Maximum likelihood estimation of Gaussian cluster weighted models and relationships with mixtures of regression

Salvatore Ingrassia[a,1], Simona C. Minotti[b]

[a]*Dipartimento di Economia e Impresa, Università di Catania (Italy)*
Corso Italia, 55 - 95129 Catania (Italy), s.ingrassia@unict.it
[b]*Dipartimento di Statistica, Università di Milano-Bicocca (Italy)*
simona.minotti@unimib.it

Abstract

Cluster-weighted modeling (CWM) is a mixture approach for modeling the joint probability of a response variable and a set of explanatory variables. The parameters are estimated by means of the expectation-maximization algorithm according to the maximum likelihood approach. We show that, under suitable hypotheses, the maximization of the likelihood function of Gaussian cluster weighted models leads to the same parameter estimates of finite mixtures of regression and finite mixtures of regression with concomitant variables. In this sense, the latter ones can be considered as nested models of Gaussian cluster weighted models.

Keywords: Cluster-weighted modeling, finite mixtures of regression, EM-algorithm

1. Introduction

Cluster-weighted modeling (CWM) is a mixture approach to modeling the joint probability density of a response variable and a set of explanatory variables. The original formulation, proposed by Gershenfeld (1997) under Gaussian and linear assumptions, was developed in the context of media technology in order to build a digital violin with traditional inputs and realistic sound (Gershenfeld *et al.*, 1999; Gershenfeld, 1999; Schöner, 2000; Schöner and Gershenfeld, 2001). Wedel (2002) refers to such a model as the saturated mixture regression model. Ingrassia *et al.* (2011) reformulated CWM from a statistical point of view in a wide framework; in particular, under Gaussian assumptions (Gaussian CWM) we investigate the relationships between CWM and both finite mixtures of regression (FMR) (De Sarbo and Cron,

1988; McLachlan and Peel, 2000; Frühwirth-Schnatter, 2005) and finite mixtures of regression with concomitant variables (FMRC) (Dayton and Macready, 1988; Wedel, 2002).

The parameters of cluster weighted models are estimated through the EM algorithm according to the maximum likelihood approach. In this paper, we show that, under suitable hypotheses, the maximization of the likelihood function of Gaussian CWM leads to the same parameter estimates of FMR and FMRC. In this sense, FMR and FMRC can be considered as nested models of Gaussian CWM.

The remainder of the paper is organized as follows. In Section 2, we review CWM as a general framework for mixture modeling. In Section 3, we analyse the complete-data likelihood function of Gaussian CWM and derive the main steps of the EM algorithm for parameter estimation. In Section 4 we show that, under suitable hypotheses, the maximization of the likelihood function of Gaussian CWM leads to the same parameter estimates of FMR and FMRC. Finally, in Section 5, we provide some conclusions and further research.

2. Cluster-Weighted Modeling

Let $(\boldsymbol{X}, Y)$ be the pair of random vector $\boldsymbol{X}$ and random variable Y defined on Ω with joint probability distribution $p(\boldsymbol{x}, y)$, where $\boldsymbol{X}$ 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, $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d+1}$. Suppose that Ω can be partitioned into G disjoint groups, say $\Omega_1, \ldots, \Omega_G$, that is $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$. CWM decomposes the joint probability $p(\boldsymbol{x}, y)$ as follows:

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

where $p(y|\boldsymbol{x}, \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\boldsymbol{x}$ and Ω_g, $p(\boldsymbol{x}|\Omega_g)$ is the probability density of $\boldsymbol{x}$ given Ω_g, $\pi_g = p(\Omega_g)$ is the mixing weight of Ω_g, ($\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$), $g = 1, \ldots, G$, and $\boldsymbol{\theta}$ denotes the set of all parameters of the model. Hence, the joint density of $(\boldsymbol{X}, Y)$ can be viewed as a mixture of local models $p(y|\boldsymbol{x}, \Omega_g)$ weighted (in a broader sense) on both local densities $p(\boldsymbol{x}|\Omega_g)$ and mixing weights π_g.

The posterior probability $p(\Omega_g|\boldsymbol{x}, y)$ of unit $(\boldsymbol{x}, y)$ to come from the g-th group $(g = 1, ..., G)$ is given by:

$$p(\Omega_g|\boldsymbol{x}, y) = \frac{p(\boldsymbol{x}, y, \Omega_g)}{p(\boldsymbol{x}, y)} = \frac{p(y|\boldsymbol{x}, \Omega_g)p(\boldsymbol{x}|\Omega_g)\pi_g}{\sum_{j=1}^{G} p(y|\boldsymbol{x}, \Omega_j)p(\boldsymbol{x}|\Omega_j)\pi_j}. \tag{2}$$

In particular, the classification of each unit depends on both marginal and conditional densities.

In the traditional framework, local densities $p(\boldsymbol{x}|\Omega_g)$ are assumed to be multivariate Gaussian with parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, that is $\boldsymbol{X}|\Omega_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $g = 1, \ldots, G$. Moreover, conditional densities $p(y|\boldsymbol{x}, \Omega_g)$ are modeled by Gaussian distributions with variance $\sigma^2_{\varepsilon,g}$ around some deterministic function of $\boldsymbol{x}$, say $\mu(\boldsymbol{x}; \boldsymbol{\beta}_g)$, $g = 1, \ldots, G$, so that the relationship between Y and $\boldsymbol{X}$ in the g-th group can be written as $Y = \mu(\boldsymbol{x}, \boldsymbol{\beta}_g) + \varepsilon_g$ where $\varepsilon_g \sim N(0, \sigma^2_{\varepsilon,g})$. Such model will be referred to as Gaussian CWM:

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \phi(y; \mu(\boldsymbol{x}, \boldsymbol{\beta}_g), \sigma^2_{\varepsilon,g})\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\, \pi_g,$$

where $\phi(\cdot)$ denotes the probability density of Gaussian distributions.

For sake of simplicity, we consider the case concerning conditional densities based on linear mappings, that is $\mu(\boldsymbol{x}; \boldsymbol{\beta}_g) = \boldsymbol{b}'_g\boldsymbol{x} + b_{g0}$, with $\boldsymbol{\beta} = (\boldsymbol{b}'_g, b_{g0})'$, $\boldsymbol{b}_g \in \mathbb{R}^d$ and $b_{g0} \in \mathbb{R}$:

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \phi(y; \boldsymbol{b}'_g\boldsymbol{x} + b_{g0}, \sigma^2_{\varepsilon,g})\, \phi_d(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\, \pi_g, \tag{3}$$

which will be referred to as *linear Gaussian CWM*.

3. The likelihood function of Gaussian CWM

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ be a sample of N independent observation pairs drawn from model in (3). Then, the corresponding likelihood function is given by:

$$L_0(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, y_n; \boldsymbol{\theta}) = \prod_{n=1}^{n} \left[\sum_{g=1}^{G} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g) \phi_d(\boldsymbol{x}_n; \boldsymbol{\psi}_g) \pi_g \right],$$

where $\boldsymbol{\chi}_g = (\boldsymbol{\beta}_g, \sigma^2_g)$ and $\boldsymbol{\psi}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Maximization of $L_0(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$ with respect to $\boldsymbol{\theta}$ yields the maximum likelihood estimate of $\boldsymbol{\theta}$.

Let us consider fully categorized data:

$$\{\boldsymbol{w}_n \,:\, n = 1, \ldots, N\} = \{(\boldsymbol{x}_n, y_n, \boldsymbol{z}_n) \,:\, n = 1, \ldots, N\},$$

where $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{ng})'$, with $z_{ng} = 1$ if $(\boldsymbol{x}_n, y_n)$ comes from the g-th population and $z_{ng} = 0$ otherwise. Then, the complete-data likelihood function corresponding to $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)$ can be written in the form:

$$L_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = \prod_{n=1}^{N} \prod_{g=1}^{G} [\phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g)]^{z_{ng}} [\phi_d(\boldsymbol{x}_n; \boldsymbol{\psi}_g)]^{z_{ng}} \pi_g^{z_{ng}}. \tag{4}$$

Taking the logarithm of (4) after some algebra we get:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) &= \ln L_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) \\ &= \sum_{n=1}^{N} \sum_{g=1}^{G} [z_{ng} \ln \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g) + z_{ng} \ln \phi_d(\boldsymbol{x}_n; \boldsymbol{\psi}_g) + z_{ng} \ln \pi_g] \\ &= \mathcal{L}_{1c}(\boldsymbol{\chi}) + \mathcal{L}_{2c}(\boldsymbol{\psi}) + \mathcal{L}_{3c}(\boldsymbol{\pi}), \end{aligned} \tag{5}$$

where

$$\mathcal{L}_{1c}(\boldsymbol{\chi}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[-\ln 2\pi - \ln \sigma^2_{\epsilon,g} - \frac{[y_n - (\boldsymbol{b}'_g \boldsymbol{x}_n + b_{0g})]^2}{\sigma^2_{\epsilon,g}} \right] \tag{6}$$

$$\mathcal{L}_{2c}(\boldsymbol{\psi}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[-p \ln 2\pi - \ln |\boldsymbol{\Sigma}_g| - (\boldsymbol{x}_n - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_g) \right] \tag{7}$$

$$\mathcal{L}_{3c}(\boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} [\ln \pi_g]. \tag{8}$$

Log-likelihood function (5) can be maximized through the EM algorithm in order to obtain the parameter estimates $\boldsymbol{\theta} = \{\boldsymbol{\chi}, \boldsymbol{\psi}, \boldsymbol{\pi}\}$. The *E-step* on the $(k+1)$-th iteration of the EM algorithm requires the calculation of the conditional expectation of the complete-data log-likelihood function $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$ in (5), say $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$, evaluated using the current fit $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$. Since $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$ is linear in the unobservable data z_{ng}, this means calculating the current conditional expectation of Z_{ng} given $\boldsymbol{X}$ and $\boldsymbol{y}$, where Z_{ng} is the random variable correspond-

ing to z_{ng}, that is

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\theta^{(k)}}\{\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})\} \\ &= \sum_{n=1}^{N}\sum_{g=1}^{G} \mathbb{E}_{\theta^{(k)}}\{Z_{ng}|\boldsymbol{x}_n, y_n\}[Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)}) + Q_2(\boldsymbol{\psi}_g; \boldsymbol{\theta}^{(k)}) + \ln \pi_g] \\ &= \sum_{n=1}^{N}\sum_{g=1}^{G} \tau_{ng}^{(k)}[Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)}) + Q_2(\boldsymbol{\psi}_g; \boldsymbol{\theta}^{(k)}) + \ln \pi_g], \end{aligned}$$

where

$$\tau_{ng}^{(k)} = \frac{\pi_g^{(k)}\phi(y_n|\boldsymbol{x}_n; \boldsymbol{\beta}_g^{(k)}, \sigma_g^{2(k)})\phi_d(\boldsymbol{x}_n; \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\Sigma}_g^{(k)})}{\sum_{j=1}^{G}\pi_j^{(k)}\phi(y_n|\boldsymbol{x}_n; \boldsymbol{\beta}_j^{(k)}, \sigma_j^{2(k)})\phi_d(\boldsymbol{x}_n; \boldsymbol{\mu}_j^{(k)}, \boldsymbol{\Sigma}_j^{(k)})}$$

provides the current value of (2) on the k-iteration and

$$\begin{aligned} Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)}) &= \frac{1}{2}\left[-\ln 2\pi - \ln \sigma_{\epsilon,g}^2 - \frac{[y_n - (\boldsymbol{b}_g'\boldsymbol{x}_n - b_{0g})]^2}{\sigma_{\epsilon,g}^2}\right], \\ Q_2(\boldsymbol{\psi}_g; \boldsymbol{\theta}^{(k)}) &= \frac{1}{2}\left[-p\ln 2\pi - \ln|\boldsymbol{\Sigma}_g| - (\boldsymbol{x}_n - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_g)\right]. \end{aligned}$$

The *M-step* on the $(k+1)$-th iteration of the EM algorithm requires the maximization of the conditional expectation of the complete-data log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. The solutions for posterior probabilities $\pi_g^{(k+1)}$ and parameters $(\boldsymbol{\mu}_g^{(k+1)}, \boldsymbol{\Sigma}_g^{(k+1)})$ of local densities $\phi_d(\boldsymbol{x}_n|\boldsymbol{\psi}_g)$, $g = 1, \ldots, G$, exist in closed form (e.g. McLachlan and Peel, 2000), that is:

$$\begin{aligned} \pi_g^{(k+1)} &= \frac{1}{N}\sum_{n=1}^{N}\tau_{ng}^{(k)}, \\ \boldsymbol{\mu}_g^{(k+1)} &= \frac{\sum_{n=1}^{N}\tau_{ng}^{(k)}\boldsymbol{x}_n}{\sum_{n=1}^{N}\tau_{ng}^{(k)}}, \\ \boldsymbol{\Sigma}_g^{(k+1)} &= \frac{\sum_{n=1}^{N}\tau_{ng}^{(k)}(\boldsymbol{x}_n - \boldsymbol{\mu}_g^{(k+1)})(\boldsymbol{x}_n - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{n=1}^{N}\tau_{ng}^{(k)}}. \end{aligned} \tag{9}$$

The updates $\boldsymbol{b}_g^{(k+1)}$, $b_{g0}^{(k+1)}$ and $\sigma_{\epsilon,g}^{2(k+1)}$ for parameters of local densities $\phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g)$,

$g = 1, \ldots, G$, are obtained by solving the equations:

$$\frac{\partial \mathbb{E}_{\theta^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\boldsymbol{x}_n, y_n)\}}{\partial b_{g0}} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)})}{\partial b_{g0}} = 0, \tag{10}$$

$$\frac{\partial \mathbb{E}_{\theta^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\boldsymbol{x}_n, y_n)\}}{\partial \boldsymbol{b}_g'} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{b}_g'} = \mathbf{0}', \tag{11}$$

$$\frac{\partial \mathbb{E}_{\theta^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\boldsymbol{x}_n, y_n)\}}{\partial \sigma_{\epsilon,g}^2} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\theta}^{(k)})}{\partial \sigma_{\epsilon,g}^2} = 0, \tag{12}$$

yielding

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

See Appendix for computational details.

4. Maximum likelihood estimates of Gaussian CWM and relationships with FMR and FMRC

In this section, we analyse the relationships between maximum likelihood estimates of Gaussian CWM and both FMR and FMRC. To begin with, we show in the following that, under suitable hypotheses, maximization of the likelihood function of Gaussian CWM leads to the same parameter estimates of FMR and FMRC. In this sense, FMR and FMRC can be considered as nested models of Gaussian CWM.

4.1. Relationship with FMR

Let us consider the density function of FMR (De Sarbo and Cron, 1988; McLachlan and Peel, 2000; Frühwirth-Schnatter, 2005):

$$f(y|\boldsymbol{x};\boldsymbol{\psi}_\circ) = \sum_{g=1}^{G} p(y|\boldsymbol{x},\Omega_g)\pi_g = \sum_{g=1}^{G} \phi(y;\boldsymbol{b}_g'\boldsymbol{x}+b_{g0},\sigma^2_{\varepsilon,g})\,\pi_g,$$

where $\boldsymbol{\psi}_\circ$ denotes the overall parameters of the model.

The corresponding complete-data log-likelihood function is:

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\psi}_\circ;\boldsymbol{X},\boldsymbol{y}) &= \sum_{n=1}^{N}\sum_{g=1}^{G}\left(z_{ng}\ln\phi(y_n|\boldsymbol{x}_n;\boldsymbol{\chi}_g)+z_{ng}\ln\pi_g\right)\\ &= \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\ln\phi(y_n|\boldsymbol{x}_n;\boldsymbol{\chi}_g)+\sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\ln\pi_g\\ &= \mathcal{L}_{1c}(\boldsymbol{\chi})+\mathcal{L}_{3c}(\boldsymbol{\pi}).\end{aligned} \tag{13}$$

Proposition 1. In model (3), if local densities $\phi_d(\boldsymbol{x};\boldsymbol{\psi}_g)$ have the same parameters $\boldsymbol{\psi}_g=(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)=(\boldsymbol{\mu},\boldsymbol{\Sigma})=\boldsymbol{\psi}$, that is

$$\phi_d(\boldsymbol{x};\boldsymbol{\psi}_g)=\phi_d(\boldsymbol{x};\boldsymbol{\psi}),\qquad g=1,\ldots G, \tag{14}$$

then maximum likelihood estimate of $(\boldsymbol{\chi},\boldsymbol{\pi})$ in (13) coincides with the corresponding estimate in (5).

Proof. In order to prove the proposition, it is sufficient to show that, under the assumption that $\boldsymbol{\psi}_g=(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)=(\boldsymbol{\mu},\boldsymbol{\Sigma})=\boldsymbol{\psi}$, terms $\mathcal{L}_{1c}(\boldsymbol{\chi})$ and $\mathcal{L}_{3c}(\boldsymbol{\pi})$ in (5) do not depend on $(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$, $g=1,\ldots,G$. Indeed, under (14), the complete-data log-likelihood function becomes:

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{y}) &= \ln L_c(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{y})\\ &= \sum_{n=1}^{N}\sum_{g=1}^{G}\left[z_{ng}\ln\phi(y_n|\boldsymbol{x}_n;\boldsymbol{\chi}_g)+z_{ng}\ln\phi_d(\boldsymbol{x}_n;\boldsymbol{\psi})+z_{ng}\ln\pi_g\right]\\ &= \mathcal{L}_{1c}(\boldsymbol{\chi})+\mathcal{L}^*_{2c}(\boldsymbol{\psi})+\mathcal{L}_{3c}(\boldsymbol{\pi}),\end{aligned} \tag{15}$$

where $\mathcal{L}_{2c}(\boldsymbol{\psi})$ in (7) is now replaced by

$$\mathcal{L}^*_{2c}(\boldsymbol{\psi})=\sum_{n=1}^{N}\frac{1}{2}\left[-p\ln 2\pi-\ln|\boldsymbol{\Sigma}|-(\boldsymbol{x}_n-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right],$$

since $\sum_{g=1}^{G} z_{ng} = 1$ for $n = 1, \ldots, N$.

Moreover, in the E-step, the posterior probability $\tau_{ng}^{(k)}$ in (3) becomes:

$$\tau_{ng}^{(k)} = \frac{\pi_g^{(k)} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g^{(k)}) \phi_d(\boldsymbol{x}_n|\boldsymbol{\psi}^{(k)})}{\sum_{j=1}^{G} \pi_j^{(k)} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_j^{(k)}) \phi_d(\boldsymbol{x}_n|\boldsymbol{\psi}^{(k)})} = \frac{\pi_g^{(k)} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g^{(k)})}{\sum_{j=1}^{G} \pi_j^{(k)} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_j^{(k)})},$$

$n = 1, \ldots, N$ and $g = 1, \ldots, G$.

Then, according to (9), term $\mathcal{L}_{3c}(\boldsymbol{\pi})$ does not depend on $\boldsymbol{\psi}_g$. Thus, maximization of (5) can be attained by independently maximizing the three terms $\mathcal{L}_{1c}(\boldsymbol{\chi}), \mathcal{L}_{2c}^{*}(\boldsymbol{\psi})$ and $\mathcal{L}_{3c}(\boldsymbol{\pi})$ and hence, maximization of (13) and (15) in the M-step leads to the same estimates of $(\boldsymbol{\chi}, \boldsymbol{\pi})$. This completes the proof. □

4.2. Relationship with FMRC

Let us consider the density function of FMRC (e.g. Dayton and Macready, 1988):

$$f^{*}(y|\boldsymbol{x}; \boldsymbol{\psi}^{*}) = \sum_{g=1}^{G} \phi(y; \boldsymbol{b}_g'\boldsymbol{x} + b_{g0}, \sigma_{\varepsilon,g}^2)\, p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi}) \,, \tag{16}$$

where the mixing weight $p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi})$ is now a function depending on $\boldsymbol{x}$ through some parameters $\boldsymbol{\xi}$ and $\boldsymbol{\psi}^{*}$ is the augmented set of all parameters of the model.

Probability $p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi})$ is usually modeled by a multinomial logistic distribution with the first component as baseline, that is:

$$p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi}) = \frac{\exp(\boldsymbol{w}_g'\boldsymbol{x} + w_{g0})}{\sum_{j=1}^{G} \exp(\boldsymbol{w}_j'\boldsymbol{x} + w_{j0})} \,. \tag{17}$$

In particular, equation (17) is satisfied if local densities $p(\boldsymbol{x}|\Omega_g)$, $g = 1, \ldots, G$, are assumed to be Gaussian with the same covariance matrices (Anderson, 1972).

The complete-data log-likelihood function corresponding to (16) is:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\psi}_{\circ}; \boldsymbol{X}, \boldsymbol{y}) &= \sum_{n=1}^{N} \sum_{g=1}^{G} \left[z_{ng} \ln \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g) + z_{ng} \ln p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi}) \right] \\ &= \mathcal{L}_{1c}(\boldsymbol{\chi}) + \mathcal{L}_{3c}(\boldsymbol{\xi}) \,. \end{aligned} \tag{18}$$

Proposition 2. In model (3), if local densities $\phi_d(\boldsymbol{x}; \boldsymbol{\psi}_g)$ have the same covariance matrices $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, $g = 1, \ldots, G$, and equal prior probabilities $\pi_g = 1/G$, then maximum likelihood estimate of $(\boldsymbol{\chi}, \boldsymbol{\xi})$ in (18) can be derived from the estimate of $(\boldsymbol{\chi}, \boldsymbol{\psi})$ in (5).

Proof. In order to prove the proposition, it is sufficient to show that, under assumptions $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = 1/G$, $g = 1, \ldots, G$, terms $\mathcal{L}_{1c}(\boldsymbol{\chi})$ and $\mathcal{L}_{3c}(\boldsymbol{\pi})$ in (5) do not depend on $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $g = 1, \ldots, G$. Indeed, we have:

$$L_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = \prod_{n=1}^{N} \prod_{g=1}^{G} \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g)^{z_{ng}} \phi_d(\boldsymbol{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma})^{z_{ng}} \pi^{z_{ng}} \tag{19}$$

and taking the logarithm of (19), after some algebra we get

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) &= \ln L_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) \\ &= \sum_{n=1}^{N} \sum_{g=1}^{G} \left[z_{ng} \ln \phi(y_n|\boldsymbol{x}_n; \boldsymbol{\chi}_g) + z_{ng} \ln \phi_d(\boldsymbol{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}) \right] + \pi \\ &= \mathcal{L}_{1c}(\boldsymbol{\chi}) + \mathcal{L}_{2c}^{**}(\boldsymbol{\psi}) + \pi, \end{aligned} \tag{20}$$

where $\mathcal{L}_{2c}(\boldsymbol{\psi})$ in (7) is now replaced by

$$\mathcal{L}_{2c}^{**}(\boldsymbol{\psi}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[-p \ln 2\pi - \ln |\boldsymbol{\Sigma}| - (\boldsymbol{x}_n - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_g) \right].$$

Once the estimates of $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ have been obtained, quantity $p(\Omega_g|\boldsymbol{x}, \boldsymbol{\xi})$ in (18) can be obtained immediately, that is:

$$p(\Omega_g|\boldsymbol{x}_n, \boldsymbol{\xi}) = \frac{\phi_d(\boldsymbol{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma})\pi}{p(\boldsymbol{x}_n)} = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_g)\right]}{\sum_{j=1}^{G} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j)\right]},$$

which can be written in form (17) for suitable constants $\boldsymbol{w}_g, w_{g0}$, $g = 1, \ldots, G$. This completes the proof. □

5. Conclusions

In this paper, we presented an analysis of the complete-data likelihood function of Gaussian CWM and derived the parameter estimates according to the EM algorithm. Afterwards, theoretical results showed that, under suitable assumptions, both FMR and FMRC are nested models of Gaussian CWM. This implies that CWM is a quite general framework for local statistical modeling.

Appendix

From equation (10), for $b_{g0}^{(k+1)}$ $(g = 1, \ldots, G)$ we obtain:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\psi}^{(k)})}{\partial b_{g0}} = 0$$

yielding

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \left[y_n - \left(\boldsymbol{b}_g^{\prime(k)} \boldsymbol{x}_n + b_{g0}^{(k)} \right) \right] = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} \tau_{ng}^{(k)} (y_n - \boldsymbol{b}_g^{\prime(k)} \boldsymbol{x}_n) = b_{g0}^{(k)} \sum_{n=1}^{N} \tau_{ng}^{(k)}$$

and then we get

$$b_{g0}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \boldsymbol{b}_g^{(k+1)} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.$$

For $\boldsymbol{b}_g^{(k+1)}$ $(g = 1, \ldots, G)$, equation (11) leads to:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{b}_g'} = \mathbf{0}' \tag{21}$$

which implies

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \left[y_n - \left(\boldsymbol{b}_g' \boldsymbol{x}_n + b_{g0}^{(k)} \right) \right] \boldsymbol{x}_n' = \mathbf{0}'$$

yielding

$$\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} = \boldsymbol{b}_g' \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \right)^2 \right)$$

and finally

$$\begin{aligned} \boldsymbol{b}_g^{\prime(k+1)} = & \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \right) \cdot \\ & \cdot \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \boldsymbol{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \right)^2 \right)^{-1}. \end{aligned} \tag{22}$$

Furthermore, equation (12) leads to the current estimate of the variance $\sigma_{\epsilon,g}^{(k)}$ $(g = 1, \ldots, G)$:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\psi}^{(k)})}{\partial \sigma_{\epsilon,g}^2} = 0$$

leading to

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \left\{ -\frac{1}{\sigma_{\epsilon,g}^{2(k)}} + \frac{1}{\sigma_{\epsilon,g}^{4(k)}} \left[y_n - \left(\boldsymbol{b}_g^{\prime(k)} \boldsymbol{x}_n + b_{g0}^{(k)} \right) \right]^2 \right\} = 0$$

and furthermore

$$\sigma_{\epsilon,g}^{2(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \left[y_n - \left(\boldsymbol{b}_g^{(k+1)} \boldsymbol{x}'_n + b_{g0}^{(k+1)} \right) \right]^2}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}. \tag{23}$$

Finally, we remark that in general case the equations (5) and (21) are replaced by

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial Q_1(\boldsymbol{\chi}_g; \boldsymbol{\psi}^{(k)})}{\partial \mu(\boldsymbol{x}, \boldsymbol{\beta}_g)} \frac{\partial \mu(\boldsymbol{x}, \boldsymbol{\beta}_g)}{\partial \boldsymbol{\beta}'_g} = \mathbf{0}'$$

and (23) is replaced by

$$\sigma_{\epsilon,g}^{2(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \left[y_n - \mu(\boldsymbol{x}, \boldsymbol{\beta}_g) \right]^2}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.$$

References

Anderson, J.A., 1972. Separate sample logistic discrimination. Biomtrka. 59, 19-35.

De Sarbo, W.S., Cron, W.L., 1988. A maximum likelihood methodology for clusterwise linear regression. JClass. 5, 248-282.

Dayton, C.M., Macready, G.B., 1988. Concomitant-Variable Latent-Class Models, JASA. 83, 173-178.

Frühwirth-Schnatter, S., 2005. Finite Mixture and Markov Switching Models. Springer, Heidelberg.

Gershenfeld, N., 1997. Non linear inference and Cluster-Weighted Modeling. Annals of the New York Academy of Sciences. 808, 18-24.

Gershenfeld, N., Schöner, B., Metois, E., 1999. Cluster-weighted modelling for time-series analysis. Nature. 397, 329-332.

Gershenfeld, N., 1999. The Nature of Mathematical Modelling. Cambridge University Press, Cambridge.

Ingrassia, S., Minotti, S.C., Vittadini, G., 2011. Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. http://EconPapers.repec.org/RePEc: mis:wpaper:20111001, submitted to JClass, 2nd revision.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Schöner, B., Gershenfeld, N., 2001. Cluster Weighted Modeling: Probabilistic Time Series Prediction, Characterization, and Synthesis, in: Mees, A.I. (Ed.), Nonlinear Dynamics and Statistics. Birkhauser, Boston, 365-385.

Schöner, B., 2000. Probabilistic Characterization and Synthesis of Complex Data Driven Systems. Ph.D. Thesis, MIT.

Wedel, M., 2002. Concomitant variables in finite mixture models. StNeerla. 56(3), 362-375.