5 Evaluating the Main Aspects of the Fitted Model

### 5.1 Normalized Three-Term Decomposition

Starting from the three-term decomposition given in (32), it is possible to define normalized summary measures aiming to evaluate the main aspects of the fitted model. In particular, dividing both sides of (32) by TSS yields:

$$\begin{aligned} \frac{\text{BSS}}{\text{TSS}} + \frac{\text{EWSS}}{\text{TSS}} + \frac{\text{RWSS}}{\text{TSS}} &= 1 \\ \text{NBSS} + \text{NEWSS} + \text{NRWSS} &= 1, \end{aligned} \quad (33)$$

where NBSS, NEWSS, and NRWSS are the normalized versions, with respect to TSS, of BSS, EWSS, and RWSS, respectively. In terms of interpretation, NBSS is the proportion of the total variability of  $Y$  in the sample explained by the weighted differences between the weighted group means  $\bar{y}_g$  and the overall mean  $\bar{y}$ ; hence, NBSS can be meant as a sort of correlation ratio, being a measure of association between the dependent variable  $Y$  and the latent group variable  $G$ . NEWSS is the proportion of the total variability of  $Y$  explained by the inclusion of the covariates  $X$  via the slope(s) of the local regressions. On the contrary, NRWSS represents the proportion of the total variability of  $Y$  in the sample which remains unexplained by the fitted model.

### 5.2 Graphical Representation of the Three-Term Decomposition

With reference to a fitted model, the triplet (NBSS, NEWSS, NRWSS) can be seen as a point  $p$  in the probability simplex  $\mathbb{S}^3$ , defined as the 2-dimensional subset of the 3-dimensional space containing vectors with non-negative coordinates summing to one. As illustrated in Aitchison (2003), Chapter 1.4, a convenient way of displaying points in  $\mathbb{S}^3$  is represented by the *ternary diagram* in Fig. 1, an equilateral triangle having unit altitude.

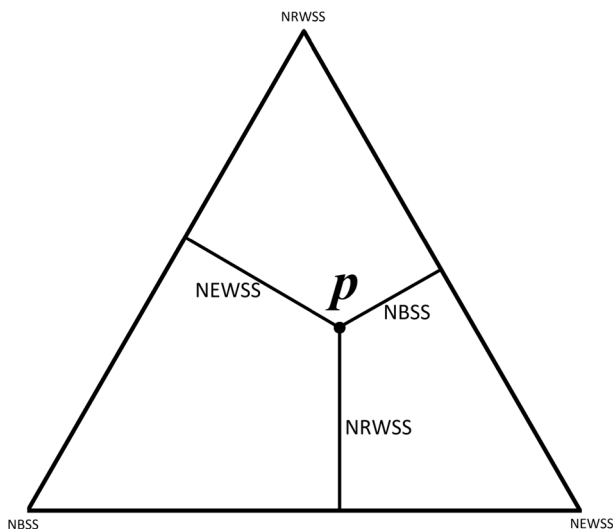


Fig. 1 A point  $p = (\text{NBSS}, \text{NEWSS}, \text{NRWSS})$  in the ternary diagram

Here, for any point  $p$ , the lengths of the perpendiculars NBSS, NEWSS, and NRWSS from  $p$  to the sides opposite to the vertices NBSS, NEWSS, and NRWSS are all greater than, or equal to, 0 and have a unitary sum. Since there is a unique point with these perpendicular values, there is a one-to-one correspondence between  $S^3$  and points in the triangle. In such a representation, the larger a component, say NBSS, the further the point  $p$  is away from the side opposite the vertex NBSS or, in other words, the nearer the point is to the vertex NBSS. Moreover, points with two components, say NBSS and NEWSS, in constant ratio are represented by points on a straight line through the complementary vertex NRWSS. Finally, points with one component, say NRWSS, in constant value are represented by points on a straight line which is parallel to the side opposite to the vertex NRWSS.

### 5.3 Normalized Explained Sum of Squares

According to (33), it is natural to introduce the quantity:

$$NESS = NBSS + NEWSS = 1 - NRWSS \tag{34}$$

representing the proportion of the total variability of  $Y$  explained by the fitted model. NESS desirably assumes values in the interval  $[0, 1]$ : large values of NESS, hence small values of NRWSS, indicate a mixture of regressions that “fits”, or comes closer to, the observed data.

Provided that  $TSS > 0$ , the limit cases  $NESS = 0$  and  $NESS = 1$  are respectively obtained when  $NBSS = NEWSS = 0$  and  $NRWSS = 0$ . Cases where each of the three terms NBSS, NEWSS, and NRWSS is null are analyzed below.

- NBSS = 0 when BSS = 0, i.e., when  $\bar{y}_1 = \dots = \bar{y}_k = \bar{y}$ , regardless of the group sizes  $\hat{n}_1, \dots, \hat{n}_k$ ; refer to (25).
- NEWSS = 0 when EWSS = 0, i.e., when  $\hat{\beta}_{11} = \dots = \hat{\beta}_{k1} = \mathbf{0}$  so that  $\hat{\beta}_{g0} = \bar{y}_g$ ,  $g = 1, \dots, k$ , regardless of the values of  $\hat{z}_{ig}$ ; refer to (28) and (30).
- NRWSS = 0 when RWSS = 0. A sufficient condition for the latter equality to be true, regardless of the values of  $\hat{z}_{ig}$ , is represented by  $k$  overlapped component regression lines (i.e.,  $\hat{\beta}_{10} = \dots = \hat{\beta}_{k0} = \hat{\beta}_0$  and  $\hat{\beta}_{11} = \dots = \hat{\beta}_{k1} = \hat{\beta}_1$ ) with all the  $n$  data points lying on the resulting common regression line (i.e.,  $y_i = \mu(\mathbf{x}_i; \hat{\beta})$ ,  $i = 1, \dots, n$  and  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ ); refer to (29) and (31).

### 5.4 Local and Overall Coefficients of Determination

Since  $\hat{\beta}_g$  is a WLS estimate of  $\beta_g$ , it is natural to define the local coefficient of determination for the  $g$ th group as:

$$R_g^2 = \frac{ESS_g}{ESS_g + RSS_g} = \frac{ESS_g}{SS_g} \tag{35}$$

(see, e.g., Will et al. 1988).  $R_g^2$  can be interpreted as the proportion of response variation in the  $g$ th group that cannot be explained in the model with the only intercept  $\hat{\beta}_{g0}$ , i.e., by  $\mu(\mathbf{x}; \hat{\beta}_{g0}) = \hat{\beta}_{g0}$ , but can be explained by the covariates  $X$  included into the linear model  $\mu(\mathbf{x}; \hat{\beta}_g) = \hat{\beta}_{g0} + \hat{\beta}'_{g1}\mathbf{x}$ . In general, the higher the  $R_g^2$ , the better the  $g$ th linear model fits the data in the  $g$ th group under the idea that the more response variability that is accounted for by the regression model, the closer the data points will fall to the fitted regression line.

With the same principle, it is natural to define the overall coefficient of determination as:

$$R^2 = \frac{\text{EWSS}}{\text{WSS}}. \quad (36)$$

It can be interpreted as the proportion of the within-group response variation explained (accounted for) by the fitted mixture of regression. Based on (28),  $R^2$  is related to  $R_1^2, \dots, R_k^2$  by the following relation:

$$R^2 = \frac{\sum_{g=1}^k \text{ESS}_g}{\text{WSS}} = \frac{\sum_{g=1}^k \text{SS}_g \frac{\text{ESS}_g}{\text{SS}_g}}{\text{WSS}} = \frac{\sum_{g=1}^k \text{SS}_g R_g^2}{\text{WSS}} = \sum_{g=1}^k \frac{\text{SS}_g}{\text{WSS}} R_g^2. \quad (37)$$

According to (37),  $R^2$  can be seen as a weighted average of the local coefficients of determination  $R_1^2, \dots, R_k^2$  with weights  $\text{SS}_1/\text{WSS}, \dots, \text{SS}_k/\text{WSS}$  being the proportion of the within-group sum of squares due to each group.

## 6 Analyses on Artificial Data

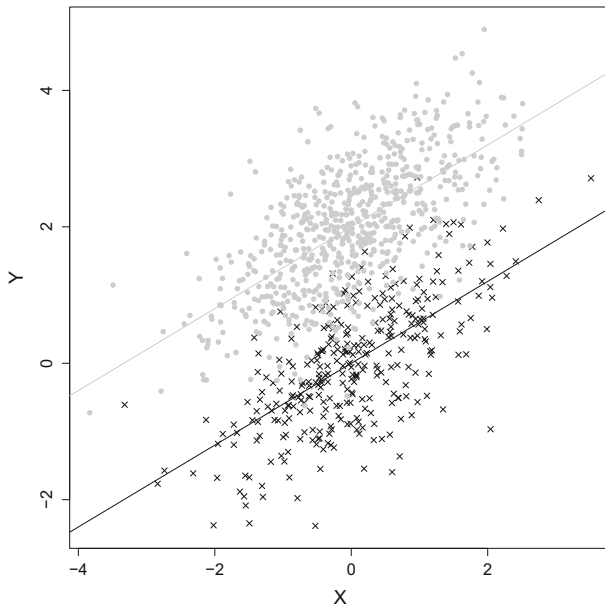
To find out more about the three terms of the decomposition proposed in (33), and to evaluate the behavior of these terms under violations of the model assumptions, applications on artificial data are here considered. The analyses are performed in R (R Core Team 2016). MRFC and MRRC are fitted via the `cwm()` function of the **flexCWM** package (Mazza et al. 2018), while MRCV are fitted via the `flexmix()` function of the **flexmix** package (Leisch 2004; Grun et al. 2008). These functions implement the EM algorithm to find ML estimates of the parameters (cf. Section 3). Among the possible initialization strategies for the EM algorithm (see, e.g., Biernacki et al. 2003; Karlis and Xekalaki 2003; Bagnato and Punzo 2013), a random initialization is repeated 20 times from different random positions and the value maximizing the observed-data log-likelihood among these 20 runs is selected.

### 6.1 Understanding the Decomposition

In the first illustrative example, artificial data are considered to find out more about the role of the three terms of the decomposition in (33). To simplify the graphical representations, a single covariate  $X$  ( $d = 1$ ) and two groups ( $k = 2$ ) are taken into account. The data generating process is a mixture of regressions where:

- The weights are  $\pi_1 = 0.3$  and  $\pi_2 = 0.7$ ;
- A standard normal distribution is used to generate the values of  $X$  in both groups;
- A normal distribution is adopted to generate the values of the dependent variable  $Y$ ;
- The two regression lines have intercepts  $\beta_{10} = 0$  and  $\beta_{20}$ , the same slope  $\beta_{11} = \beta_{21} = \beta_1$ , and the same conditional standard deviation  $\sigma_1 = \sigma_2 = \sigma$ .

The experimental conditions are the intercept in the second group ( $\beta_{20} \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ ), the common slope ( $\beta_1 \in \{0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2\}$ ), and the common conditional standard deviation ( $\sigma \in \{0.1, 0.4, 0.7, 1\}$ ). These experimental conditions cover the aspects the three terms of the decomposition are based on: the difference between  $\beta_{10}$  and  $\beta_{20}$  is related to the BSS term in (25), and the slope  $\beta_1$  of the parallel regression lines



**Fig. 2** Section 6.1. Example of scatter plot in the case  $\beta_{20} = 2$ ,  $\beta_1 = 0.6$ , and  $\sigma = 0.7$

affects the EWSS term in formula (28), while the conditional standard deviation  $\sigma$  impacts the RWSS term in (29).

One hundred datasets, each of size  $n = 1000$ , have been generated for each of the 264 combinations of the conditions above. Figure 2 shows an example of generated dataset related to the following combination of experimental conditions:  $\beta_{20} = 2$ ,  $\beta_1 = 0.6$ , and  $\sigma = 0.7$ .

On each generated dataset, a MRFC with  $k = 2$  components is fitted and the three terms NBSS, NEWSS, and NRWSS are computed. Figure 3 displays the ternary diagrams of the obtained results. Each of these diagrams contains the same points, but their color (in grayscale) changes based on the considered experimental factor. In these diagrams, each of the 264 points is related to a particular combination  $(\beta_{20}, \beta_1, \sigma)$ , and the point is obtained by averaging the triplets (NBSS, NEWSS, NRWSS) related to the 100 replications for the considered combination.

From Fig. 3a, we note that points roughly depart from the vertex NBSS as the second regression line approaches the first one (i.e., as  $\beta_{20} \rightarrow \beta_{10} = 0$ ). This happens because if the parallel lines move closer, then the group means of  $Y$ , i.e.,  $\bar{y}_1$  and  $\bar{y}_2$ , move closer too; consequently, the separation (on  $Y$ ) between groups reduces and the BSS term in (25) decreases too. From Fig. 3b, we note that points roughly depart from the vertex NEWSS as the positive common slope of the regression lines tends to 0. This happens because if  $\beta_1 \rightarrow 0$ , then  $\mu(x; \beta_g) \rightarrow \bar{y}_g$ , with  $\beta_g = (\beta_{g0}, \beta_1)'$ ,  $g = 1, 2$ ; consequently,  $X$  is not useful (via the linear model) to explain  $Y$  in each group, and  $\text{EWSS} \rightarrow 0$ ; refer to formula (28). Finally, from Fig. 3c, we note that points roughly depart from the vertex NRWSS as the local conditional variability of  $Y$  tends to vanish (i.e., as  $\sigma \rightarrow 0$ ). This happens because if  $\sigma \rightarrow 0$ , then the observed couples  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , tend to lie on one of the local regression lines and  $\text{RWSS} \rightarrow 0$ ; refer to formula (29).

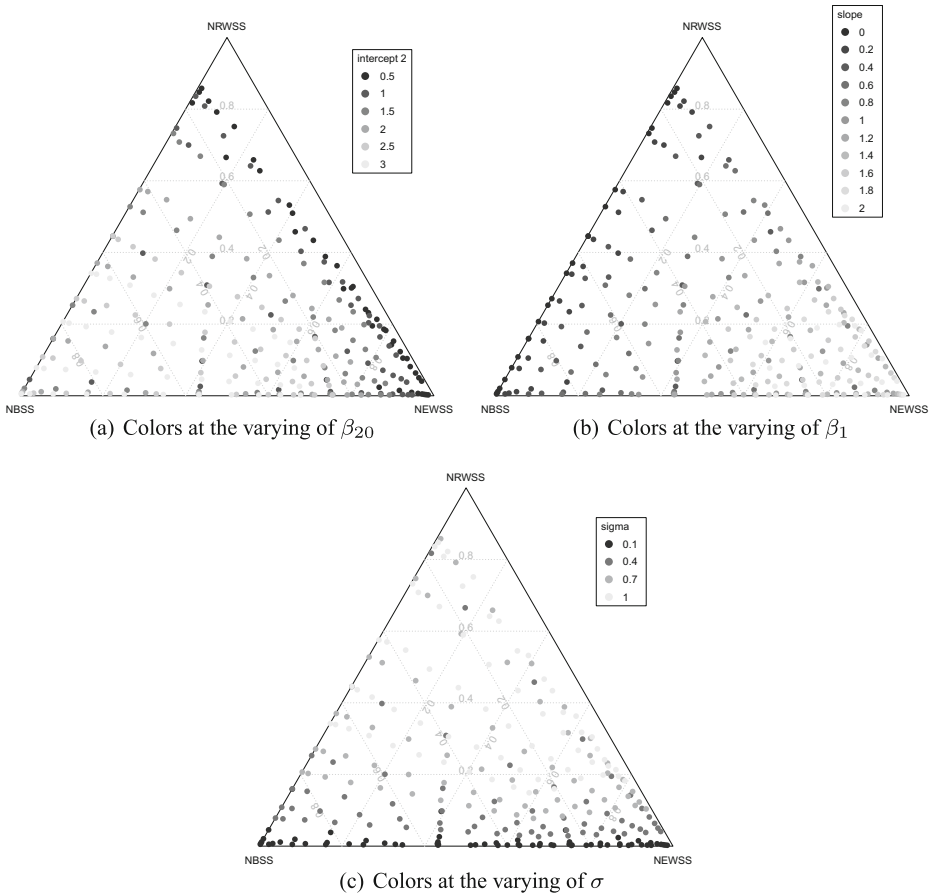


Fig. 3 Section 6.1. Average (over 100 replications) of the decomposition terms from the fitted MRFC models

### 6.2 Atypical Points and Departures from Conditional Local Normality

In the second illustrative example, artificial data are considered to evaluate the behavior of the decomposition in (33) with respect to the presence of atypical observations and departures from conditional normality of  $Y|X = x, G = g, g = 1, \dots, k$  (local conditional normality).

In regression analysis, atypical observations in  $Y|x$  represent model failure, and such observations are called outliers, while atypical observations with respect to  $X$  are called leverage points. There are two types of leverage points: good and bad. A bad leverage point is a regression outlier that has an  $x$  value that is atypical among the values of  $X$  as well. A good leverage point is a point that is unusually large or small among the  $X$  values but is not a regression outlier, i.e.,  $x$  is atypical but the corresponding  $y$  fits the model quite well. A point like this is called good because it improves the precision of the regression coefficients (Rousseeuw and Van Zomeren 1990, p. 635). Each point  $(x', y)$  can be considered as belonging to one of the four categories indicated in Table 1.

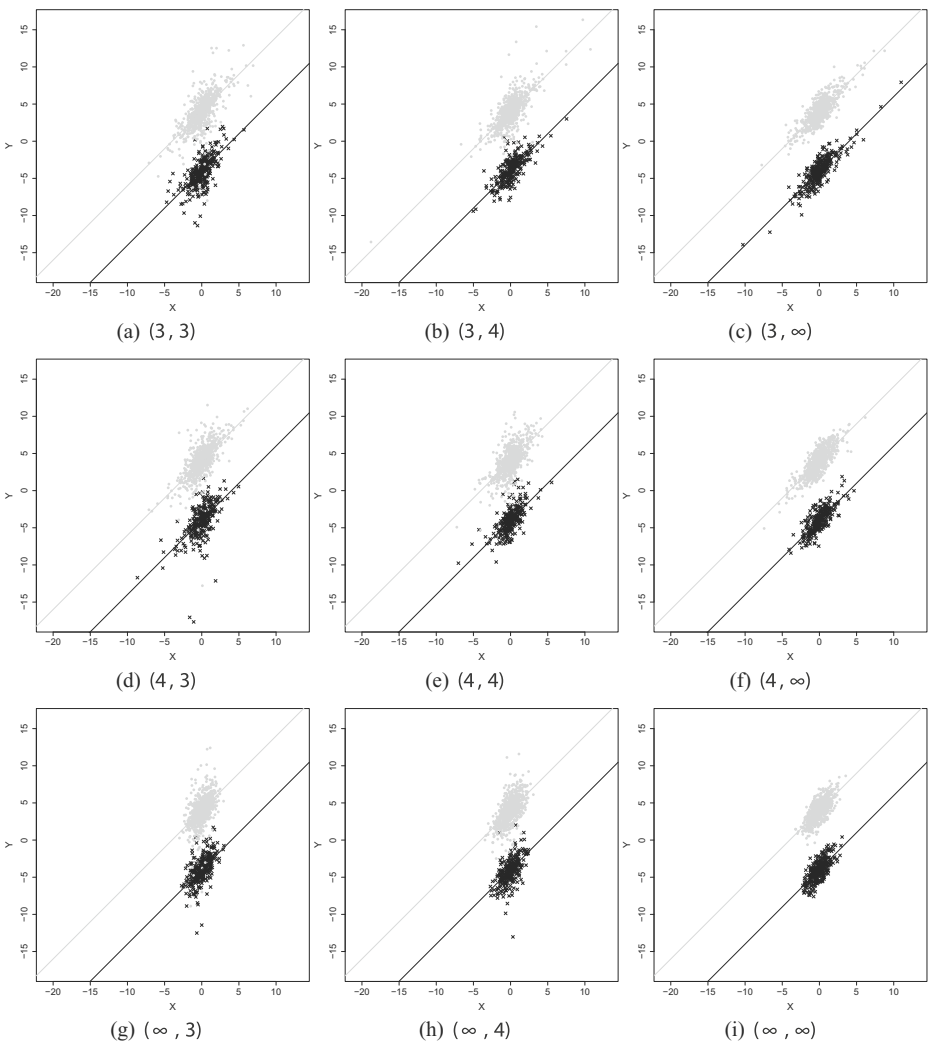
As in Section 6.1, a single covariate  $X$  ( $d = 1$ ) and two groups ( $k = 2$ ) are taken into account. The  $t$  distribution—with mean  $\mu \in \mathbb{R}$ , scale parameter  $\tau > 0$ , and degrees

**Table 1** Categorization for points in a regression analysis

	Leverage (on $X$ )	
Outlier (on $Y x$ )	Yes	No
Yes	Bad leverage	Outlier
No	Good leverage	Typical (bulk of the data)

of freedom  $\nu > 2$ —is considered to introduce departures from normality and the possible presence of atypical observations. It is important to recall that the  $t$ -distribution approaches the normal distribution, with mean  $\mu$  and standard deviation  $\tau$ , as  $\nu \rightarrow \infty$  (see, e.g., Lange et al. 1989). The data generating process is a mixture of regressions where:

- The weights are  $\pi_1 = 0.3$  and  $\pi_2 = 0.7$ ;



**Fig. 4** Section 6.2. Examples of generated scatters as a function of  $(\nu_X, \nu_Y)$

- The values of  $X$  are generated by a  $t$ -distribution with mean  $\mu_X = 0$ , scale parameter  $\tau_X = 1$ , and  $\nu_X$  degrees of freedom;
- Two different  $t$ -distributions are adopted to generate the values of the dependent variable  $Y$  in the two groups;
- The two regression lines have intercepts  $\beta_{10} = -4$  and  $\beta_{20} = 4$ , the same slope  $\beta_{11} = \beta_{21} = 1$ , the same conditional scale parameter  $\tau_Y = 1$ , and the same degrees of freedom  $\nu_Y$ .

The experimental conditions are  $\nu_X \in \{3, 4, \infty\}$  and  $\nu_Y \in \{3, 4, \infty\}$ . Their combination gives rise to nine different scenarios. These scenarios cover all the types of data categorized in Table 1: typical data (with respect to MRFC, MRCV, and MRRC models) when  $\nu_X \rightarrow \infty$  and  $\nu_Y \rightarrow \infty$ , good leverage points when  $\nu_X < \infty$ , outliers when  $\nu_Y < \infty$ , and bad leverage points when  $\nu_X < \infty$  and  $\nu_Y < \infty$ .

One hundred datasets, each of size  $n = 1000$ , have been generated for each of the 9 scenarios. Figure 4 shows examples of generated data for each scenario. On each generated dataset, MRFC, MRCV, and MRRC models, all with  $k = 2$  components, are fitted and the terms NBSS, NEWSS, and NRWSS are computed. Figure 5 displays the ternary diagrams of the obtained results for each model. Each of these diagrams contains 9 triplets (NBSS, NEWSS, NRWSS), averaged over the 100 replications, each related to a particular scenario. Points into the diagrams are denoted as  $(\nu_X, \nu_Y)$ , with  $\nu_X, \nu_Y \in \{3, 4, \infty\}$ .

By comparing the three diagrams in Fig. 5, it is possible to note that the position of the points is pretty much the same regardless from the considered model, with a slightly worse performance, in terms of NESS, for the MRCV model. Therefore, the following considerations will apply to all the considered models. Taking  $(\infty, \infty)$  as a reference scenario, it is possible to note that the points go up into the ternary diagram (consequently, the NESS decreases) as  $\nu_Y$  goes down. This means that, as expected, outliers get NESS values worse. At the same time, it is also interesting to note as good leverage points make NESS values slightly better; compare the position of the pairs  $(4, \infty)$  and  $(3, \infty)$  with respect to  $(\infty, \infty)$ . Finally, pairs where both  $\nu_X$  and  $\nu_Y$  are finite, i.e., scenarios including bad leverage points, are located closer to the NRWSS vertex, as expected.

## 7 Illustration on Tourism Data

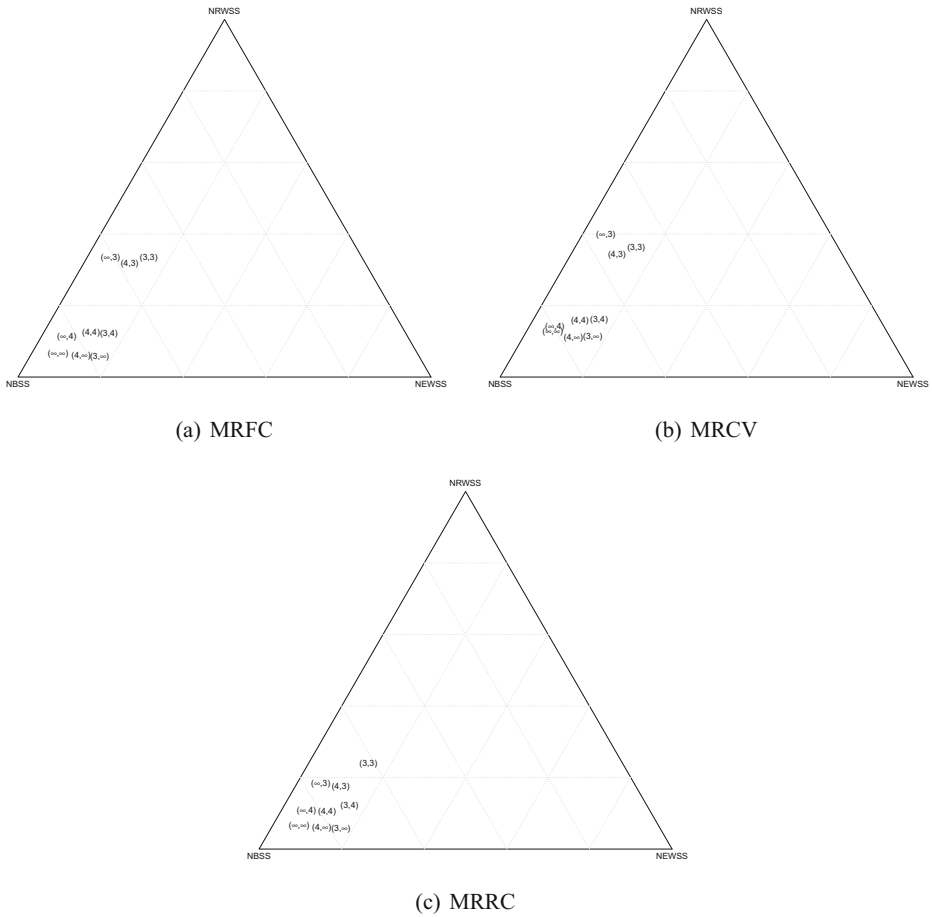
This application focuses on  $n = 180$  monthly data concerning *tourist overnights* ( $X$ , data in millions) and *attendance at museums and monuments* ( $Y$ , data in millions) in Italy over the 15-year period spanning from January 1996 to December 2010. These data, available at <http://www.economia.unict.it/punzo/Data.htm>, have been recently analyzed by Cellini and Cuccia (2013) and Ingrassia et al. (2014).

The scatter plot of the data is shown in Fig. 6; it gives strong evidence of both group-structure and relationships of  $Y$  on  $X$ .

Motivated by this consideration, we fit MRFC, MRCV, and MRRC for  $k \in \{1, \dots, 4\}$ , resulting in 12 different models. As concerns the MRRC, a normal distribution is considered for  $X$  in each group (see, e.g., Punzo and Ingrassia 2016 and Dang et al. 2017).

When using mixtures of regressions, and mixture models in general, some objective criterion is necessary for selecting the number of mixture components  $k$  for data under consideration. The Bayesian information criterion (BIC; Schwarz 1978) is the most commonly used for this purpose and is given by:

$$\text{BIC} = -2l(\hat{\psi}) + m \ln(n),$$

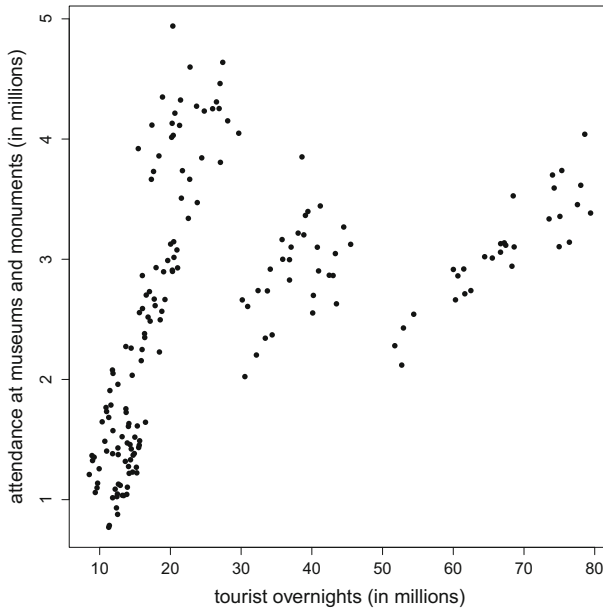


**Fig. 5** Section 6.2. Averages (over 100 replications) of the decomposition terms from the fitted models. Each point is represented by a pair  $(\nu_X, \nu_Y)$ , with  $\nu_X, \nu_Y \in \{3, 4, \infty\}$

where  $l(\hat{\psi})$  is the (maximized) observed-data log-likelihood while  $m$  is the number of free parameters. Note that, while the likelihood for MRFC and MRCV is a product of conditional probabilities  $p(y_i | \mathbf{x}_i; \psi)$ , the likelihood for MRRC is a product of joint probabilities  $p(\mathbf{x}_i, y_i; \psi)$ ; therefore, values of  $l(\hat{\psi})$  and BIC can be compared between MRFC and MRCV, but not with respect to the MRRC. Operationally, this means we can use the BIC to select between MRFC and MRCV too. With respect to these latter models, it is finally important to underline that, given  $k$ , MRFC in (1) can be thought as nested in the MRCV in (2).

Values of  $m$ ,  $l(\hat{\psi})$ , and BIC for the fitted models are reported in Table 2. Bold numbers in Table 2(c) highlight the best BIC value among the fitted MRFC and MRCV models (whose likelihoods can be compared) and among the fitted MRRC models. The selected models are the MRCV and the MRRC with  $k = 4$  components; they are represented in Fig. 7 in terms of regression lines and MAP classification of the observations; points are displayed as numbers denoting the MAP group membership.



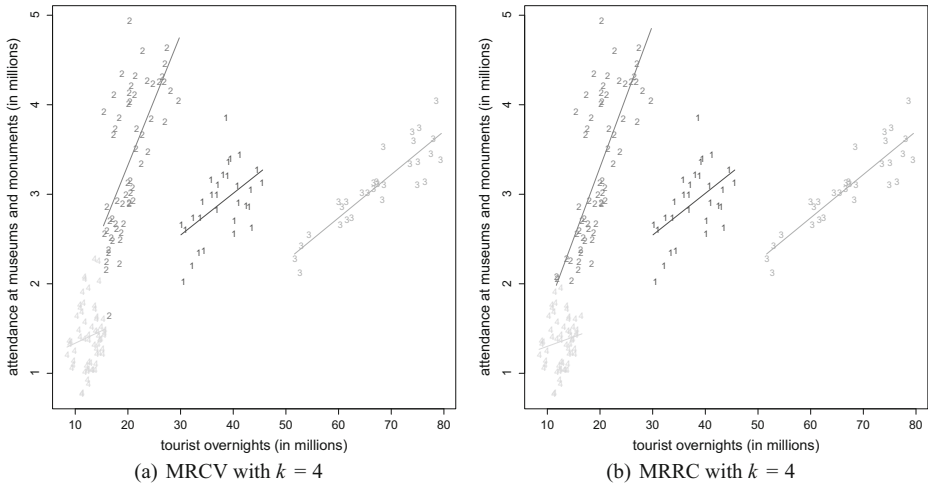


**Fig. 6** Tourism data. Scatter plot

The classifications from the two models are similar enough, with slight differences only with respect to the composition of groups 2 and 4. With respect to these models, it is interesting to note the good agreement between clusters and months (see Table 3). In detail, with

**Table 2** Tourism data. Values of  $m$ ,  $l(\hat{\psi})$ , and BIC for the fitted mixtures of regressions. Bold numbers in Table 2(c) highlight the BIC values of the selected models

$k$	MRFC	MRCV	MRRC
(a) $m$			
1	3	3	5
2	7	9	11
3	11	15	17
4	15	21	23
(b) $l(\hat{\psi})$			
1	-236.807	-236.812	-1029.649
2	-202.885	-132.560	-855.708
3	-182.515	-101.732	-816.155
4	-179.357	-79.588	-782.143
(c) BIC			
1	489.192	489.203	2085.262
2	442.120	306.664	1768.539
3	422.152	270.972	1720.591
4	436.609	<b>252.650</b>	<b>1683.725</b>



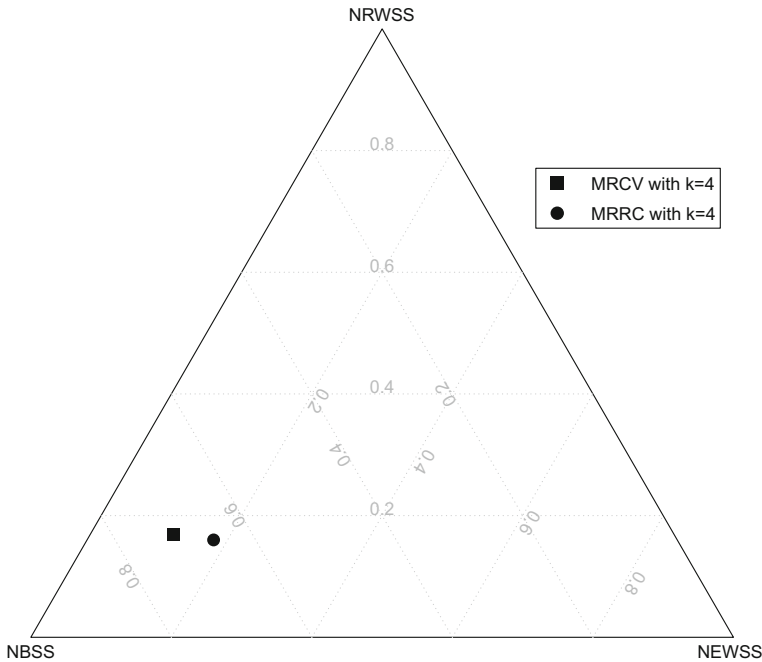
**Fig. 7** Tourism data. Scatter plots with regression lines and MAP classification of the observations from the models selected by the BIC

reference to the MRCV, only one unit in February, which concerns the year 2008, and three units in March, which concern the years 1996, 1998, and 1999, are assigned to a different group (see Table 3(a)). With reference to the MRRC, only two units in November, which concern the years 2006 and 2010, are assigned to a different group (see Table 3 and compare with Ingrassia et al. 2014, p. 170).

To determine how much the obtained clusters are separated on  $Y$ , and how well the selected models fit the data, it is useful to consider the measures introduced in Section 4. Figure 8 shows the ternary diagram containing the triplets (NBSS, NEWSS, NRWSS) of the selected models. The displayed triplets are (0.712, 0.119, 0.169) for the MRCV, and (0.660, 0.180, 0.160) for the MRRC. In terms of proportion of the total variability on  $Y$  explained by the fitted model, as measured by NESS in (34), the MRRC (with NESS = 0.840) performs slightly better than the MRCV (with NESS = 0.831), and this is visually

**Table 3** Tourism data. Relation between the clusters from the selected MRCV and MRRC and the months

Group	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
(a) MRCV with $k = 4$ components												
1	0	0	0	0	0	15	0	0	15	0	0	0
2	0	1	12	15	15	0	0	0	0	15	0	0
3	0	0	0	0	0	0	15	15	0	0	0	0
4	15	14	3	0	0	0	0	0	0	0	15	15
(b) MRRC with $k = 4$ components												
1	0	0	0	0	0	15	0	0	15	0	0	0
2	0	0	15	15	15	0	0	0	0	15	2	0
3	0	0	0	0	0	0	15	15	0	0	0	0
4	15	15	0	0	0	0	0	0	0	0	13	15



**Fig. 8** Tourism data. Ternary diagram of the triplets (NBSS, NEWSS, NRWSS) from the mixtures of regressions selected by the BIC

confirmed by a point for the MRRC which is slightly further from the NRWSS vertex. Even if the two models have similar NESS values, they have different behaviors in terms of NBSS and NEWSS which, according to (34), are the components of NESS. In particular, as concerns the MRCV, the  $\text{NBSS}/\text{NESS} \cdot 100 = 85.7\%$  of the explained variability is due to the clustering on  $Y$ , as measured by NBSS. The analogous percentage for the MRRC is lower (78.6%). Indeed, the MRCV point in Fig. 8 lies closer to the vertex NBSS than the MRRC point.

Given the clustering provided by the fitted model, i.e., given the values of  $\hat{z}_{ig}$ , to evaluate how close the data are to the fitted regression lines, it is useful to refer to the local coefficients of determination introduced in Section 5.4. For the MRCV, the local coefficients of determination are  $R_1^2 = 0.267$ ,  $R_2^2 = 0.446$ ,  $R_3^2 = 0.807$ , and  $R_4^2 = 0.023$ . A good fit can be noted in the third group, where the regression lines account for 80.7% of the local sum of squares  $SS_3$ . The overall coefficient of determination is  $R^2 = 0.412$ . The third group contributes to this value with weight  $SS_3/\text{WSS} = 0.106$ ; refer to (37). The other groups take part in the overall  $R^2$  with weights  $SS_1/\text{WSS} = 0.084$ ,  $SS_2/\text{WSS} = 0.674$ , and  $SS_4/\text{WSS} = 0.136$ . For the MRRC, the local coefficients of determination are  $R_1^2 = 0.267$ ,  $R_2^2 = 0.572$ ,  $R_3^2 = 0.807$ , and  $R_4^2 = 0.020$ . With respect to the MRCV,  $R_1^2$  and  $R_3^2$  are the same (see also Fig. 6), and  $R_4^2$  is slightly lower, while  $R_2^2$  is quite greater. The overall coefficient of determination is  $R^2 = 0.530$ , quite greater than the overall  $R^2$  for the MRCV. This improvement is due to the greater weight (0.764) associated with  $R_2^2$ . The other groups participate with weights  $SS_1/\text{WSS} = 0.071$ ,  $SS_3/\text{WSS} = 0.090$ , and  $SS_4/\text{WSS} = 0.075$ .

## 8 Conclusions and Discussion

When we use mixtures of regressions, the aim is twofold. First, as in the classical use of clustering/classification techniques, we want a method explaining the unobserved heterogeneity via the identification of homogeneous groups of observations. Second, as in the classical use of regression models, we hope that the inclusion of covariates explains more variation in the dependent variable. A mixture of regressions performs well if both these aspects are accounted for.

In this paper, for classical classes of mixtures of linear regressions, we proposed a three-term decomposition of the total sum of squares when the parameters are estimated with the expectation-maximization (EM) algorithm, within a maximum likelihood framework, under normally distributed errors in each mixture component. Based on this decomposition, we also introduced a measure for the explained within-group response variation (NEWSS), a measure of association between the response variable and the latent groups (NBSS), and an overall measure, collectively referred to as explained variation (NESS), considering NEWSS and NBSS together. Moreover, we introduced local and overall coefficients of determination to further evaluate how well the model fits the data group-by-group but also taken as a whole. The application to real data in Section 7 illustrated the use and the usefulness of our measures.

Finally, we remark that a natural extension of the ideas proposed herein would be the definition of “adjusted” local and overall coefficients of determination to be used—as the classical adjusted  $R^2$  for the standard linear regression model whose parameters are estimated by least squares—as comparative measures of suitability of models with alternative nested/nonnested sets of covariates (de Amorim 2016). However, groups are unknown and they change every time the model is estimated with a different set of covariates; this would make adjusted indexes senseless in our context.

## References

- Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell: Blackburn Press.
- Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1), 243–256.
- Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the  $k$ -bumps algorithm. *Computational Statistics*, 28(4), 1571–1597.
- Berta, P., Ingrassia, S., Punzo, A., Vittadini, G. (2016). Multilevel cluster-weighted models for the evaluation of hospitals. *METRON*, 74(3), 275–292.
- Biernacki, C., Celeux, G., Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3–4), 561–575.
- Buse, A. (1973). Goodness of fit in generalized least squares estimation. *The American Statistician*, 27(3), 106–108.
- Cameron, A.C., & Windmeijer, F.A.G. (1996).  $R$ -squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2), 209–220.
- Cameron, A.C., & Windmeijer, F.A.G. (1997). An  $R$ -squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2), 329–342.
- Cellini, R., & Cuccia, T. (2013). Museum and monument attendance and tourism flow: a time series approach. *Applied Economics*, 45, 3473–3482.
- Cerdeira, J.O., Martins, M.J., Silva, P.C. (2012). A combinatorial approach to assess the separability of clusters. *Journal of Classification*, 29(1), 7–22.
- Chatterjee, S., & Hadi, A.S. (2006). *Regression Analysis by Example, volume 607 of Wiley Series in Probability and Statistics*. Hoboken: Wiley.

- Dang, U.J., Punzo, A., McNicholas, P.D., Ingrassia, S., Browne, R.P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, 34(1), 4–34.
- Davidson, R., & MacKinnon, J.G. (2004). *Econometric Theory and Methods*. Oxford: Oxford University Press.
- Dayton, C.M., & Macready, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401), 173–178.
- de Amorim, R.C. (2016). A survey on feature weighting based k-means algorithms. *Journal of Classification*, 33(2), 210–242.
- Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1), 1–38.
- DeSarbo, W.S., & Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2), 249–282.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.
- Gershensfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, 808(1), 18–24.
- Grün, B., & Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4), 1–35.
- Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2–3), 107–145.
- Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2), 273–296.
- Hosmer, D.W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics-Theory and Methods*, 3(10), 995–1006.
- Huitema, B.E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, volume 608 of Wiley Series in Probability and Statistics*. New Jersey: Wiley.
- Ingrassia, S., & Punzo, A. (2016). Decision boundaries for mixtures of regressions. *Journal of the Korean Statistical Society*, 45(2), 295–306.
- Ingrassia, S., Minotti, S., Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, 29(3), 363–401.
- Ingrassia, S., Minotti, S.C., Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, 71, 159–182.
- Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S.C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, 32(1), 85–113.
- Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3–4), 577–590.
- Lange, K.L., Little, R.J.A., Taylor, J.M.G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, 84(408), 881–896.
- Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 1–18.
- Maddala, G.S. (1986). *Limited-Dependent and Qualitative Variables in Econometrics. Econometric Society Monographs*. Cambridge: Cambridge University Press.
- Mazza, A., & Punzo, A. (2018). Mixtures of multivariate contaminated normal regression models. *Statistical Papers*. <https://doi.org/10.1007/s00362-017-0964-y>.
- Mazza, A., Punzo, A., Ingrassia, S. (2018). **flexCWM**: Flexible cluster-weighted modeling. *Journal of Statistical Software*, 86(2), 1–30.
- Mazza, A., Battisti, M., Ingrassia, S., Punzo, A. (2019). Modeling return to education in heterogeneous populations. An application to Italy. In Greselin, I., Deldossi, L., Vichi, M., Bagnato, L. (Eds.) *Advances in Statistical Models for Data Analysis, Studies in Classification, Data Analysis and Knowledge Organization. Switzerland: Springer International Publishing*.
- McNicholas, P.D. (2016). Model-based clustering. *Journal of Classification*, 33(3), 331–373.
- Milligan, G.W., & Cheng, R. (1996). Measuring the influence of individual data points in a cluster analysis. *Journal of Classification*, 13(2), 315–335.
- Panagiotakis, C. (2015). Point clustering via voting maximization. *Journal of Classification*, 32(2), 212–240.
- Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, 14(3), 257–291.
- Punzo, A., & Ingrassia, S. (2015). Parsimonious generalized linear Gaussian cluster-weighted models. In Morlini, I.s., Minerva, T., Vichi, M. (Eds.) *Advances in Statistical Models for Data Analysis, Studies in Classification, Data Analysis and Knowledge Organization* (pp. 201–209). Switzerland: Springer International Publishing.

- Punzo, A., & Ingrassia, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, 31(3), 989–1013.
- Punzo, A., & McNicholas, P.D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, 34(2), 249–293.
- Punzo, A., Ingrassia, S., Maruotti, A. (2018). Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Statistics in Medicine*, 37(19), 2797–2808.
- Quandt, R.E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338), 306–310.
- Quandt, R.E., & Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364), 730–738.
- R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rezaee, M.R., Relieveltd, B.P.F., Reiber, J.H.C. (1998). A new cluster validity index for the fuzzy  $c$ -mean. *Pattern Recognition Letters*, 19(3–4), 237–246.
- Rousseeuw, P.J., & Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633–639.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Steinley, D., Hendrickson, G., Brusco, M.J. (2015). A note on maximizing the agreement between partitions: a stepwise optimal algorithm and some properties. *Journal of Classification*, 32(1), 114–126.
- Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P.D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7(1), 5–40.
- Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P.D. (2015). Cluster-weighted  $t$ -factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, 24(4), 623–649.
- Theodoridis, S., & Koutroumbas, K. (2008). *Pattern Recognition*. London: Academic Press.
- Veall, M.R., & Zimmermann, K.F. (1996). Pseudo- $R^2$  measures for some common limited dependent variable models. *Journal of Economic Surveys*, 10(3), 241–259.
- Wedel, M. (1990). *Clusterwise Regression and Market Segmentation: Developments and Applications*. Landbouwniversiteit te Wageningen.
- Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, 56(3), 362–375.
- Wedel, M., & De Sarbo, W. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(3), 21–55.
- Wedel, M., & Kamakura, W.A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*, 2nd edn. Boston: Kluwer Academic Publishers.
- Willett, J.B., & Singer, J.D. (1988). Another cautionary note about  $r^2$ : Its use in weighted least-squares regression analysis. *The American Statistician*, 42(3), 236–238.
- Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews*, 14(1), 101–116.
- Zarei, S., Mohammadpour, A., Ingrassia, S., Punzo, A. (2018). On the use of the sub-Gaussian  $\alpha$ -stable distribution in the cluster-weighted model. *Iranian Journal of Science and Technology, Transactions A: Science*. <https://doi.org/10.1007/s40995-018-0526-8>.