

# An EM Algorithm for the Student- $t$ Cluster-Weighted Modeling

Salvatore Ingrassia, Simona C. Minotti, and Giuseppe Incarbone

**Abstract** Cluster-Weighted Modeling is a flexible statistical framework for modeling local relationships in heterogeneous populations on the basis of weighted combinations of local models. Besides the traditional approach based on Gaussian assumptions, here we consider Cluster Weighted Modeling based on Student- $t$  distributions. In this paper we present an EM algorithm for parameter estimation in Cluster-Weighted models according to the maximum likelihood approach.

## 1 Introduction

The functional dependence between some input vector  $\mathbf{X}$  and output variable  $Y$  based on data coming from a heterogeneous population  $\Omega$ , supposed to be constituted by  $G$  homogeneous subpopulations  $\Omega_1, \dots, \Omega_g$ , is often estimated using methods able to model local behavior like Finite Mixtures of Regressions (FMR) and Finite Mixtures of Regressions with Concomitant variables (FMRC), see e.g. [Frühwirth-Schnatter \(2005\)](#). As a matter of fact, the purpose of these models is to identify groups by taking into account the local relationships between some response variable  $Y$  and some  $d$ -dimensional explanatory variables  $\mathbf{X} = (X_1, \dots, X_d)$ . Here we focus on a different approach called *Cluster-Weighted Modeling* (CWM) proposed first in the context of media technology in

---

S. Ingrassia · G. Incarbone

Dipartimento di Impresa, Culture e Società, Università di Catania Corso Italia 55, - 95129  
Catania, Italy

e-mail: [s.ingrassia@unict.it](mailto:s.ingrassia@unict.it); [gincarbo@diit.unict.it](mailto:gincarbo@diit.unict.it)

S.C. Minotti (✉)

Dipartimento di Statistica, Università di Milano-Bicocca Via Bicocca degli Arcimboldi 8 - 20126  
Milano, Italy

e-mail: [simona.minotti@unimib.it](mailto:simona.minotti@unimib.it)

order to recreate a digital violin with traditional inputs and realistic sound, see [Gershenfeld et al. \(1999\)](#).

From a statistical point of view, CWM can be regarded as a flexible statistical framework for capturing local behavior in heterogeneous populations. In particular, while FMR considers the conditional probability density  $p(y|\mathbf{x})$ , the CWM approach models the joint probability density  $p(\mathbf{x}, y)$  which is factorized as a weighted sum over  $G$  clusters. More specifically, in CWM each cluster contains an input distribution  $p(\mathbf{x}|\Omega_g)$  (i.e. a local model for the input variable  $\mathbf{X}$ ) and an output distribution  $p(y|\mathbf{x}, \Omega_g)$  (i.e. a local model for the dependence between  $\mathbf{X}$  and  $Y$ ). Some statistical properties of the Cluster-Weighted (CW) models have been established by [Ingrassia et al. \(2010\)](#); in particular it is shown that, under suitable hypotheses, CW models generalize FMR and FMRC. In the literature, CW models have been developed under Gaussian assumptions; moreover, a wider setting based on Student- $t$  distributions has been introduced and will be referred to as the Student- $t$  CWM. In this paper, we focus on the problem of parameter estimation in CW models according to the likelihood approach; in particular to this end we present an EM algorithm.

The rest of the paper is organized as follows: the Cluster-Weighted Modeling is introduced in Sect. 2; the EM algorithm for the estimation of the CWM is described in Sect. 3; in Sect. 4 we provide conclusions and ideas for further research.

## 2 The Cluster-Weighted Modeling

Let  $(\mathbf{X}, Y)$  be a pair of a random vector  $\mathbf{X}$  and a random variable  $Y$  defined on  $\Omega$  with joint probability distribution  $p(\mathbf{x}, y)$ , where  $\mathbf{X}$  is the  $d$ -dimensional input vector with values in some space  $\mathcal{X} \subseteq \mathbb{R}^d$  and  $Y$  is a response variable having values in  $\mathcal{Y} \subseteq \mathbb{R}$ . Thus  $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$ . Assume that  $\Omega$  can be partitioned into  $G$  disjoint groups, say  $\Omega_1, \dots, \Omega_G$ , that is  $\Omega = \Omega_1 \cup \dots \cup \Omega_G$ . *Cluster-Weighted Modeling* (CWM) decomposes the joint probability of  $(\mathbf{X}, Y)$  as:

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^G p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}|\Omega_g) \pi_g, \quad (1)$$

where  $\pi_g = p(\Omega_g)$  is the mixing weight of group  $\Omega_g$ ,  $p(\mathbf{x}|\Omega_g)$  is the probability density of  $\mathbf{x}$  given  $\Omega_g$  and  $p(y|\mathbf{x}, \Omega_g)$  is the conditional density of the response variable  $Y$  given the predictor vector  $\mathbf{x}$  and the group  $\Omega_g$ ,  $g = 1, \dots, G$ , see [Gershenfeld et al. \(1999\)](#). Vector  $\boldsymbol{\theta}$  denotes the set of all parameters of the model. Hence, the joint density of  $(\mathbf{X}, Y)$  can be viewed as a mixture of local models  $p(y|\mathbf{x}, \Omega_g)$  weighted (in a broader sense) on both the local densities  $p(\mathbf{x}|\Omega_g)$  and the mixing weights  $\pi_g$ . Throughout this paper we assume that the input-output relation can be written as  $Y = \mu(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon$ , where  $\mu(\mathbf{x}; \boldsymbol{\beta})$  is a given function of  $\mathbf{x}$  (depending also on some parameters  $\boldsymbol{\beta}$ ) and  $\varepsilon$  is a random variable with zero mean and finite variance.

Usually, in the literature about CWM, the local densities  $p(\mathbf{x}|\Omega_g)$  are assumed to be multivariate Gaussians with parameters  $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ , that is  $\mathbf{X}|\Omega_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ ,  $g = 1, \dots, G$ ; moreover, also the conditional densities  $p(y|\mathbf{x}, \Omega_g)$  are often modeled by Gaussian distributions with variance  $\sigma_{\varepsilon,g}^2$  around some deterministic function of  $\mathbf{x}$ , say  $\mu_g(\mathbf{x}; \boldsymbol{\beta}_g)$ ,  $g = 1, \dots, G$ . Thus  $p(\mathbf{x}|\Omega_g) = \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ , where  $\phi_d$  denotes the probability density of a  $d$ -dimensional multivariate Gaussian and  $p(y|\mathbf{x}, \Omega_g) = \phi(y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_{\varepsilon,g}^2)$ , where  $\phi$  denotes the probability density of a uni-dimensional Gaussian. Such model will be referred to as the Gaussian CWM. In the simplest case, the conditional densities are based on linear mappings  $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = \mathbf{b}'_g \mathbf{x} + b_{g0}$ , with  $\boldsymbol{\beta} = (\mathbf{b}'_g, b_{g0})'$ ,  $\mathbf{b}_g \in \mathbb{R}^d$  and  $b_{g0} \in \mathbb{R}$ , yielding the linear Gaussian CWM:

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^G \phi(y; \mathbf{b}'_g \mathbf{x} + b_{g0}, \sigma_{\varepsilon,g}^2) \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \pi_g. \quad (2)$$

Under suitable assumptions, one can show that model (2) leads to the same posterior probability of Finite Mixtures of Gaussians (FMG), Finite Mixtures of Regressions (FMR) and Finite Mixtures of Regressions with Concomitant variables (FMRC). In this sense, we shall say that CWM contains FMG, FMR and FMRC, as summarized in the following table, see [Ingrassia et al. \(2010\)](#) for details:

	$p(\mathbf{x} \Omega_g)$	$p(y \mathbf{x}, \Omega_g)$	assumption
FMG	Gaussian	Gaussian	linear relationship between $\mathbf{X}$ and $Y$
FMR	none	Gaussian	$(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , $g=1, \dots, G$
FMRC	none	Gaussian	$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = \pi$ , $g=1, \dots, G$

In order to provide more realistic tails for real-world data with respect to the Gaussian models and rely on a more robust estimation of parameters, the CWM with Student- $t$  components has been introduced by [Ingrassia et al. \(2010\)](#). Let us assume that in model (1) both  $p(\mathbf{x}|\Omega_g)$  and  $p(y|\mathbf{x}, \Omega_g)$  are Student- $t$  densities. In particular, we assume that  $\mathbf{X}|\Omega_g$  has a multivariate  $t$  distribution with location parameter  $\boldsymbol{\mu}_g$ , inner product matrix  $\boldsymbol{\Sigma}_g$  and  $\nu_g$  degrees of freedom, that is  $\mathbf{X}|\Omega_g \sim t_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ , and  $Y|\mathbf{x}, \Omega_g$  has a  $t$  distribution with location parameter  $\mu(\mathbf{x}; \boldsymbol{\beta}_g)$ , scale parameter  $\sigma_g^2$  and  $\zeta_g$  degrees of freedom, that is  $Y|\mathbf{x}, \Omega_g \sim t(\mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_g^2, \zeta_g)$ , so that

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^G t(y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_g^2, \zeta_g) t_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) \pi_g. \quad (3)$$

This implies that, for  $g = 1, \dots, G$ ,

$$\mathbf{X}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g, U_g \sim N_d\left(\boldsymbol{\mu}_g, \frac{\boldsymbol{\Sigma}_g}{u_g}\right),$$

$$Y|\mu(\mathbf{x}, \boldsymbol{\beta}_g), \sigma_g, \zeta_g, W_g \sim N\left(\mu(\mathbf{x}, \boldsymbol{\beta}_g), \frac{\sigma_g}{W_g}\right),$$

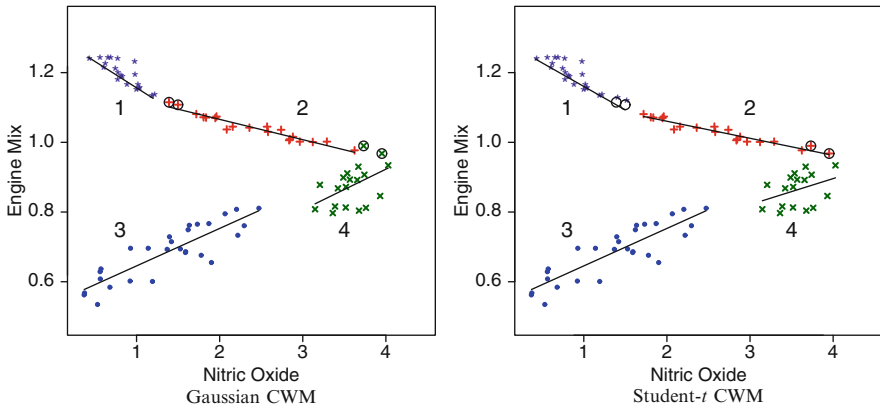
where  $U_g$  and  $W_g$  are independent random variables such that

$$U_g|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g \sim \Gamma\left(\frac{\nu_g}{2}, \frac{\nu_g}{2}\right) \quad \text{and} \quad W_g|\mu(\mathbf{x}, \boldsymbol{\beta}_g), \sigma_g, \zeta_g \sim \Gamma\left(\frac{\zeta_g}{2}, \frac{\zeta_g}{2}\right). \quad (4)$$

Model (3) will be referred to as the Student- $t$  CWM; the special case in which  $\mu(\mathbf{x}; \boldsymbol{\beta}_g)$  is some linear mapping will be called the *linear  $t$ -CWM*. In particular, we remark that the linear  $t$ -CWM defines a wide family of densities which includes mixtures of multivariate  $t$ -distributions and mixtures of regressions with Student- $t$  errors.

Obviously, the two models have a different behavior. An example is illustrated using the NO dataset, which relates the concentration of nitric oxide in engine exhaust to the equivalence ratio, see [Hurn et al. \(2003\)](#). Data have been fitted using both Gaussian and Student- $t$  CWM. The two classifications differ by four units, which are indicated by circles around them (two units classified in either group 1 or group 2; two units classified in either group 2 or group 4), see Fig. 1. In particular, there are two units that the Gaussian CWM classifies in group 4 but which are a little bit far from the other points of the same group; instead, such units are classified in group 2 by means of the Student- $t$  CWM. If we consider the following *index of weighted model fitting* (IWF)  $\mathcal{E}$  defined as:

$$\mathcal{E} = \left( \frac{1}{N} \sum_{n=1}^N \left[ y_n - \left( \sum_{g=1}^G \mu(\mathbf{x}_n; \boldsymbol{\beta}_g) p(\Omega_g | \mathbf{x}_n, y_n) \right) \right]^2 \right)^{1/2}, \quad (5)$$



**Fig. 1** Gaussian and Student- $t$  CW models for the NO dataset

we get  $\mathcal{E} = 0.108$  in the Gaussian case and  $\mathcal{E} = 0.086$  in the Student CWM. Thus the latter model showed a slightly better fit than the Gaussian CWM.

### 3 CWM Parameter Estimation Via the EM Algorithm

In this section we present the main steps of the EM algorithm in order to estimate the parameters of CWM according to the likelihood approach. We illustrate here the Student- $t$  distribution case; similar ideas can be easily developed in the Gaussian case. Let  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$  be a sample of  $N$  independent observation pairs drawn from (3) and set  $\tilde{\mathbf{X}} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ ,  $\mathbf{Y} = (y_1, \dots, y_N)$ . Then, the likelihood function of the Student- $t$  CWM is given by

$$L_0(\boldsymbol{\psi}; \tilde{\mathbf{X}}, \mathbf{Y}) = \prod_{n=1}^N p(\mathbf{x}_n, y_n; \boldsymbol{\psi}) = \prod_{n=1}^n \left[ \sum_{g=1}^G p(y_n | \mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g) p_d(\mathbf{x}_n; \boldsymbol{\theta}_g, v_g) f(w; \zeta_g) f(u; v_g) \pi_g \right], \quad (6)$$

where  $f(w; \zeta_g)$  and  $f(u; v_g)$  denote the density functions of  $W_g$  and  $U_g$  respectively given in (4). Maximization of  $L_0(\boldsymbol{\psi}; \tilde{\mathbf{X}}, \mathbf{Y})$  with respect to  $\boldsymbol{\psi}$ , for given data  $(\tilde{\mathbf{X}}, \mathbf{Y})$ , yields the estimate of  $\boldsymbol{\psi}$ . If we consider fully categorized data  $\{(\mathbf{x}_n, y_n, \mathbf{z}_n, u_n, w_n) : n = 1, \dots, N\}$ , then the complete-data likelihood function can be written in the form

$$L_c(\boldsymbol{\psi}; \tilde{\mathbf{X}}, \mathbf{Y}) = \prod_{n,g} p(y_n | \mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)^{z_{ng}} p_d(\mathbf{x}_n; \boldsymbol{\theta}_g, v_g)^{z_{ng}} f(w; \zeta_g)^{z_{ng}} f(u; v_g)^{z_{ng}} \pi_g^{z_{ng}},$$

where  $z_{ng} = 1$  if  $(\mathbf{x}_n, Y_n)$  comes from the  $g$ -th population and  $z_{ng} = 0$  elsewhere.

Let us take the logarithm, then after some algebra we get:

$$\mathcal{L}_c(\boldsymbol{\psi}; \tilde{\mathbf{X}}, \mathbf{Y}) = \ln L_c(\boldsymbol{\psi}; \tilde{\mathbf{X}}, \mathbf{Y}) = \mathcal{L}_{1c}(\boldsymbol{\beta}) + \mathcal{L}_{2c}(\boldsymbol{\theta}) + \mathcal{L}_{3c}(\zeta) + \mathcal{L}_{4c}(v) + \mathcal{L}_{5c}(\pi)$$

where

$$\mathcal{L}_{1c}(\boldsymbol{\beta}) = \sum_{n=1}^N \sum_{g=1}^G z_{ng} \frac{1}{2} \left[ -\ln 2\pi + \ln w_n - \ln \sigma_{\epsilon,g}^2 - w_n \frac{(y_n - \mathbf{b}'_g \mathbf{x}_n - b_{0g})^2}{\sigma_{\epsilon,g}^2} \right]$$

$$\begin{aligned}
\mathcal{L}_{2c}(\boldsymbol{\theta}) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} \frac{1}{2} \left[ -p \ln 2\pi + p \ln u_n - \ln |\boldsymbol{\Sigma}_g| - u_n (\mathbf{x}_n - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_g) \right] \\
\mathcal{L}_{3c}(\zeta) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} \left[ -\ln \Gamma \left( \frac{\zeta_g}{2} \right) + \left( \frac{\zeta_g}{2} \right) \ln \left( \frac{\zeta_g}{2} \right) + \frac{\zeta_g}{2} (\ln w_n - w_n) - \ln w_n \right] \\
\mathcal{L}_{4c}(v) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} \left[ -\ln \Gamma \left( \frac{v_g}{2} \right) + \left( \frac{v_g}{2} \right) \ln \left( \frac{v_g}{2} \right) + \frac{v_g}{2} (\ln u_n - u_n) - \ln u_n \right] \\
\mathcal{L}_{5c}(\pi) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} [\ln \pi_g].
\end{aligned}$$

The **E-step** on the  $(k + 1)$ -th iteration of the EM algorithm requires the calculation of the conditional expectation of the complete-data loglikelihood function  $\ln L_c(\boldsymbol{\psi}; \mathbf{X}, \mathbf{Y})$ , say  $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)})$ , evaluated using the current fit  $\boldsymbol{\psi}^{(k)}$  for  $\boldsymbol{\psi}$ . To this end, the quantities  $\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{Z_{ng} | \mathbf{x}_n, y_n\}$ ,  $\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{U_n | \mathbf{x}_n, \mathbf{z}_n\}$ ,  $\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{\ln U_n | \mathbf{x}_n, \mathbf{z}_n\}$ ,  $\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{W_n | y_n, \mathbf{z}_n\}$  and  $\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{\ln W_n | y_n, \mathbf{z}_n\}$  have to be evaluated, for  $n = 1, \dots, N$  and  $g = 1, \dots, G$ . It follows that

$$\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{Z_{ng} | \mathbf{x}_n, y_n\} = \tau_{ng}^{(k)} = \frac{\pi_g^{(k)} p(y_n | \mathbf{x}_n; \boldsymbol{\beta}_g^{(k)}, \zeta_g^{(k)}) p_d(\mathbf{x}_n; \boldsymbol{\theta}_g^{(k)}, v_g^{(k)})}{\sum_{j=1}^G \pi_j^{(k)} p(y_n | \mathbf{x}_n; \boldsymbol{\beta}_j^{(k)}, \zeta_j^{(k)}) p_d(\mathbf{x}_n; \boldsymbol{\theta}_j^{(k)}, v_j^{(k)})}$$

is the posterior probability that the  $n$ -th observation  $(\mathbf{x}_n, y_n)$  belongs to the  $g$ -th component of the mixture. Further, we have that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{U_n | \mathbf{x}_n, \mathbf{z}_n\} &= u_{ng}^{(k)} = \frac{v_g^{(k)} + d}{v_g^{(k)} + (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k)})' \boldsymbol{\Sigma}_g^{-1(k)} (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k)})} \\
\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{\ln U_n | \mathbf{x}_n, \mathbf{z}_n\} &= \ln u_{ng}^{(k)} + \left\{ \psi \left( \frac{v_g^{(k)} + d}{2} \right) - \ln \left( \frac{v_g^{(k)} + d}{2} \right) \right\} \\
\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{W_n | y_n, \mathbf{z}_n\} &= w_{ng}^{(k)} = \frac{\zeta_g^{(k)} + 1}{\zeta_g^{(k)} + \frac{(y_n - \mathbf{b}_g^{(k)} \mathbf{x}_n - b_{g0}^{(k)})^2}{\sigma_{\epsilon, g}^{2(k)}}} \\
\mathbb{E}_{\boldsymbol{\psi}^{(k)}} \{\ln W_n | y_n, \mathbf{z}_n\} &= \ln w_{ng}^{(k)} + \left\{ \psi \left( \frac{\zeta_g^{(k)} + 1}{2} \right) - \ln \left( \frac{\zeta_g^{(k)} + 1}{2} \right) \right\},
\end{aligned}$$

where  $\psi(s) = \{\partial \Gamma(s) / \partial s\} / \Gamma(s)$ .

In the **M-step**, on the  $(k + 1)$ -th iteration of the EM algorithm, we maximize the conditional expectation of the complete-data loglikelihood  $Q$ , with respect to

$\boldsymbol{\psi}$ . The solutions for the posterior probabilities  $\pi_g^{(k+1)}$  and the parameters  $\boldsymbol{\mu}_g^{(k+1)}$ ,  $\boldsymbol{\Sigma}_g^{(k+1)}$  of the local input densities  $p_d(\mathbf{x}_n | \boldsymbol{\theta}_g, \nu_g)$ , for  $g = 1, \dots, G$ , exist in a closed form and coincide with the case of mixtures of multivariate  $t$  distributions (Peel and McLachlan 2000), that is:

$$\begin{aligned}\pi_g^{(k+1)} &= \frac{1}{N} \sum_{n=1}^N \tau_{ng}^{(k)} \\ \boldsymbol{\mu}_g^{(k+1)} &= \frac{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{u}_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{u}_{ng}^{(k)}} \\ \boldsymbol{\Sigma}_g^{(k+1)} &= \frac{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{u}_{ng}^{(k)} (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{u}_{ng}^{(k)}}.\end{aligned}$$

The updates  $\mathbf{b}_g^{(k+1)}$ ,  $b_{g0}^{(k+1)}$  and  $\sigma_{\epsilon,g}^{2(k+1)}$  of the parameters for the local output densities  $p(y_n | \mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)$ , for  $g = 1, \dots, G$ , are obtained by the solution of the equations

$$\begin{aligned}\frac{\partial \mathbb{E}_{\psi^{(k)}} \{ \mathcal{L}_c(\boldsymbol{\psi} | \mathbf{x}_n, y_n) \}}{\partial \mathbf{b}'_g} &= \mathbf{0}' \\ \frac{\partial \mathbb{E}_{\psi^{(k)}} \{ \mathcal{L}_c(\boldsymbol{\psi} | \mathbf{x}_n, y_n) \}}{\partial b_{g0}} &= 0 \\ \frac{\partial \mathbb{E}_{\psi^{(k)}} \{ \mathcal{L}_c(\boldsymbol{\psi} | \mathbf{x}_n, y_n) \}}{\partial \sigma_{\epsilon,g}^2} &= 0\end{aligned}$$

and it can be proved that they are given by

$$\begin{aligned}\mathbf{b}'_g{}^{(k+1)} &= \left( \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} y_n \mathbf{x}'_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} - \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} y_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} \mathbf{x}'_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} \right) \\ &\quad \cdot \left( \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} \mathbf{x}_n \mathbf{x}'_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} - \left( \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} \mathbf{x}'_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} \right)^2 \right)^{-1} \\ b_{g0}^{(k+1)} &= \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} y_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} - \mathbf{b}_g^{(k+1)} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} \mathbf{x}'_n}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}} \\ \sigma_{\epsilon,g}^{2(k+1)} &= \frac{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)} [y_n - (\mathbf{b}_g^{(k+1)} \mathbf{x}'_n + b_{g0}^{(k+1)})]^2}{\sum_{n=1}^N \tau_{ng}^{(k)} w_{ng}^{(k)}}.\end{aligned}$$

The updates  $v_g^{(k+1)}$  and  $\zeta_g^{(k+1)}$  for the degrees of freedom  $v_g$  and  $\zeta_g$  need to be computed iteratively and are given by the solutions of the following equations, respectively:

$$-\psi\left(\frac{v_g}{2}\right) + \ln\left(\frac{v_g}{2}\right) + 1 + \frac{1}{\sum_{n=1}^N \tau_{ng}^{(k)}} \cdot \sum_{n=1}^N \tau_{ng}^{(k)} \left( \ln u_{ng}^{(k)} - u_{ng}^{(k)} \right) + \\ + \psi\left(\frac{v_g^{(k)} + d}{2}\right) - \ln\left(\frac{v_g^{(k)} + d}{2}\right) = 0$$

and

$$-\psi\left(\frac{\zeta_g}{2}\right) + \ln\left(\frac{\zeta_g}{2}\right) + 1 + \frac{1}{\sum_{n=1}^N \tau_{ng}^{(k)}} \cdot \sum_{n=1}^N \tau_{ng}^{(k)} \left( \ln w_{ng}^{(k)} - w_{ng}^{(k)} \right) + \\ + \psi\left(\frac{\zeta_g^{(k)} + 1}{2}\right) - \ln\left(\frac{\zeta_g^{(k)} + 1}{2}\right) = 0$$

where  $\psi(s) = \{\partial \Gamma(s) / \partial s\} / \Gamma(s)$ .

## 4 Concluding Remarks

CWM is a flexible approach for clustering and modeling functional relationships based on data coming from a heterogeneous population. Besides traditional modeling based on Gaussian assumptions, in order to provide more realistic tails for real-world data and rely on a more robust estimation of parameters, we considered also CWM based on Student- $t$  distributions. Here we focused on a specific issue concerning parameter estimates according to the likelihood approach. For this aim, in this paper we presented an EM algorithm which has been implemented in Matlab and R language for both the Gaussian and the Student- $t$  case.

Our simulations showed that the initialization of the algorithms is quite critical, in particular the initial guess has been chosen according to both a preliminary clustering of data using a  $k$ -means algorithm and a random grouping of data, but our numerical studies pointed out that there is no overwhelming strategy. Similar results have been obtained by [Faria and Soromenho \(2010\)](#) in the area of mixtures of regressions, where the performance of the algorithm depends essentially on the configuration of the true regression lines and the initialization of the algorithms.

Finally we remark that, in order to reduce such critical aspects, suitable constraints on the eigenvalues of the covariance matrices could be implemented. This provides ideas for future work.



## References

- Faria S, Soromenho G (2010) Fitting mixtures of linear regressions. *J Stat Comput Simulat* 80:201–225
- Frühwirth-Schnatter S (2005) Finite mixture and markov switching models. Springer, Heidelberg
- Gershenfeld N, Schöner B, Metois E (1999) Cluster-weighted modeling for time-series analysis. *Nature* 397:329–332
- Hurn M, Justel A, Robert CP (2003) Estimating mixtures of regressions. *J Comput Graph Stat* 12:55–79
- Ingrassia S, Minotti SC, Vittadini G (2010) Local statistical modeling via the cluster-weighted approach with elliptical distributions. *ArXiv*: 0911.2634v2
- Peel D, McLachlan GJ (2000) Robust mixture modelling using the  $t$  distribution. *Stat Comput* 10:339–348

Challenges at the Interface of Data Analysis, Computer  
Science, and Optimization

Proceedings of the 34th Annual Conference of the  
Gesellschaft für Klassifikation e. V., Karlsruhe, July 21 -  
23, 2010

Gaul, W.; Geyer-Schulz, A.; Schmidt-Thieme, B.; Kunze, J.  
(Eds.)

2012, XIV, 598 p. 163 illus., Softcover

ISBN: 978-3-642-24465-0