TESI DI DOTTORATO

Dipartimento di Economia e Impresa

Statistical algorithms for cluster weighted models

Algoritmi statistici per i cluster weighted models

Giuseppe Incarbone

Tutor: Prof. Salvatore Ingrassia

Coordinatore Dottorato: Prof. Salvatore Greco

Dottorato di Ricerca in "Matematica per le decisioni economiche e finanziarie"

XXIV Ciclo – 2012

Settore Scientifico Disciplinare: SECS-S/01

Università degli Studi di Catania

Contents

Chapter 1

Cluster Weighted Models

1.1 Introduction

Linear systems theory has produced a many results which can be used in practically all engineering and scientific disciplines. Most signal processing, system engineering, control and characterization techniques rely on linear assumptions and apply the results produced by decades of research in this field. However the limitations is the fact that non-linear behaviour of any kind can not be handled.

Cluster Weighted Modeling (CWM) is a modeling tool that allows to characterize systems of arbitrary character. In the original formulation, Cluster Weighted Models (CWM) have been proposed by Gershenfeld (1997) under Gaussian and linear assumptions in the context of media technology to build a digital violin with traditional inputs and realistic sound (Gershenfeld (1997), Gershenfeld *et al.* (1999), Gershenfeld (1998), Schöner and Gershenfeld (2001)).

The use of CWM has also been propsed for evaluating the quality of public sector activities(Minotti and Vittadini (2010), Minotti S.C. (2011)). The framework is based on density estimation around Gaussian kernels which contain simple local models describ-

ing the system behaviour of data subspace. In the extreme case where only one kernel is used the framework collapses to a simple model that is linear in the coefficients. In the opposite extreme it allows one to embed and forecast data that may be non-Gaussian, discontinuous, high dimensional, and chaotic. In between CWM covers a multitude of models, each of which is characterized by linear models with transparent local structures through the embedding of past practice and mature techniques in the general non-linear framework.

The limitations of Artificial Neural Networks (ANNs) have become apparent almost as quickly as their modeling power: networks take long to converge, coefficients are only meaningful in the context of the entire model and failure and success of an architecture are unpredictable beforehand.

More recently a new family of networks has been developed, which interpret data probabilistically and are often represented in graphical networks (Buntine (1996),Heckerman and Wellman (1995), Jordan (1999)). As a meta-class of models, graphical models are conceptually unbounded. They unify existing network architectures, for example classical ANNs in a single theory Neal (1995), provide new insights and extensions to conventional networks and open up new application domains. Graphical models are also referred to as independence networks, since the graphical representation really describes dependence and independence among random variables. They are called Bayesian belief networks since dependencies between variables are expressed in terms of conditional probability functions that have implicit or explicit prior beliefs built into them. They are furthermore named influence diagrams since causal dependences between variables are clearly illustrated. "Influence" is meant probabilistically, which contains deterministic causality as a special case.

Unfortunately graphical models lack a systematic search algorithm that maps a given

problem into a network architecture. Instead, before the networks parameters can be trained on new data, the architecture needs to be redesigned node by node from scratch. Cluster Weighted Modeling is a special case of a probabilistic model that gives up some of the generality of graphical models in favour of ease of use, a minimal number of hyper-parameters and a fast parameter search. It has been designed as an architecture that is as general as reasonably possible, but as specific to a particular application as necessary. As opposed to ANNs it provides transparent local structures and meaningful parameters, it allows one to identify and analyse data subspaces and converges quickly.

1.2 Architecture

Cluster-Weighted Modeling (CWM) is an input-output inference frame-work based on probability density estimation of a joint set of input feature and output target data. It is similar to mixture-of-experts type architectures (Jordan and Jacobs (1994)) and can be interpreted as a flexible and transparent technique to approximate an arbitrary function. Unlike conventional Kernel based techniques, CWM requires only one hyper-parameter to be fixed beforehand, and provides data parameters such as the length scale (band-width) of the local approximation as an output rather than an input of the algorithm (Cleveland and Devlin (1988)).

Let us consider the general framework of CWM.

Let $(\mathbf{X}, Y)$be the pair of random vector $\mathbf{X}$ and random varaible Y defined on Ω with joint probability distribution $p(\mathbf{x}, y)$,where $\mathbf{X}$ is a d-dimensional input vector with values in some space $\mathcal{X}$ and Y is a response variable having values in $\mathcal{Y} \subseteq \mathbb{R}$. Therefore we have:

$$(\mathbf{x}, y) \in \mathbb{R}^{d+1} \tag{1.1}$$

Let us assume that Ω can be partitioned into G disjoint groups, say $\Omega_1, \ldots, \Omega_G$, that is:

$$\Omega = \Omega_1 \cup \cdots \cup \Omega_G \tag{1.2}$$

Given the joint density $p(\mathbf{x}, y)$ we can expand it in terms of explanatory clusters containing three terms:

- a weight $p(\Omega_g)$
- a domain of influence in the input space $p(\mathbf{x}|\Omega_g)$
- a dependence in the output space $p(y|\mathbf{x}, \Omega_g)$

More specifically CWM decomposes the joint probability as follows:

$$\begin{aligned} p(\mathbf{x}, y; \boldsymbol{\theta}) &= \sum_{m=1}^{M} p(\mathbf{x}, y, \Omega_g) \\ &= \sum_{m=1}^{M} p(\mathbf{x}, y|\Omega_g)\pi_g \\ &= \sum_{g=1}^{G} p(y|\mathbf{x}, \Omega_g)p(\mathbf{x}|\Omega_g)\pi_g \end{aligned} \tag{1.3}$$

where $p(y|\mathbf{x}, \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\mathbf{x}$ and Ω_g, $\pi_g = p(\Omega_g)$ is the mixing weight of Ω_g, ($\pi_g > 0$ and $\sum_{n=1}^{G} \pi_g = 1$), $g = 1, \ldots, G$, and $\boldsymbol{\theta}$ denotes the set of all parameters of the model. Hence, the joint density of $(\mathbf{X}, Y)$ can be viewed as a mixture of local models $p(y|\mathbf{x}, \Omega_g)$ weighted (in a broader sense) on both local densities $p(\mathbf{x}|\Omega_g)$ and mixing weights π_g. In the spirit of Titterington *et al.* (1985), we can distinguish three types of applicaton for CWM in (1.3):

1. *Direct application of type A*. We assume that each group Ω_g is characterized by an input-output relation that can be written as

$$Y|\boldsymbol{x} = \mu(\boldsymbol{x}; \beta_g) + \epsilon_g \tag{1.4}$$

where ϵ_g is a random variable with zero mean and finite variance $\sigma_{\epsilon,g}$, and β_g denotes the set of parameters of the $\mu(\cdot)$ function, $g = 1, \ldots, G$.

2. *Direct application of type B*. We assume that a random vector $\mathbf{Z}$ is defined on $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$ with values in $\mathbb{R}^{d+1}$ belongs to one of these groups. Further, vector $\mathbf{z}$ is partitioned as $\mathbf{z} = (\mathbf{x}', y)'$ and we assume that within each group we write:

$$p(\mathbf{z}; \Omega_g) = p((\mathbf{x}', y)'; \Omega_g) = p(y|\mathbf{x}, \Omega_g)p(\mathbf{x}|\Omega_g). \tag{1.5}$$

   In other words, CWM in (1.3) is another form of density of FMD given by:

$$p(\mathbf{z}; \boldsymbol{\theta}) = \sum_{g=1}^{G} p(\mathbf{z}|\Omega_g)\pi_g = \sum_{g=1}^{G} p(y|\mathbf{x}, \Omega_g)p(\mathbf{x}|\Omega_g)\pi_g. \tag{1.6}$$

3. *Indirect application*. In this case, CWM in (1.3) is simply used as a mathematical tool for density estimation.

In this thesis we will concentrate on direct applications that essentially have classification purposes. In this case, posterior probability $p(\Omega_g|\mathbf{x}, y)$ of unit $(\mathbf{x}, y)$ belonging to the g-th group ($g = 1, \ldots, G$) is given by;

$$p(\Omega_g|\mathbf{x}, y) = \frac{p(\mathbf{x}, y, \Omega_g)}{p(\mathbf{x}, y)} = \frac{p(y|\mathbf{x}, \Omega_g)p(\mathbf{x}|\Omega_g)\pi_g}{\sum_{j=1}^{G} p(y|\mathbf{x}, \Omega_j)p(\mathbf{x}|\Omega_j)\pi_j}, \quad g = 1, \ldots, G \tag{1.7}$$

that is, the classification of each unit depends on both marginal and conditional densities. Because $p(\mathbf{x}|\Omega_g)\pi_g = p(\Omega_g|\mathbf{x})p(\mathbf{x})$, from (1.7) we get:

$$p(\Omega_g|\mathbf{x}, y) = \frac{p(y|\mathbf{x}, \Omega_g)p(\Omega_g|\mathbf{x})p(\mathbf{x})}{\sum_{j=1}^{G} p(y|\mathbf{x}, \Omega_j)p(\Omega_j|\mathbf{x})p(\mathbf{x})} = \frac{p(y|\mathbf{x}, \Omega_g)p(\Omega_g|\mathbf{x})}{\sum_{j=1}^{G} p(y|\mathbf{x}, \Omega_j)p(\Omega_j|\mathbf{x})} \tag{1.8}$$

with

$$p(\Omega_g|\mathbf{x}) = \frac{p(\mathbf{x}|\Omega_g)\pi_g}{\sum_{j=1}^{G} p(\mathbf{x}|\Omega_j)\pi_j} = \frac{p(\mathbf{x}|\Omega_g)\pi_g}{p(\mathbf{x})} \tag{1.9}$$

1.3 Gaussian CWM

In the first approach of CWM, both marginal and conditional densities are assumed to be Gaussian, with $\mathbf{X}|\Omega_g \sim N_d(\mu_g, \Sigma_g)$ and $Y|\mathbf{x}, \Omega_g \sim N(\mu(\mathbf{x}, \beta_g), \sigma_{\epsilon,g})$, so that we shall write:

$$\begin{aligned} p(\mathbf{x}|\Omega_g) &= \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \\ &= \frac{|\Sigma_g^{-1}|^{1/2}}{(2\pi)^{d/2}} e^{-(\mathbf{x}-\boldsymbol{\mu}_g)'\dot{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)/2} \end{aligned} \tag{1.10}$$

and

$$\begin{aligned} p(y|\mathbf{x}, \Omega_g) &= \phi(y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma^2_{\epsilon,g}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{\epsilon,g}}} e^{-[y-\mu(\mathbf{x};\boldsymbol{\beta}_g)]^2/2\sigma^2_{\epsilon,g}} \end{aligned} \tag{1.11}$$

with $g = 1, \ldots, G$.

Let us observe that in (1.11) the mean value of the Gaussian output is replaced by the function $\mu(\boldsymbol{x}, \boldsymbol{\beta}_g)$ with unknown parameters $\boldsymbol{\beta}_g$.

With this assumption, the conditional forecast $\langle y|\mathbf{x}\rangle$ will be:

$$\begin{aligned} \langle y|\mathbf{x}\rangle &= \int y p(y|\mathbf{x}) dy \\ &= \int y \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} dy \\ &= \frac{\sum_{j=1}^G \int y p(y|\mathbf{x}, \Omega_j) dy \; p(\mathbf{x}|\Omega_j)\pi_j}{\sum_{j=1}^G p(\mathbf{x}|\Omega_j)\pi_j} \\ &= \frac{\sum_{j=1}^G \mu(\mathbf{x}, \boldsymbol{\beta}_j) p(\mathbf{x}|\Omega_j)\pi_j}{\sum_{j=1}^G p(\mathbf{x}|\Omega_j)\pi_j}. \end{aligned} \tag{1.12}$$

Let us observe that the predicted y is a superposition of all the local functionals, where the weight of each contribution depends on the posterior probability that an input point was generated by a particular cluster. The denominator assures that the sum of the weights of all contributions equals unity.

Similarly the conditional error in terms of the expected covariance of y given $\boldsymbol{x}$ will be:

$$\begin{aligned}\langle \sigma_y^2 | \mathbf{x} \rangle &= \int (y - \langle y|\mathbf{x}\rangle)^2 p(y|\mathbf{x}) dy \\ &= \int (y^2 - \langle y|\mathbf{x}\rangle^2) p(y|\mathbf{x}) dy \\ &= \frac{\sum_{j=1}^G [\sigma_{\epsilon,j}^2 + \mu(\mathbf{x}, \boldsymbol{\beta}_j)^2] p(\mathbf{x}|\Omega_j)\pi_j}{\sum_{j=1}^G p(\mathbf{x}|\Omega_j)\pi_j} - \langle y|\mathbf{x}\rangle^2.\end{aligned} \tag{1.13}$$

Finally the posterior probability in (1.7) specializes as:

$$\begin{aligned}p(\Omega_g|\mathbf{x}, y) &= \frac{p(\mathbf{x}, y, \Omega_g)}{p(\mathbf{x}, y)} \\ &= \frac{p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}|\Omega_g)\pi_g}{\sum_{j=1}^G p(\mathbf{x}, y, \Omega_j)} \\ &= \frac{p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}|\Omega_g)\pi_g}{\sum_{j=1}^G p(\mathbf{x}|y, \Omega_j) p(\mathbf{x}|\Omega_j)\pi_j} \\ &= \frac{\phi(y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_{\epsilon,g}^2)\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g}{\sum_{j=1}^G \phi(y; \mu(\mathbf{x}; \boldsymbol{\beta}_j), \sigma_{\epsilon,j}^2)\phi_d(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} \quad g = 1, \ldots, G.\end{aligned} \tag{1.14}$$

There are two parameters to be determined beforehand: the number of clusters G and the form of the local models μ which together control the model resources and hence under versus over-fitting. We trade off the complexity of the local models against the complexity of the global architecture, which is nicely illustrated in the case of a local polynomial expansion: if we use locally constant models together with a large number of clusters, the predictive power is determined by the number of Gaussian kernels. If, alternatively, we use a high-order polynomial model and a single kernel, the model reduces to a global polynomial model.

The choice of local models depends on the application. In general μ expresses prior beliefs about the nature of the data or insights in the mechanics of a system an thus functions as a regularizer of the model. Machine learning architectures and estimation algorithms typically depend on global regularizers that handle prior beliefs about what is a good model. This is problematic since global statements may not apply locally. For

example, the maximum entropy principle is good at handling discontinuities, but has no notion of local smoothness, whereas integrated curvature is good in enforcing local smoothness but rounds out discontinuities. In our approach the model is constrained only by the local architecture which may enforce local smoothness but at the same time allows for discontinuities where needed.

1.4 Linear Gaussian CWM and relationships with Traditional Mixture Models

In this section and in the following one we will assume that the conditional densities are based on linear mappings, so that $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = \boldsymbol{b}_{\boldsymbol{g}}'\mathbf{x} + b_{g0}$, for some $\boldsymbol{\beta}_g = (\boldsymbol{b}_{\boldsymbol{g}}', b_{g0})'$, with $\mathbf{b} \in \mathbb{R}^d$ and $b_{g0} \in \mathbb{R}$. Thus we get:

$$p(\mathbf{x}, y, \boldsymbol{\theta}) = \sum_{g=1}^{G} \phi(y; \boldsymbol{b}_{\boldsymbol{g}}'\mathbf{x} + b_{g,0}, \sigma_{\epsilon,g}^2)\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{\boldsymbol{g}})\pi_g \tag{1.15}$$

with $\phi(\cdot)$ denoting the probability density of Gaussian distributions. The approach in (1.15) will be referred to as *linear Gaussian CWM*.

We will now consider the relationships between *linear Gaussian CWM* and traditional Gaussian-based mixture models, considering both probability density functions and posterior probabilities (Ingrassia *et al.* (2012b)). In particular we will prove that, under suitable assumptions, *linear Gaussian CWM* in (1.15) leads to the same posterior probability of such mixture models. In this sense we say that CWM contains other gaussian mixture models. In particular, *linear Gaussian CWM* leads to the same family of probability distributions generated by FMG.

1.4.1 Finite Mixtures of Gaussian Distributions

Let $\mathbf{Z}$ be a random vector defined on $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$ with joint probability distribution $p(\mathbf{z})$, where $\mathbf{Z}$ assumes values in some space $\mathcal{Z} \subset \mathbb{R}^{d+1}$. Assume that the density $p(\mathbf{z})$ of $\mathbf{Z}$ has the form of a mixture of Gaussian distribution (FMG), i.e.

$$p(\mathbf{z}) = \sum_{g=1}^{G} p(\mathbf{z}|\Omega_g)\pi_g \tag{1.16}$$

where $p(\mathbf{z}|\Omega_g)$ is the probability density of $\mathbf{Z}|\Omega_g$ and $\pi_g = p(\Omega_g)$ is the mixing weight of group Ω_g, $g = 1, \ldots, G$. Finally, denote with $\boldsymbol{\mu}_g^{(\mathbf{z})}$ and $\boldsymbol{\Sigma}_g^{(\mathbf{z})}$ the mean vector and the covariance matrix of $\mathbf{Z}|\Omega_g$, respectively. Now let us set $\mathbf{Z} = (\mathbf{X}', Y)$, where $\mathbf{X}$ is a random vector with values in $\mathbb{R}^d$ and Y is a random variable. Thus, we can write

$$\boldsymbol{\mu}_g^{(\mathbf{z})} = \begin{pmatrix} \boldsymbol{\mu}_g^{(\mathbf{x})} \\ \mu_g^{(y)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_g^{(\mathbf{z})} = \begin{pmatrix} \boldsymbol{\Sigma}_g^{(\mathbf{xx})} & \boldsymbol{\Sigma}_g^{(\mathbf{x}y)} \\ \boldsymbol{\Sigma}_g^{(y\mathbf{x})} & \sigma_g^{2(y)} \end{pmatrix} \tag{1.17}$$

Further, the posterior probability in the g-group is given by:

$$p(\Omega_g|\mathbf{z}) = \frac{p(\mathbf{z}|\Omega_g)\pi_g}{\sum_{j=1}^{G} p(\mathbf{z}|\Omega_j)\pi_j} \qquad g = 1, \ldots, G. \tag{1.18}$$

Proposition 1. *Let $\mathbf{Z}$ be a random vector defined on $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$ with values in $\mathbb{R}^{d+1}$, and assume that $\mathbf{Z}|\Omega_g \sim N_{d+1}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)(g = 1, \ldots, G)$. In particular, the density $p(\mathbf{z})$ of $\mathbf{Z}$ is a FMG:*

$$p(\mathbf{z}) = \sum_{g=1}^{G} \phi_{d+1}(\mathbf{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g. \tag{1.19}$$

Then $p(\mathbf{z})$ can be written similar to (1.15), that is as a linear Gaussian CWM.

Proof. Let us set $\mathbf{Z} = (\mathbf{X}', Y)'$, where $\mathbf{X}$ is a d-dimensianl random vector and Y is a random variable. According to well-known results of multivariate statistics (eg. Kent

et al. (1979)), from (1.19) we get:

$$
\begin{aligned}
p(\mathbf{z}) &= \sum_{g=1}^{G} \phi_{d+1}(\mathbf{z}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g = \sum_{g=1}^{G} \phi_{d+1}((\mathbf{x}', y)'; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g \\
&= \sum_{g=1}^{G} \phi_d(\mathbf{x}; \boldsymbol{\mu}_g^{(\mathbf{x})}, \boldsymbol{\Sigma}_g^{(\mathbf{xx})})\phi(y; \mu_g^{(y|\mathbf{x})}, \sigma_g^{2(y|\mathbf{x})})\pi_g,
\end{aligned}
\tag{1.20}
$$

where $\mu_g^{(y|\mathbf{x})} = \mu_g^{(y)} + \boldsymbol{\Sigma}_g^{(y\mathbf{x})}\boldsymbol{\Sigma}_g^{(\mathbf{xx})^{-1}}(\mathbf{x} - \boldsymbol{\mu}_g^{(\mathbf{x})})$ and $\sigma_g^{2(y|\mathbf{x})} = \boldsymbol{\Sigma}_g^{(yy)}$.
If we set $\mathbf{b}_g = \Sigma_g^{(y\mathbf{x})}\Sigma_g^{(\mathbf{xx})^{-1}}$, $b_{g0} = \mu_g^{(y)} - \Sigma_g^{(y\mathbf{x})}\Sigma_g^{(\mathbf{xx})^{-1}}\boldsymbol{\mu}_g^{(\mathbf{x})}$ and $\sigma_{\epsilon,g}^2 = \sigma_g^{2(y|\mathbf{x})}$, the (1.19) can be written in the form of (1.15). □

Using similar arguments, FMG can be shown to lead to the same distribution of posterior probabilities and, thus, CWM contains FMG.
We remark that the equivalence between FMG and CWM holds only for linear mappings $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = \mathbf{b}_g'\mathbf{x} + b_{g0} \quad (g = 1, \ldots, G)$, while, more generally, Gaussian CWM

$$
p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \phi(y; \mu(\mathbf{x}; \boldsymbol{\beta}_g), \sigma_{\epsilon,g}^2)\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g \tag{1.21}
$$

includes a quite wide family of models.

1.4.2 Finite Mixtures of Regression Models

Let us consider Mixtures of Regression Models (FMR) (DeSarbo and Cron (1988), McLachlan and Peel (2000), Frühwirth-Schnatter (2006)):

$$
f(y|\mathbf{x}; \psi) = \sum_{g=1}^{G} \phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2)\pi_g, \tag{1.22}
$$

where vector $\boldsymbol{\psi}$ denotes the overall parameters of the model. Posterior probability $p(\Omega_g|\mathbf{x}, y)$ of the g-th group $(g = 1, \ldots, G)$ for FMR is:

$$
\begin{aligned}
p(\Omega_g|\mathbf{x}, y) &= \frac{f(y|\mathbf{x}; \boldsymbol{\psi}, \Omega_g)}{f(y|\mathbf{x}; \boldsymbol{\psi})} \\
&= \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2)\pi_g}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma_{\epsilon,j}^2)\pi_j}
\end{aligned}
\tag{1.23}
$$

that is the classification of each observation depends on the local model and the mixing weight. We have the following result:

Proposition 2. *Let us consider linear Gaussian CWM in (1.15), with* $\boldsymbol{X}|\Omega_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ *for* $g = 1, \ldots, G$*. If the probability density of* $\boldsymbol{X}|\Omega_g$ *does not depend on group* g*, i.e.,* $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *for every* $g = 1, \ldots, G$*, then it follows:*

$$p(\boldsymbol{x}, y, \boldsymbol{\theta}) = \phi_d(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) f(y|\boldsymbol{x}; \boldsymbol{\psi}). \tag{1.24}$$

where $f(y|\boldsymbol{x}; \boldsymbol{\psi})$ *is the FMR model in (1.22).*

Proof. Assume that $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), g = 1, \ldots, G$. Then (1.15) yields:

$$\begin{aligned} p(\boldsymbol{x}, y, \boldsymbol{\theta}) &= \sum_{g=1}^{G} \phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}) \phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \pi_g \\ &= \phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sum_{g=1}^{G} \phi(y; \boldsymbol{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}) \pi_g \\ &= \phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) f(y|\mathbf{x}; \boldsymbol{\psi}), \end{aligned} \tag{1.25}$$

where $f(y|\mathbf{x}; \boldsymbol{\psi})$ is the FMR model in (1.22)

□

The second result of this section shows that, under the same hypothesis, CWM contains FMR.

Corollary 3. *If the probability density of* $\boldsymbol{X}|\Omega_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ *in (1.15) does not depend on the* g*-th group, i.e.,* $\phi_d(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \phi_d(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *for every* $g = 1, \ldots, G$*, then the posterior probability in (1.14) coincides with (1.23)*

Proof. Assume that $\phi_d(\mathbf{x}; \mu_g \Sigma_g) = \phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), g = 1, \ldots, G$. Thus from (1.14) we

get:

$$
\begin{aligned}
p(\Omega_g|\mathbf{x}, y) &= \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g})\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\pi_g}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j})\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\pi_j} \\
&= \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g})\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\pi_g}{\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j})\pi_j} \\
&= \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g})\pi_g}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j})\pi_j}
\end{aligned}
\tag{1.26}
$$

for $g = 1, \ldots, G$ which coincides with (1.23) □

1.4.3 Finite Mixtures of Regression Models with Concomitant Variables

Mixture of Regression Models with Concomitant Variables(FMRC) (Dayton and Macready (1988), Wedel and DeSarbo (2002)) are extension of FMR:

$$
f^*(y|\mathbf{x}; \boldsymbol{\psi}^*) = \sum_{g=1}^{G} \phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}) p(\Omega_g|\mathbf{x}, \boldsymbol{\xi}), \tag{1.27}
$$

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

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

$$
p(\Omega_g|\mathbf{x}, \boldsymbol{\xi}) = \frac{\exp(\mathbf{w}_g'\mathbf{x} + w_{g0})}{\sum_{j=1}^{G} \exp(\mathbf{w}_j'\mathbf{x} + w_{j0})} \tag{1.28}
$$

Equation (1.28) is satisfied if local densities $p(\mathbf{x}|\Omega_g), \quad g = 1, \ldots, G$ are assumed to be multivariate Gaussian with the same covariance matrices (Anderson (1972)). Posterior probability $p(\Omega_g|\mathbf{x}, y)$ of the g-th group ($g = 1, \ldots, G$) for FMRC is:

$$
p(\Omega_g|\mathbf{x}, y) = \frac{f^*(y|\mathbf{x}; \boldsymbol{\psi}^*, \Omega_g)}{f^*(y|\mathbf{x}; \boldsymbol{\psi}^*)} = \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}) p(\Omega_g|\mathbf{x}, \boldsymbol{\xi})}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j}) p(\Omega_j|\mathbf{x}, \boldsymbol{\xi})} \tag{1.29}
$$

Under suitable assumptions, *linear Gaussian CWM* leads to the same estimates of $\mathbf{b}_g, b_{g0} (g = 1, \ldots, G)$ in (1.27).

Proposition 4. *Let us consider linear Gaussian CWM in (1.15), with* $\boldsymbol{X}|\Omega_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)(g = 1, \ldots, G)$. *If* $\boldsymbol{\Sigma}_g = \Sigma$ *and* $\pi_g = \pi = 1/G$ *for every* $g = 1, \ldots, G$, *then it follows that*

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = p(\boldsymbol{x}) f^*(y|\boldsymbol{x}; \boldsymbol{\psi}^*), \tag{1.30}$$

where $f^*(y|\boldsymbol{x}; \boldsymbol{\psi}^*)$ *is the FMRC model in (1.27) based on the multinomial logistic in (1.28) and* $p(\boldsymbol{x}) = \sum_{g=1}^{G} p(\boldsymbol{x}|\Omega_g)\pi_g$

Proof. Assume $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = \pi = 1/G$ for every $g = 1, \ldots, G$; thus, the density in (1.15) yields:

$$\begin{aligned}
&p(\mathbf{x}, y; \theta) \\
&= \sum_{g=1}^{G} \phi(\mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2)\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \Sigma)\pi \\
&= p(\mathbf{x}) \sum_{g=1}^{G} \phi(\mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2) \frac{\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \Sigma)\pi}{p(\mathbf{x})} \\
&= p(\mathbf{x}) \sum_{g=1}^{G} \phi(\mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2) \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)\right]}{\sum_{j=1}^{G} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\right]}
\end{aligned} \tag{1.31}$$

where

$$\begin{aligned}
&\frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)\right]}{\sum_{j=1}^{G} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\right]} \\
&= \frac{1}{1 + \sum_{j \neq g} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)\right]} \\
&= \frac{1}{1 + \sum_{j \neq g} \exp\left[(\boldsymbol{\mu}_j - \boldsymbol{\mu}_g)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_g)'\Sigma^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_g)\right]}
\end{aligned} \tag{1.32}$$

and we recognize that (1.32) can be written in form (1.28) for suitable constants $\mathbf{w}_g, w_{g0} \quad (g = 1, \ldots, G)$. □

Based on similar arguments, we can immediately prove that, under the same hypotheses, CWM contains FMRC.

Corollary 5. *Let us consider the linear Gaussian CWM in 1.15. If* $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ *and* $\pi_g = \pi = 1/G$ *for every* $g = 1, \ldots, G$*, then the posterior probability in (1.14) coincides with (1.29).*

Proof. First, based on (1.28), let us rewrite (1.29) as

$$p(\Omega_g|\mathbf{x}, y) = \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g})\exp(\mathbf{w}_g'\mathbf{x} + w_{g0})}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j})\exp(\mathbf{w}_j'\mathbf{x} + w_{j0})} \tag{1.33}$$

Assume $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = \pi = 1/G$ for every $g = 1, \ldots, G$. Thus (1.14) reduces to

$$p(\Omega_g|\mathbf{x}, y) = \frac{\phi(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g})\phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma})}{\sum_{j=1}^{G} \phi(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j})\phi_d(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})} \tag{1.34}$$

and after some algebra we find a quantity similar to (1.33). □

As for the relation between FMRC and *linear Gaussian CWM*, consider that joint density $p(\mathbf{x}, \Omega_g)$ can be written in either form:

$$p(\mathbf{x}, \Omega_g) = p(\mathbf{x}|\Omega_g)p(\Omega_g) \quad \text{or} \quad p(\mathbf{x}, \Omega_g) = p(\Omega_g|\mathbf{x})p(\mathbf{x}), \tag{1.35}$$

where quantity $p(\mathbf{x}|\Omega_g)$ is involved in CWM (left-hand side), while FMRC contains conditional probability $p(\Omega_g|\mathbf{x})$ (right-hand side). In other words, CWM is a Ω_g-to-$\mathbf{x}$ model, while FMRC is a $\mathbf{x}$-to-Ω_g model. According to Jordan *et al.* (1995), they are called the *generative direction* model and the *diagnostic direction* model, respectively, in the framework of neural networks.

The results of this section are provided in Table 1.1, which summarizes the relationships between *linear Gaussian CWM* and traditional Gaussian mixture models.

model	$p(\mathbf{x}\|\Omega_g)$	$p(y\|\mathbf{x},\Omega_g)$	parameterisation of π_g	assumptions
FMG	Gaussian	Gaussian	none	
FMR	none	Gaussian	none	$(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, g=1,...,G
FMRC	none	Gaussian	logistic	$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = \pi$, g=1,...,G

Table 1.1: Relationships between *linear Gaussian* CWM and traditional Gaussian mixtures.

Finally, we remark that if conditional distributions

$$p(y|\mathbf{x},\Omega_g) = \phi(y;\mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\varepsilon,g}) \quad (g = 1,\ldots,G) \tag{1.36}$$

do not depend on group g, that is

$$\phi(y;\mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\varepsilon,g}) = \phi(y;\mathbf{b}'\mathbf{x} + b_0, \sigma^2_{\varepsilon}) \quad (g = 1,\ldots,G), \tag{1.37}$$

then (1.15) specializes as:

$$\begin{aligned} p(\mathbf{x}, y, \theta) &= \sum_{g=1}^{G} \phi(y;\mathbf{b}'\mathbf{x} + b_0, \sigma^2_{\varepsilon}) \phi_d(\mathbf{x}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g \\ &= \phi(y;\mathbf{b}'\mathbf{x} + b_0, \sigma^2_{\varepsilon}) \sum_{g=1}^{G} \phi_d(\mathbf{x}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g, \end{aligned} \tag{1.38}$$

and this implies, from (1.22) and (1.27), that FMR and FMRC are reduced to a single straight line:

$$f(y|\mathbf{x};\psi) = f^*(y|\mathbf{x},\psi^*) = \phi(y;\mathbf{b}'\mathbf{x} + b_0), \tag{1.39}$$

because $\sum_{g=1}^{G} \pi_g = \sum_{g=1}^{G} p(\Omega_g|\mathbf{x},\boldsymbol{\xi}) = 1$.

1.5 Decision surfaces of linear Gaussian CWM

The potential of CWM as a general and flexible framework for classification purposes can also be illustrated from a geometrical point of view, by considering the decision

surfaces that separate the groups. In the following we will discuss the binary case and will prove that these decision surfaces belong to the family of quadrics.

In the specific case of two groups, a decision surface is the set of $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$ such that $p(\Omega_0|\mathbf{x}, y) = p(\Omega_1|\mathbf{x}, y) = 0.5$. Given that $p(\mathbf{x}|\Omega_g)\pi_g = p(\Omega_g|\mathbf{x})p(\mathbf{x})$, we can rewrite $p(\Omega_1|\mathbf{x}, y)$ as:

$$\begin{aligned} p(\Omega_1|\mathbf{x}, y) &= \frac{p(y|\mathbf{x}, \Omega_1)p(\Omega_1|\mathbf{x})}{p(y|\mathbf{x}, \Omega_0)p(\Omega_0|\mathbf{x}) + p(y|\mathbf{x}, \Omega_1)p(\Omega_1|\mathbf{x})} \\ &= \frac{1}{1 + \dfrac{p(y|\mathbf{x}, \Omega_0)p(\Omega_0|\mathbf{x})}{p(y|\mathbf{x}, \Omega_1)p(\Omega_1|\mathbf{x})}} \\ &= \frac{1}{1 + \exp\left\{-\ln\dfrac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} - \ln\dfrac{p(\Omega_1|\mathbf{x})}{p(\Omega_0|\mathbf{x})}\right\}}. \end{aligned} \tag{1.40}$$

Thus it results that $p(\Omega_1|\mathbf{x}, y) = 0.5$ when

$$\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} + \ln\frac{p(\Omega_1|\mathbf{x})}{p(\Omega_0|\mathbf{x})} = 0, \tag{1.41}$$

which may be rewritten as:

$$\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} + \ln\frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_0)} + \ln\frac{\pi_1}{\pi_0} = 0. \tag{1.42}$$

In the *linear Gaussian-CWM*, the first and the second term in (1.42) are, respectively:

$$\begin{aligned} \ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} &= \ln\frac{\sqrt{2\pi\sigma^2_{\epsilon,0}}}{\sqrt{2\pi\sigma^2_{\epsilon,1}}} + \frac{(y - \boldsymbol{b_0'}\mathbf{x} - b_{00})^2}{2\sigma^2_{\epsilon,0}} - \frac{(y - \boldsymbol{b_1'}\mathbf{x} - b_{10})^2}{2\sigma^2_{\epsilon,1}} \\ \ln\frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_0)} &= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} \\ &\quad + \frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_0)'\Sigma_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) - (\mathbf{x} - \boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)]. \end{aligned} \tag{1.43}$$

Then, equation (1.42) is satisfied for $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$ such that:

$$\begin{aligned} &\ln\frac{\sigma_{\epsilon,0}}{\sigma_{\epsilon,1}} + \frac{(y - \boldsymbol{b_0'}\mathbf{x} - b_{00})^2}{2\sigma^2_{\epsilon,0}} - \frac{(y - \boldsymbol{b_1'}\mathbf{x} - b_{10})^2}{2\sigma^2_{\epsilon,1}} + \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma_0}|}{|\boldsymbol{\Sigma_1}|} + \\ &\frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_0)'\Sigma_0^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) - (\mathbf{x} - \boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] + \ln\frac{\pi_1}{\pi_0} = 0 \end{aligned} \tag{1.44}$$

which defines quadratic surfaces, i.e., *quadrics*. Examples of quadrics are spheres, circular cylinders, and circular cones. (Se possibile inserire le figure) In the homoschedastic case $\boldsymbol{\Sigma_0} = \boldsymbol{\Sigma_1} = \boldsymbol{\Sigma}$, it is well known that:

$$\begin{aligned} \ln\frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_0)} &= \frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) - (\mathbf{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)] \\ &= \mathbf{w'x} + w_0 \end{aligned} \tag{1.45}$$

where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) \quad \text{and} \quad w_0 = \frac{1}{2}(\boldsymbol{\mu_0} + \boldsymbol{\mu_1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_0} - \boldsymbol{\mu_1}). \tag{1.46}$$

In this case, according to (1.45) equation (1.42) yields:

$$\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} + \mathbf{w}'\mathbf{x} + w_0 + \ln\frac{\pi_1}{\pi_0} = 0 \tag{1.47}$$

(see Figure 1.1).

1.6 Parameter Estimation of CWM via the EM algorithm - Gaussian case

Given a sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ of N independent observation pairs, the cluster-weighted likelihood function is

$$L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) = \prod_{n=1}^{N} p(\mathbf{x}_n, y_n; \boldsymbol{\psi}) = \prod_{n=1}^{n}\left[\sum_{g=1}^{G}\phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g)\phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g)\pi_g\right]. \tag{1.48}$$

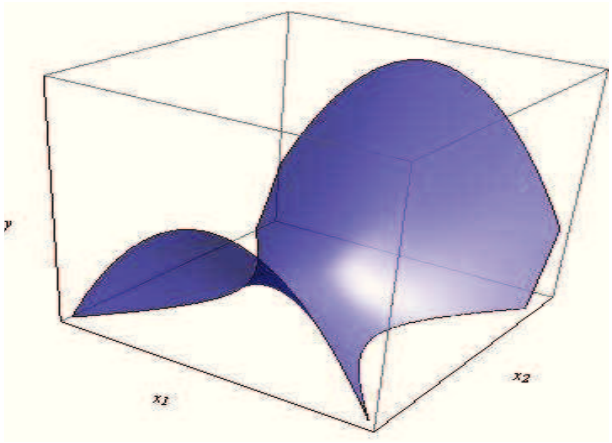

Figure 1.1: Examples of decision surfaces for *linear Gaussian* CWM (homoscedastic case).

Maximization of $L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y})$ with respect to $\boldsymbol{\psi}$, for given data $(\underset{\sim}{\mathbf{X}}, \mathbf{Y})$, yields the maximum likelihood estimate of $\boldsymbol{\psi}$. Equivalently the quantity maximized is the log-likelihood $\mathcal{L}_0 = \ln L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y})$.

If we consider fully categorized data:

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

then the likelihood corresponding to $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ can be written in the form

$$L_c(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) = \prod_{n=1}^{N} \prod_{g=1}^{G} \phi(y_n | \mathbf{x}_n; \boldsymbol{\beta}_g)^{z_{ng}} \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g)^{z_{ng}} \pi_g^{z_{ng}}, \tag{1.49}$$

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

Consider the logarithm

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) &= \ln \prod_{n=1}^{N} \prod_{g=1}^{G} \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g)^{z_{ng}} \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g)^{z_{ng}} \pi_g^{z_{ng}} = \\
&= \sum_{n=1}^{N} \sum_{g=1}^{G} [\ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g)^{z_{ng}} + \ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g)^{z_{ng}} + \ln \pi_g^{z_{ng}}] = \\
&= \sum_{n=1}^{N} \sum_{g=1}^{G} [z_{ng} \ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g) + z_{ng} \ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g) + z_{ng} \ln \pi_g] = \\
&= \sum_{n=1}^{N} [\mathbf{z}_n' \ln \boldsymbol{\phi}(y_n|\mathbf{x}_n; \mathbf{B}) + \mathbf{z}_n' \ln \boldsymbol{\phi}_d(\mathbf{x}_n; \boldsymbol{\Theta}) + \mathbf{z}_n' \ln \boldsymbol{\pi}] = \\
&= \sum_{n=1}^{N} \mathbf{z}_n' \mathbf{W}_n(\mathbf{B}) + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{U}_n(\boldsymbol{\Theta}) + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{V}(\boldsymbol{\pi}), \qquad (1.50)
\end{aligned}$$

where $\mathbf{W}_n(\mathbf{B})$ is a G-component vector having the g-th component $\ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g)$, $\mathbf{U}_n(\boldsymbol{\Theta})$ is a G-component vector having the g-th component $\ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g)$ and $\mathbf{V}(\boldsymbol{\pi})$ is a G-component vector having the g-th component $\ln \pi_g$.

The form of the cluster-weighted likelihood function in (1.48) corresponds to the marginal density of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ obtained summing (1.49) over $\mathbf{z}_1, \ldots, \mathbf{z}_N$. This emphasizes the interpretation of cluster-weighted data as incomplete data with the indicator vectors as missing values. In this formulation, maximum likelihood fitting of CWM can be performed by the EM algorithm.

Remembering that in this case, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are the missing quantities, from (1.50) the

E-step can be described as follows:

$$
\begin{aligned}
Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(k)}) &= \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi};\underset{\sim}{\mathbf{X}},\mathbf{Y})\} = \\
&= \mathbb{E}_{\boldsymbol{\psi}^{(k)}}\{\sum_{n=1}^{N}\mathbf{Z}_n'\mathbf{W}_n(\mathbf{B}) + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{U}_n(\boldsymbol{\Theta}) + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{V}(\boldsymbol{\pi})\} = \\
&= \sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{\psi}^{(k)}}\{\mathbf{Z}_n|\mathbf{x}_n,y_n;\boldsymbol{\psi}^{(k)}\}[\mathbf{W}_n(\mathbf{B}) + \mathbf{U}_n(\boldsymbol{\Theta}) + \mathbf{V}(\boldsymbol{\pi})] = \\
&= \sum_{n=1}^{N}\boldsymbol{\tau}_n^{(k)\prime}\mathbf{W}_n(\mathbf{B}) + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{U}_n(\boldsymbol{\Theta}) + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{V}_n(\boldsymbol{\pi}), \qquad (1.51)
\end{aligned}
$$

where

$$
\boldsymbol{\tau}_n^{(k)} = \boldsymbol{\tau}_n(\boldsymbol{\psi}^{(k)}) = \mathbb{E}_{\psi^{(k)}}\{\mathbf{Z}_n|\mathbf{x}_n,y_n;\boldsymbol{\psi}^{(k)}\},
$$

that is

$$
\tau_{ng}^{(k)} = \frac{\pi_g^{(k)}\phi(y_n|\mathbf{x}_n,\boldsymbol{\beta}_g^{(k)})\phi_d(\mathbf{x}_n|\boldsymbol{\theta}_g^{(k)})}{\sum_{j=1}^{G}\pi_j^{(k)}\phi(y_n|\mathbf{x}_n,\boldsymbol{\beta}_j^{(k)})\phi_d(\mathbf{x}_n|\boldsymbol{\theta}_j^{(k)})} \quad n=1,\ldots,N, g=1,\ldots,G.
$$

These weights are the posterior probabilities of group membership for the n-th observation, conditional on $(\mathbf{x}_n, y_n)$ and given the current parameter estimates $\boldsymbol{\psi}_g^{(k)}$.

If the z_{ng} were observable, then the MLE of π_g would be simply given by

$$
\hat{\pi}_g = \frac{1}{n}\sum_{n=1}^{N} z_{ng} \qquad g=1,\ldots,G.
$$

The **M-step** on the $(k+1)$-th iteration simply requires replacing each z_{ng} in the previous relation by $\tau_{ng}^{(k)}$ to give

$$
\pi_g^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N}\tau_{ng}^{(k)} \qquad g=1,\ldots,G. \qquad (1.52)
$$

The estimates of the mean vectors $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G$ and covariance matrices $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_G$ for the local input densities $\phi_g(\mathbf{x}_n|\boldsymbol{\theta}_g)$ at the $(k+1)$-th iteration are then given by:

$$
\boldsymbol{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^{N}\tau_{ng}^{(k)}\mathbf{x}_n}{\sum_{n=1}^{N}\tau_{ng}^{(k)}} \quad g=1,\ldots,G \qquad (1.53)
$$

$$
\boldsymbol{\Sigma}_g^{(k+1)} = \frac{\sum_{n=1}^{N}\tau_{ng}^{(k)}(\mathbf{x}_n-\boldsymbol{\mu}_g^{(k+1)})(\mathbf{x}_n-\boldsymbol{\mu}_g^{(k+1)})'}{\sum_{n=1}^{N}\tau_{ng}^{(k)}} \quad g=1,\ldots,G. \qquad (1.54)
$$

Thus, the current estimates of the mean vectors and covariance matrices coincide with the estimates obtained in the case of Mixtures of multivariate Gaussian distributions.

Now we compute the estimates of parameters $b_{10}, \ldots, b_{G0}$ and $\mathbf{b}_1, \ldots, \mathbf{b}_G$ and variances $\sigma^2_{\epsilon,1}, \ldots, \sigma^2_{\epsilon,G}$, for the local models $\phi_g(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g)$, at the $(k+1)$-th iteration, by means of the usual statistical approach introduced for Mixtures of distributions.
The M-step computes the solutions of the equations

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial b_{g0}} = 0 \quad g = 1, \ldots, G \tag{1.55}$$

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial \mathbf{b}_g} = \mathbf{0} \quad g = 1, \ldots, G \tag{1.56}$$

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial \sigma_{\epsilon,g}} = 0 \quad g = 1, \ldots, G \tag{1.57}$$

where $\mathbb{E}_{\psi^{(k)}}$ is defined in (1.51).

From equation (1.55), for $b^{(k+1)}_{g0}$ $(g = 1, \ldots, G)$ we obtain:

$$\begin{aligned}
&\sum_{n=1}^{N} \tau^{(k)}_{ng} \frac{\partial \ln \phi(y_n|\mathbf{x}_n, \boldsymbol{\beta}^{(k)}_g)}{\partial b_{g0}} = 0 \\
&\sum_{n=1}^{N} \tau^{(k)}_{ng} \frac{y_n - (\mathbf{b}^{'(k)}_g \mathbf{x}_n + b^{(k)}_{g0})}{\sigma^{2(k)}_{yg}} = 0 \\
&\sum_{n=1}^{N} \tau^{(k)}_{ng} (y_n - \mathbf{b}^{'(k)}_g \mathbf{x}_n) = b^{(k)}_{g0} \sum_{n=1}^{N} \tau^{(k)}_{ng}
\end{aligned}$$

and then we get

$$b^{(k+1)}_{g0} = \frac{\sum_{n=1}^{N} \tau^{(k)}_{ng} (y_n - \mathbf{b}^{'(k)}_g \mathbf{x}_n)}{\sum_{n=1}^{N} \tau^{(k)}_{ng}} = \bar{y}_g - \mathbf{b}^{'(k)}_g \bar{\mathbf{x}}_g, \tag{1.58}$$

where

$$\bar{y}_g = \frac{\sum_{n=1}^{N} \tau^{(k)}_{ng} y_n}{\sum_{n=1}^{N} \tau^{(k)}_{ng}} \qquad \text{and} \qquad \bar{\mathbf{x}}_g = \frac{\sum_{n=1}^{N} \tau^{(k)}_{ng} \mathbf{x}_n}{\sum_{n=1}^{N} \tau^{(k)}_{ng}}.$$

For $\mathbf{b}_g^{(k+1)}$ $(g = 1, \ldots, G)$, equation (1.56) yields:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial \ln \phi(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g^{(k)})}{\partial \mathbf{b}_g'} = \mathbf{0}'$$

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

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

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} (y_n \mathbf{x}_n' - \mathbf{b}_g' \mathbf{x}_n \mathbf{x}_n' - \bar{y}_g \mathbf{x}_n' - \mathbf{b}_g' \bar{\mathbf{x}}_g \mathbf{x}_n')$$

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \mathbf{x}_n' - \sum_{n=1}^{N} \tau_{ng}^{(k)} \bar{y}_g \mathbf{x}_n' = \mathbf{b}_g' \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n' - \sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \bar{\mathbf{x}}_g' \right]$$

$$\overline{y\mathbf{x}}_g' - \bar{y}_g \bar{\mathbf{x}}_g' = \mathbf{b}_g' (\overline{\mathbf{x}\mathbf{x}}_g' - \bar{\mathbf{x}}_g \bar{\mathbf{x}}_g')$$

that is

$$\mathbf{b}_g^{'(k+1)} = (\overline{y\mathbf{x}}_g' - \bar{y}_g \bar{\mathbf{x}}_g')(\overline{\mathbf{x}\mathbf{x}}_g' - \bar{\mathbf{x}}_g \bar{\mathbf{x}}_g')^{-1}, \tag{1.59}$$

where

$$\overline{y\mathbf{x}}_g' = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \qquad \text{and} \qquad \overline{\mathbf{x}\mathbf{x}}_g' = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.$$

It can be demonstrated that (1.59) can be written as:

$$\mathbf{b}_g^{'(k+1)} = \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n (\mathbf{x}_n - \bar{\mathbf{x}}_g)' \right] \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}}_g)' \right]^{-1}. \tag{1.60}$$

Finally, equation (1.57) leads to the current estimate of the variance $\sigma_{\epsilon,g}^{(k)}$ $(g = 1, \ldots, G)$:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial \ln \phi(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g^{(k)})}{\partial \sigma_{\epsilon,g}^{(k)}} = 0$$

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

and solving the above equation we get

$$\sigma_{\epsilon,g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} [y_n - (\mathbf{b}_g^{'(k)} \mathbf{x}_n + b_{g0}^{(k)})]^2}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}. \tag{1.61}$$

Analogously, for the multivariate case, we have:

$$\mathbf{b}_{g0}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{y}_n - \mathbf{b}_g^{'(k)} \mathbf{x}_n)}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} = \bar{\mathbf{y}}_g - \mathbf{b}_g^{'(k)} \bar{\mathbf{x}}_g. \tag{1.62}$$

$$\mathbf{B}_g^{(k+1)} = \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{y}_n (\mathbf{x}_n - \bar{\mathbf{x}}_g)' \right] \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}}_g)' \right]^{-1}. \tag{1.63}$$

$$\mathbf{\Sigma}_{\epsilon,g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} [\mathbf{y}_n - (\mathbf{B}_g^{'(k)} \mathbf{x}_n + \mathbf{B}_{g0}^{(k+1)})][\mathbf{y}_n - (\mathbf{B}_g^{'(k+1)} \mathbf{x}_n + \mathbf{B}_{g0}^{(k+1)})]^{-1}}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}. \tag{1.64}$$

Chapter 2

Student-t CWM

2.1 Introduction

Let us introduce CWM based on another important type of elliptical distribution: the Student-t distribution. This type of data modeling has been proposed to provide more robust fitting for groups of observations with no longer than normal tails or noise data (e.g. Zellner (1976),Lange *et al.* (1989),Bernardo and Girón (1992),McLachlan and Peel (1998),McLachlan and Peel (2000),Peel and McLachlan (2000),Nadarajah and Kotz (2005),Andrews and McNicholas (2011),Baek and McLachlan (2011)). Recent applications also include analysis of orthodontic data via linear effect models (Pinheiro *et al.* (2001)), marketing data analysis (Andrews *et al.* (2002)), and asset pricing (Kan and Zhou (2003)). In particular, and different from the Gaussian case, we prove that *linear Student-t CWM* defines a wide family of probability distributions, which, under suitable assumptions, strictly includes Mixture of t-distributions (FMT) as a special case.

To begin with, we recall that a q variate random vector $\mathbf{Z}$ has a multivariate t distribution with degrees of freedom $\nu \in (0, \infty)$, location parameter $\boldsymbol{\mu} \in \mathbb{R}^q$, and $q \times q$ positive

definite inner product matrix $\boldsymbol{\Sigma}$ if it has density

$$p(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma((\nu+q)/2)\nu^{\nu/2}}{\Gamma(\nu/2)|\pi\boldsymbol{\Sigma}|^{1/2}[\nu+\delta(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})]^{(\nu+q)/2}} \tag{2.1}$$

where $\delta(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})=(\mathbf{z}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu})$ denotes the squared Mahalonobis distance between $\mathbf{z}$ and $\boldsymbol{\mu}$, with respect to matrix $\boldsymbol{\Sigma}$, and $\Gamma(\cdot)$ is the Gamma function. In this case, we write $\mathbf{Z}\sim t_q(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$, and then $\mathbb{E}(\mathbf{Z})=\boldsymbol{\mu}$ for $(\nu>1)$ and $\text{cov}(\mathbf{Z})=\nu\boldsymbol{\Sigma}/(\nu-2)$ (for $\nu>2$).

If U is a random variable, independent of $\mathbf{Z}$, such that νU has a chi-squared distribution with ν degrees of freedom, that is $\nu U\sim\chi^2_\nu$, then it is well known that $\mathbf{Z}|(U=u)\sim N_q(\boldsymbol{\mu},\Sigma/u)$.

Assume that $\mathbf{X}|\Omega_g$ has a multivariate t distribution with location parameter $\boldsymbol{\mu}_g$, inner product matrix $\boldsymbol{\Sigma}_g$, and degrees of freedom ν_g, that is, $\mathbf{X}|\Omega_g\sim t_d(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)$, and that $Y|\mathbf{x},\Omega_g$ has a t distribution with location parameter $\mu(\mathbf{x};\boldsymbol{\beta}_g)$, scale parameter $\sigma^2_{\epsilon,g}$ and degrees of freedom ζ_g, that is $Y|\mathbf{x},\Omega_g\sim t(\mu(\mathbf{x};\boldsymbol{\beta}_g),\sigma^2_{\epsilon,g}.\zeta_g),\quad g=1,\ldots,G$. Thus (1.3) specializes as:

$$p(\mathbf{x},y;\boldsymbol{\theta})=\sum_{g=1}^{G}t(y;\mu(\mathbf{x};\boldsymbol{\beta}_g),\sigma^2_{\epsilon,g},\zeta_g)t_d(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)\pi_g, \tag{2.2}$$

and this model will be referred to as *t*-CWM (Ingrassia *et al.* (2012a)).

The special case in which $\mu(\mathbf{x};\boldsymbol{\beta}_g)$ is a linear mapping will be called *linear t*-CWM:

$$p(\mathbf{x},y;\boldsymbol{\theta})=\sum_{g=1}^{G}t(y;\mathbf{b}'_g\mathbf{x}+b_{g0},\sigma^2_{\epsilon,g},\zeta_g)t_d(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)\pi_g, \tag{2.3}$$

where, according to (2.1), for $g=1,\ldots,G$, we have

$$t(y;\mathbf{b}'_g\mathbf{x}+b_{g0},\sigma^2_{\epsilon,g},\zeta_g)= \frac{\Gamma((\zeta_g+1)/2)\zeta_g^{\zeta_g/2}}{\Gamma(\zeta_g/2)\sqrt{\pi\sigma^2_{\epsilon,g}}\{\zeta_g+[y-(\mathbf{b}'_g\mathbf{x}+b_{g0})]^2/\sigma^2_{\epsilon,g}\}^{(\zeta_g+1)/2}} \tag{2.4}$$

$$t_d(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)=\frac{\Gamma((\nu_g+d)/2)\nu_g^{\nu_g/2}}{\Gamma(\nu_g/2)|\pi\boldsymbol{\Sigma}_g|^{1/2}\{\nu_g+\delta(\mathbf{x},\boldsymbol{\mu}_g,\Sigma_g)\}^{(\nu_g+d)/2}}. \tag{2.5}$$

Moreover, the posterior probability in (1.8) specializes as:

$$p(\Omega_g|\mathbf{x}, y) = \frac{t(y; \mathbf{b}_g'\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}, \zeta_g) t_d(\mathbf{x}; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}, \nu_g)\pi_g}{\sum_{j=1}^{G} t(y; \mathbf{b}_j'\mathbf{x} + b_{j0}, \sigma^2_{\epsilon,j}, \zeta_g) t_d(\mathbf{x}; \boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}, \nu_j)\pi_j} \quad g = 1, \ldots, G \quad (2.6)$$

and the decision surfaces that separate the groups are elliptical (see section 2.2) The result in the following implies that, different from the Gaussian case, *linear t*-CWM defines a larger family of probability distributions than FMT; in fact, the family of distributions generated by *linear t*-CWM strictly includes the family of distributions generated by FMT.

Proposition 6. *Let* $\mathbf{Z}$ *be a random vector defined on* $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$ *with values in* $\mathbb{R}^{d+1}$ *and set* $\mathbf{Z} = (\boldsymbol{X}', Y)'$*, where* $\boldsymbol{X}$ *is a d-dimensional input vector and Y is a random variable defined on* Ω*. Assume that the density of* $\mathbf{Z} = (\boldsymbol{X}', Y)$ *can be written in the form of a linear t-CWM (2.3), where* $\boldsymbol{X}|\Omega_g \sim t_d(\mu_g, \Sigma_g, \nu_g)$ *and* $Y|\boldsymbol{x}, \Omega_g \sim t(\boldsymbol{b}_g'\boldsymbol{x} + b_{g0}, \sigma^2_{\epsilon,g}, \zeta_g), \quad g = 1, \ldots, G$*. If* $\zeta_g = \nu_g + d$ *and* $\sigma^{2*}_{\epsilon,g} = \sigma^2_{\epsilon,g}[\nu_g + \delta(\boldsymbol{x}, \boldsymbol{\mu}_g, \Sigma_g)]/(\nu_g + d)$*, then linear t-CWM (2.3) coincides with FMT for suitable parameters* $\boldsymbol{b}_g$,b_{g0}*, and* $\sigma^2_{\epsilon,g}$*,* $g = 1, \ldots, G$*.*

Proof. Let $\mathbf{Z}$ be a q-variate random vector having multivariate t distribution (2.1) with degrees of freedom $\nu \in (0, \infty)$, location parameter $\boldsymbol{\mu}$, and positive definite inner product matrix $\boldsymbol{\Sigma}$. If $\mathbf{Z}$ is partitioned as $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2')'$, where $\mathbf{Z}_1$ takes values in $\mathbb{R}^{q_1}$ and $\mathbf{Z}_2$ in $\mathbb{R}^{q_2} = \mathbb{R}^{q-q_1}$, then $\mathbf{Z}$ con be written as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad (2.7)$$

hence, based on properties of multivariate t distribution (e.g. Dickey (1967); Liu and Rubin (1995)), it can be proven that:

$$\mathbf{Z}_1 \sim t_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \nu) \quad \text{and} \quad \mathbf{Z}_2|\mathbf{z}_1 \sim t_{q_2}(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}^*_{2|1}, \nu + q_1), \quad (2.8)$$

where

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_{2|1}(\mathbf{z}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1) \tag{2.9}$$

$$\boldsymbol{\Sigma}^*_{2|1} = \boldsymbol{\Sigma}^*_{2|1}(\mathbf{z}_1) = \frac{\nu + \delta(\mathbf{z}_1; \boldsymbol{\mu}_1, \Sigma_{11})}{\nu + q_1}\boldsymbol{\Sigma}_{2|1}, \tag{2.10}$$

with $\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\delta(\mathbf{z}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = (\mathbf{z}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1)$. In particular, if we set $\mathbf{Z} = (\mathbf{X}', Y)'$, then (2.3) coincides with FMT when $\zeta_g = \nu_g + d$ and $\sigma^{*2}_{\epsilon,g} = \sigma^2_{\epsilon,g}[\nu_g + \delta(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_g)]/(\nu_g + d)$. □

Thus, the *linear* t-CWM in (2.3) defines a wide family of densities, which strictly includes FMT as special case but is able to model more general cases.

Analogous to Gaussian case, the *linear* t-CWM in (2.3) also includes *finite mixtures of regression models with Student-t errors* (FMR-t):

$$f(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^{G} t(y; \mathbf{b}'_g\mathbf{x} + b_{g0}, \sigma^2_{\epsilon,g}, \zeta_g)\pi_g, \tag{2.11}$$

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

Moreover, because the Gaussian distribution can be regarded as the limit of the Student-t distribution, as the number of degrees of freedom tends to infinity, the *linear* t-CWM contains FMRC as a limiting special case. The relationships among $linear$ t-CWM, FMT, FMR-t and FMRC are summarized in Table 2.1. However, the analysis of finite mixtures of regressions under Student-t distribution assumptions, which have not been proposed in the literature yet (as far as the authors know), provides ideas for further research.

Finally in Table 2.2 we summarize all the models discussed in these chapters to show that linear CWM based on elliptical distributions is a general and flexible family of mixture models, which includes well-known models as special cases. We remark that if the degrees of freedom become large, *linear Gaussian* CWM can be seen as limiting special case of *linear* t-CWM

model	$p(\mathbf{x}\|\Omega_g)$	$p(y\|\mathbf{x},\Omega_g)$	parameterisation of π_g	assumptions
FMT	Student-t	Student-t	none	$\zeta_g = \nu_g + d$ and $\sigma^{2*}_{\epsilon,g} = \sigma^2_{\epsilon,g}[\nu_g + \delta(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)]/(\nu_g + d)$
FMR-t	none	Student-t	none	$(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g) = (\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$, $g = 1,\ldots,G$
FMRC	none	Student-t	logistic	$\nu_g \to \infty$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and $\pi_g = \pi$, $g = 1,\ldots,G$

Table 2.1: Relationships between *linear Student-t* CWM and Student-t mixtures.

model	$p(\mathbf{x}\|\Omega_g)$	$p(y\|\mathbf{x},\Omega_g)$	parameterisation of π_g	assumptions
CWM-t	$t_d(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)$	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	
CWM-G	$t_d(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)$	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	$\nu_g\to\infty,\zeta_g\to\infty, g=1,\ldots,G$
FMG	$t_d(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)$	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	$\nu_g\to\infty,\zeta_g\to\infty, g=1,\ldots,G$
FMT	$t_d(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)$	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	$\zeta_g=\nu_g+d$ and $\sigma^{*2}_g=\sigma^2_g[\nu_g+\delta(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)]/(\nu_g+d)$
FMR-t	none	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	$(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)=(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu), g=1,\ldots,G$
FMR	none	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	none	$\zeta_g\to\infty,\nu_g\to\infty,(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)=(\boldsymbol{\mu},\boldsymbol{\Sigma}), g=1,\ldots,G$
FMRC	none	$t(y;\mathbf{b}_g'\mathbf{x}+b_{g0},\sigma^2_{\varepsilon,g},\zeta_g)$	logistic	$\nu_g\to\infty,\boldsymbol{\Sigma}_g=\boldsymbol{\Sigma}$ and $\pi_g=\pi, g=1,\ldots,G$

Table 2.2: Overview of models included in linear CWM based on elliptical distributions.

2.2 Decision surfaces of *linear-t* CWM

In the formula (1.40) in chapter 1 we have shown that in the case of two groups, the posterior probability $p(\Omega_1|\mathbf{x}, y)$ of CWM is:

$$p(\Omega_1|\mathbf{x}, y) = \frac{1}{1 + \exp\left\{-\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} - \ln\frac{p(\Omega_1|\mathbf{x})}{p(\Omega_0|\mathbf{x})}\right\}}. \tag{2.12}$$

and that $p(\Omega_1|\mathbf{x}, y) = 0.5$ when

$$\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} + \ln\frac{p(\Omega_1|\mathbf{x})}{p(\Omega_0|\mathbf{x})} = 0, \tag{2.13}$$

which may be rewritten as:

$$\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} + \ln\frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_0)} + \ln\frac{\pi_1}{\pi_0} = 0. \tag{2.14}$$

In the *linear t*-CWM, the first and the second term in (2.14) are respectively:

$$\begin{aligned}\ln\frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_0)} = &\ln\left[\frac{\Gamma((\nu_1 + d)/2)\Gamma(\nu_0/2)}{\Gamma((\nu_0 + d)/2)\Gamma(\nu_1/2)}\right] + \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} + \\ &+ \frac{\nu_0 + d}{2}\ln\{\nu_0 + \delta(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\} \\ &- \frac{\nu_1 + d}{2}\ln\{\nu_1 + \delta(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\}\end{aligned} \tag{2.15}$$

$$\begin{aligned}\ln\frac{p(y|\mathbf{x}, \Omega_1)}{p(y|\mathbf{x}, \Omega_0)} = &\ln\left[\frac{\Gamma((\zeta_1 + 1)/2)\Gamma(\zeta_0/2)}{\Gamma((\zeta_0 + 1)/2)\Gamma(\zeta_1/2)}\right] + \ln\frac{\sigma_{\epsilon,0}}{\sigma_{\epsilon,1}} + \\ &+ \frac{\zeta_0 + 1}{2}\ln\left[\zeta_0 + \left(\frac{y - \mathbf{b}_0'\mathbf{x} - b_{00}}{\sigma_{\epsilon,0}}\right)^2\right] \\ &- \frac{\zeta_1 + 1}{2}\ln\left[\zeta_1 + \left(\frac{y - \mathbf{b}_1'\mathbf{x} - b_{10}}{\sigma_{\epsilon,1}}\right)^2\right]\end{aligned} \tag{2.16}$$

Then equation (2.14) is satisfied for $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$ such that:

$$c(\nu_0, \nu_1, \zeta_0, \zeta_1) + \ln\frac{\sigma_{\epsilon,0}}{\sigma_{\epsilon,1}} + \frac{\zeta_0 + 1}{2}\ln\left[\zeta_0 + \left(\frac{y - \mathbf{b}_0'\mathbf{x} - b_{00}}{\sigma_{\epsilon,0}}\right)^2\right] + \\ - \frac{\zeta_1 + 1}{2}\ln\left[\zeta_1 + \left(\frac{y - \mathbf{b}_1'\mathbf{x} - b_{10}}{\sigma_{\epsilon,1}}\right)^2\right] + \frac{1}{2}\ln\frac{|\Sigma_0|}{|\Sigma_1|} \\ + \frac{\nu_0 + d}{2}\ln\{\nu_0 + \delta(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\} - \frac{\nu_1 + d}{2}\ln\{\nu_1 + \delta(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} + \ln\frac{\pi_1}{\pi_0} = 0, \quad (2.17)$$

where

$$c(\nu_0, \nu_1, \zeta_0, \zeta_1) = \ln\left[\frac{\Gamma((\zeta_1 + 1)/2)\Gamma(\zeta_0/2)}{\Gamma((\zeta_0 + 1)/2)\Gamma(\zeta_1/2)}\right] + \ln\left[\frac{\Gamma((\nu_1 + d)/2)\Gamma(\nu_0/2)}{\Gamma((\nu_0 + d)/2)\Gamma(\nu_1/2)}\right]. \quad (2.18)$$

We remark that, in this case, the decision surfaces are elliptical.

2.3 Parameter Estimation of CWM via the EM algorithm - Student-t case

Given a sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ of N independent observation pairs, the cluster-weighted likelihood function is

$$L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) = \prod_{n=1}^{N} p(\mathbf{x}_n, y_n; \boldsymbol{\psi}) = \\ \prod_{n=1}^{n}\left[\sum_{g=1}^{G}\phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)\phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g) f_g(u; \zeta_g) f_g(w; \nu_g)\pi_g\right].$$

Maximization of $L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y})$ with respect to $\boldsymbol{\psi}$, for given data $(\underset{\sim}{\mathbf{X}}, \mathbf{Y})$, yields the maximum likelihood estimate of $\boldsymbol{\psi}$. Equivalently the quantity maximized is the log-likelihood $\mathcal{L}_0 = \ln L_0(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y})$.

If we consider fully categorized data:

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

then the likelihood corresponding to $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ can be written in the form

$$L_c(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) = \prod_{n=1}^{N} \prod_{g=1}^{G} \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)^{z_{ng}} \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)^{z_{ng}} f_g(u; \zeta_g)^{z_{ng}} f_g(w; \nu_g)^{z_{ng}} \pi_g^{z_{ng}}, \tag{2.19}$$

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

Consider the logarithm

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\psi}; \underset{\sim}{\mathbf{X}}, \mathbf{Y}) = \\
&= \ln \prod_{n=1}^{N} \prod_{g=1}^{G} \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)^{z_{ng}} \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)^{z_{ng}} f_g(u; \zeta_g)^{z_{ng}} f_g(w; \nu_g)^{z_{ng}} \pi_g^{z_{ng}} \\
&= \sum_{n=1}^{N} \sum_{g=1}^{G} [\ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)^{z_{ng}} + \ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)^{z_{ng}} + \ln f_g(u; \zeta_g)^{z_{ng}} + \\
&\quad + \ln f_g(w; \nu_g)^{z_{ng}} + \ln \pi_g^{z_{ng}}] = \\
&= \sum_{n=1}^{N} \sum_{g=1}^{G} [z_{ng} \ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g) + z_{ng} \ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g) + z_{ng} \ln f_g(u; \zeta_g) \\
&\quad + z_{ng} \ln f_g(w; \nu_g) + z_{ng} \ln \pi_g] = \\
&= \sum_{n=1}^{N} [\mathbf{z}_n' \ln \boldsymbol{\phi}(y_n|\mathbf{x}_n; \mathbf{B}) + \mathbf{z}_n' \ln \boldsymbol{\phi}_d(\mathbf{x}_n; \boldsymbol{\Theta}) + \mathbf{z}_n' \ln f_g(u; \boldsymbol{\zeta}) + \mathbf{z}_n' \ln f_g(w; \boldsymbol{\nu}) + \\
&\quad + \mathbf{z}_n' \ln \boldsymbol{\pi}] = \\
&= \sum_{n=1}^{N} \mathbf{z}_n' \mathbf{A}_n^{(1)}(\mathbf{B}) + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{A}_n^{(2)}(\boldsymbol{\Theta}) + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{A}_n^{(3)}(\boldsymbol{\zeta}) + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{A}_n^{(4)}(\boldsymbol{\nu}) + \\
&\quad + \sum_{n=1}^{n} \mathbf{z}_n' \mathbf{A}_n^{(5)}(\boldsymbol{\pi}),
\end{aligned} \tag{2.20}$$

where $\mathbf{A}_n^{(1)}(\mathbf{B})$ is a G-component vector having the g-th component $\ln \phi(y_n|\mathbf{x}_n; \boldsymbol{\beta}_g, \zeta_g)$, $\mathbf{A}_n^{(2)}(\boldsymbol{\Theta})$ is a G-component vector having the g-th component $\ln \phi_d(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)$, $\mathbf{A}_n^{(3)}(\boldsymbol{\zeta})$

is a G-component vector having the g-th component $\ln f_g(u;\boldsymbol{\zeta})$, $\mathbf{A}_n^{(4)}(\boldsymbol{\nu})$ is a G-component vector having the g-th component $\ln f_g(w;\boldsymbol{\nu})$ and $\mathbf{A}_n^{(5)}(\boldsymbol{\pi})$ is a G-component vector having the g-th component $\ln \pi_g$.

The form of the cluster-weighted likelihood function corresponds to the marginal density of $\mathbf{x}_1,\ldots,\mathbf{x}_N$ obtained summing (2.19) over $\mathbf{z}_1,\ldots,\mathbf{z}_N$. This emphasizes the interpretation of cluster-weighted data as incomplete data with the indicator vectors as missing values. In this formulation, maximum likelihood fitting of CWM can be performed by the EM algorithm.

Remembering that in this case, $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are the missing quantities, from (2.20) the **E-step** can be described as follows:

$$
\begin{aligned}
Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(k)}) &= \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi};\underset{\sim}{\mathbf{X}},\mathbf{Y})\} = \\
&= \mathbb{E}_{\boldsymbol{\psi}^{(k)}}\{\sum_{n=1}^{N}\mathbf{Z}_n'\mathbf{A}_n^{(1)}(\mathbf{B}) + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{A}_n^{(2)}(\boldsymbol{\Theta}) + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{A}_n^{(3)}(\boldsymbol{\zeta}) + \\
&\quad + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{A}_n^{(4)}(\boldsymbol{\nu}) + \sum_{n=1}^{n}\mathbf{Z}_n'\mathbf{A}_n^{(5)}(\boldsymbol{\pi})\} = \\
&= \sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{\psi}^{(k)}}\{\mathbf{Z}_n|\mathbf{x}_n,y_n;\boldsymbol{\psi}^{(k)}\}[\mathbf{A}_n^{(1)}(\mathbf{B}) + \mathbf{A}_n^{(2)}(\boldsymbol{\Theta}) + \mathbf{A}_n^{(3)}(\boldsymbol{\zeta}) + \mathbf{A}_n^{(4)}(\boldsymbol{\nu}) + \\
&\quad + \mathbf{V}(\boldsymbol{\pi})] = \\
&= \sum_{n=1}^{N}\boldsymbol{\tau}_n^{(k)\prime}\mathbf{A}_n^{(1)}(\mathbf{B}) + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{A}_n^{(2)}(\boldsymbol{\Theta}) + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{A}_n^{(3)}(\boldsymbol{\zeta}) + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{A}_n^{(4)}(\boldsymbol{\nu}) + \\
&\quad + \boldsymbol{\tau}_n^{(k)\prime}\mathbf{V}_n(\boldsymbol{\pi}),
\end{aligned}
\tag{2.21}
$$

where

$$
\boldsymbol{\tau}_n^{(k)} = \boldsymbol{\tau}_n(\boldsymbol{\psi}^{(k)}) = \mathbb{E}_{\psi^{(k)}}\{\mathbf{Z}_n|\mathbf{x}_n,y_n;\boldsymbol{\psi}^{(k)}\},
$$

that is

$$
\tau_{ng}^{(k)} = \frac{\pi_g^{(k)}\phi(y_n|\mathbf{x}_n,\boldsymbol{\beta}_g^{(k)})\phi_d(\mathbf{x}_n|\boldsymbol{\theta}_g^{(k)})}{\sum_{j=1}^{G}\pi_j^{(k)}\phi(y_n|\mathbf{x}_n,\boldsymbol{\beta}_j^{(k)})\phi_d(\mathbf{x}_n|\boldsymbol{\theta}_j^{(k)})} \quad n=1,\ldots,N, g=1,\ldots,G.
$$

These weights are the posterior probabilities of group membership for the n-th observation, conditional on $(\mathbf{x}_n, y_n)$ and given the current parameter estimates $\boldsymbol{\psi}_g^{(k)}$.

If the z_{ng} were observable, then the MLE of π_g would be simply given by

$$\hat{\pi}_g = \frac{1}{n}\sum_{n=1}^{N} z_{ng} \qquad g = 1, \ldots, G.$$

The **M-step** on the $(k+1)$-th iteration simply requires replacing each z_{ng} in the previous relation by $\tau_{ng}^{(k)}$ to give

$$\pi_g^{(k+1)} = \frac{1}{N}\sum_{n=1}^{N} \tau_{ng}^{(k)} \qquad g = 1, \ldots, G. \tag{2.22}$$

The estimates of the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$ and covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G$ for the local input densities $\phi_g(\mathbf{x}_n|\boldsymbol{\theta}_g)$ at the $(k+1)$-th iteration are then given by:

$$\boldsymbol{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \quad g = 1, \ldots, G \tag{2.23}$$

$$\boldsymbol{\Sigma}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)})(\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \quad g = 1, \ldots, G. \tag{2.24}$$

Thus, the current estimates of the mean vectors and covariance matrices coincide with the estimates obtained in the case of Mixtures of multivariate Gaussian distributions.

Now we compute the estimates of parameters $b_{10}, \ldots, b_{G0}$ and $\mathbf{b}_1, \ldots, \mathbf{b}_G$ and variances $\sigma^2_{\epsilon,1}, \ldots, \sigma^2_{\epsilon,G}$, for the local models $\phi_g(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g)$, at the $(k+1)$-th iteration, by means of the usual statistical approach introduced for Mixtures of distributions.
The M-step computes the solutions of the equations

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial b_{g0}} = 0 \quad g = 1, \ldots, G \tag{2.25}$$

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial \mathbf{b}_g} = \mathbf{0} \quad g = 1, \ldots, G \tag{2.26}$$

$$\frac{\partial \mathbb{E}_{\psi^{(k)}}\{\mathcal{L}_c(\boldsymbol{\psi}|\mathbf{x}_n, y_n)\}}{\partial \sigma_{\epsilon,g}} = 0 \quad g = 1, \ldots, G \tag{2.27}$$

where $\mathbb{E}_{\psi^{(k)}}$ is defined in (2.21).

From equation (2.25), for $b_{g0}^{(k+1)}$ $(g = 1, \ldots, G)$ we obtain:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial \ln \phi(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g^{(k)})}{\partial b_{g0}} = 0$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{y_n - (\mathbf{b}_g^{'(k)}\mathbf{x}_n + b_{g0}^{(k)})}{\sigma_{yg}^{2(k)}} = 0$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} (y_n - \mathbf{b}_g^{'(k)}\mathbf{x}_n) = b_{g0}^{(k)} \sum_{n=1}^{N} \tau_{ng}^{(k)}$$

and then we get

$$b_{g0}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} (y_n - \mathbf{b}_g^{'(k)}\mathbf{x}_n)}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} = \bar{y}_g - \mathbf{b}_g^{'(k)}\bar{\mathbf{x}}_g, \tag{2.28}$$

where

$$\bar{y}_g = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \qquad \text{and} \qquad \bar{\mathbf{x}}_g = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.$$

For $\mathbf{b}_g^{(k+1)}$ $(g = 1, \ldots, G)$, equation (2.26) yields:

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial \ln \phi(y_n|\mathbf{x}_n, \boldsymbol{\beta}_g^{(k)})}{\partial \mathbf{b}_g'} = \mathbf{0}'$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} [y_n - (\mathbf{b}_g'\mathbf{x}_n + b_{g0}^{(k)})]\mathbf{x}_n' = \mathbf{0}'$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} [y_n\mathbf{x}_n' - \mathbf{b}_g'\mathbf{x}_n\mathbf{x}_n' - b_{g0}^{(k)}\mathbf{x}_n'] = \mathbf{0}'$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} (y_n\mathbf{x}_n' - \mathbf{b}_g'\mathbf{x}_n\mathbf{x}_n' - \bar{y}_g\mathbf{x}_n' - \mathbf{b}_g'\bar{\mathbf{x}}_g\mathbf{x}_n')$$
$$\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n\mathbf{x}_n' - \sum_{n=1}^{N} \tau_{ng}^{(k)} \bar{y}_g\mathbf{x}_n' = \mathbf{b}_g' \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n\mathbf{x}_n' - \sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n\bar{\mathbf{x}}_g' \right]$$
$$\overline{y\mathbf{x}}_g' - \bar{y}_g\bar{\mathbf{x}}_g' = \mathbf{b}_g'(\overline{\mathbf{x}\mathbf{x}}_g' - \bar{\mathbf{x}}_g\bar{\mathbf{x}}_g')$$

that is

$$\mathbf{b}_g^{'(k+1)} = (\overline{y\mathbf{x}}_g' - \bar{y}_g\bar{\mathbf{x}}_g')(\overline{\mathbf{x}\mathbf{x}}_g' - \bar{\mathbf{x}}_g\bar{\mathbf{x}}_g')^{-1}, \tag{2.29}$$

where

$$\overline{y\mathbf{x}}'_g = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \mathbf{x}'_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \qquad \text{and} \qquad \overline{\mathbf{x}\mathbf{x}}'_g = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \mathbf{x}'_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.$$

It can be demonstrated that (2.29) can be written as:

$$\mathbf{b}_g^{\prime(k+1)} = \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n (\mathbf{x}_n - \bar{\mathbf{x}}_g)'\right] \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}}_g)'\right]^{-1}. \tag{2.30}$$

Finally, equation (2.27) leads to the current estimate of the variance $\sigma_{\epsilon,g}^{(k)}$ ($g = 1, \ldots, G$):

$$\sum_{n=1}^{N} \tau_{ng}^{(k)} \frac{\partial \ln \phi(y_n | \mathbf{x}_n, \boldsymbol{\beta}_g^{(k)})}{\partial \sigma_{\epsilon,g}^{(k)}} = 0$$

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

and solving the above equation we get

$$\sigma_{\epsilon,g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} [y_n - (\mathbf{b}_g^{\prime(k)} \mathbf{x}_n + b_{g0}^{(k)})]^2}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}. \tag{2.31}$$

Analogously, for the multivariate case, we have:

$$\mathbf{b}_{g0}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{y}_n - \mathbf{b}_g^{\prime(k)} \mathbf{x}_n)}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} = \bar{\mathbf{y}}_g - \mathbf{b}_g^{\prime(k)} \bar{\mathbf{x}}_g. \tag{2.32}$$

$$\mathbf{B}_g^{(k+1)} = \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{y}_n (\mathbf{x}_n - \bar{\mathbf{x}}_g)'\right] \left[\sum_{n=1}^{N} \tau_{ng}^{(k)} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}}_g)'\right]^{-1}. \tag{2.33}$$

$$\boldsymbol{\Sigma}_{\epsilon,g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} [\mathbf{y}_n - (\mathbf{B}_g^{\prime(k)} \mathbf{x}_n + \mathbf{B}_{g0}^{(k+1)})][\mathbf{y}_n - (\mathbf{B}_g^{\prime(k+1)} \mathbf{x}_n + \mathbf{B}_{g0}^{(k+1)})]^{-1}}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}. \tag{2.34}$$

Chapter 3

Model Based clustering via Elliptical CWM

3.1 Introduction

Let us consider a real-valued random vector $(Y, \mathbf{X}')' : \Omega \to \mathbb{R}^{d+1}$, having joint density $p(y, \mathbf{x})$ where Ω can be partitioned into G groups $\Omega_1, \ldots, \Omega_G$. Let us also assume that, for each Ω_g, the dependence of Y on $\mathbf{x}$ can be modeled by:

$$Y = \mu(\mathbf{x}, \boldsymbol{\beta}_g) + \epsilon_g = \beta_{0g} + \boldsymbol{\beta}_{1g}\mathbf{x} + \epsilon_g \tag{3.1}$$

where $\boldsymbol{\beta}_g = (\beta_{0g}, \boldsymbol{\beta}'_{1g})'$, $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = E(Y|\mathbf{X} = \mathbf{x}, \Omega_g)$ is the linear regression function and ε_g is the error variable, independent with respect to $\mathbf{X}$, with zero mean and finite constant variance σ_g^2, $g = 1, \ldots, G$.

Let us consider the class of cluster weighted models with density:

$$p(y, \mathbf{x}) = \sum_{g=1}^{G} \pi_g p(y, \mathbf{x}|\Omega_g) = \sum_{g=1}^{G} \pi_g p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}, \Omega_g). \tag{3.2}$$

Let us assume t distributions for this model. In particular we will consider:

$$p(y|\mathbf{x},\Omega_g) = h_t(y|\mathbf{x};\xi_g,\zeta_g) = \frac{\Gamma(\frac{\zeta_g+1}{2})}{(\pi\zeta_g\sigma_g^2)^{\frac{1}{2}}\{1+\delta[y,\mu(\mathbf{x};\boldsymbol{\beta}_g);\sigma_g^2]\}^{\frac{\zeta_g+1}{2}}} \tag{3.3}$$

$$p(\mathbf{x}|\Omega_g) = h_{t_d}(\mathbf{x};\boldsymbol{\theta}_g,\nu_g) = \frac{\Gamma(\frac{\nu_g+d}{2})|\boldsymbol{\Sigma}_g|^{-\frac{1}{2}}}{(\pi\nu_g)^{\frac{d}{2}}[1+\delta(\mathbf{x},\mu_g;\Sigma_g)]^{\frac{\nu_g+d}{2}}}, \tag{3.4}$$

with $\boldsymbol{\xi}_g = \{\boldsymbol{\beta}_g,\sigma_g^2\}$, $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g\}$,$\delta[y,\mu(\mathbf{x};\boldsymbol{\beta}_g);\sigma_g^2] = [y-\mu(\mathbf{x};\boldsymbol{\beta}_g)^2]/\sigma_g^2$, and $\delta(\mathbf{x},\boldsymbol{\mu}_g;\boldsymbol{\Sigma}_g) = (\mathbf{x}-\mu_g)'\Sigma_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)$. If we plug-in (3.3) and (3.4) in (3.2), we obtain the linear t CWM

$$p(y,\mathbf{x};\underset{\sim}{\psi}) = \sum_{g=1}^{G}\pi_g h_t(y|\mathbf{x};\boldsymbol{\xi}_g,\zeta_g)h_{t_d}(\mathbf{x};\boldsymbol{\theta}_g,\nu_g), \tag{3.5}$$

where the set of all unknown parameters is denoted by $\underset{\sim}{\psi} = \{\psi_1,\ldots,\psi_G\}$, with $\psi_g = \{\pi_g,\boldsymbol{\xi}_g,\zeta_g,\boldsymbol{\theta}_g,\nu_g\}$.

We will now introduce a family of twelve linear CWMs obtained from (3.5) by imposing convenient component distributional constraints. If $\zeta_g,\nu_g\to\infty$, the more famous linear Gaussian (normal) CWM is obtained as special case. The resulting models are easily interpretable and appropriate for describing various practical situations. In particular, they also allow one to infer if the group structure of the data is due to the contribution of $\mathbf{X}, Y|\mathbf{X}$, or both.

3.2 Preliminary results

We will now recall some basic ideas on model-based clustering according to the CWM approach and we will provide some preliminary results that will be useful for our family of models. Let $(y_1,\mathbf{x}_1)',\ldots,(y_N,\mathbf{x}_N')'$ be a sample of size N from (3.5). The posterior probability that a generic unit $(y_n,\mathbf{x}_n')', n = 1,\ldots,N$, comes from component Ω_g is

given by

$$\tau_{ng} = P(\Omega_g|y_n, \mathbf{x}_n; \underset{\sim}{\boldsymbol{\psi}}) = \frac{\pi_g h_t(y|\mathbf{x}; \xi_g, \zeta_g) h_{t_d}(\mathbf{x}; \boldsymbol{\theta}_g, \nu_g)}{p(y, \mathbf{x}; \underset{\sim}{\boldsymbol{\psi}})}, \quad g = 1, \ldots, G. \tag{3.6}$$

In the following propositions we will require the preliminary definition of:

$$p(y|\mathbf{x}; \underset{\sim}{\pi}, \underset{\sim}{\boldsymbol{\xi}}, \underset{\sim}{\zeta}) = \sum_{g=1}^{G} \pi_g h_t(y|\mathbf{x}, \boldsymbol{\xi}_g, \zeta_g) \tag{3.7}$$

$$p(\mathbf{x}; \underset{\sim}{\pi}, \underset{\sim}{\boldsymbol{\theta}}, \underset{\sim}{\nu}) = \sum_{g=1}^{G} \pi_g h_{t_d}(\mathbf{x}, \boldsymbol{\theta}_g, \nu_g) \tag{3.8}$$

which respectively correspond to a finite mixture of linear t regressions and a finite mixture of multivariate t distributions ($\underset{\sim}{\pi} = \{\pi_1, \ldots, \pi_{G-1}\}, \underset{\sim}{\boldsymbol{\xi}} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G\}, \underset{\sim}{\zeta} = \{\zeta_1, \ldots, \zeta_G\}, \underset{\sim}{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}, \underset{\sim}{\nu} = \{\nu_1, \ldots, \nu_G\}$).

Proposition 7. *Given $\underset{\sim}{\pi}, \underset{\sim}{\boldsymbol{\theta}}$ and $\underset{\sim}{\nu}$, if $h_t(y|\boldsymbol{x}; \boldsymbol{\xi}_1, \zeta_1) = \cdots = h_t(y|\boldsymbol{x}; \boldsymbol{\xi}_G, \zeta_G) = h_t(y|\boldsymbol{x}; \boldsymbol{\xi}, \zeta)$, then models (3.5) and (3.8) generate the same posterior probabilities.*

Proof. If the component conditional densities do not depend on Ω_g, then the posterior probabilities for the linear t CWM in (3.5) can be written as

$$\begin{aligned} \tau_{ng} &= \frac{\pi_g \cancel{h_t(y_n|\mathbf{x}_n; \boldsymbol{\xi}, \zeta)} h_{t_d}(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)}{\sum_{j=1}^{G} \pi_j \cancel{h_t(y_n|\mathbf{x}_n; \boldsymbol{\xi}, \zeta)} h_{t_d}(\mathbf{x}_n; \boldsymbol{\theta}_j, \nu_j)} \\ &= \frac{\pi_g h_{t_d}(\mathbf{x}_n; \boldsymbol{\theta}_g, \nu_g)}{\sum_{j=1}^{G} \pi_j h_{t_d}(\mathbf{x}_n; \boldsymbol{\theta}_j, \nu_j)}, \end{aligned} \tag{3.9}$$

which correspond to the posterior probabilities for model (3.8). □

Proposition 8. *Given $\underset{\sim}{\pi}, \underset{\sim}{\boldsymbol{\xi}}$ and $\underset{\sim}{\zeta}$, if $h_{t_d}(\boldsymbol{x}; \boldsymbol{\theta}_1, \nu_1) = \cdots = h_{t_d}(\boldsymbol{x}; \boldsymbol{\theta}_G, \nu_G) = h_{t_d}(\boldsymbol{x}; \boldsymbol{\theta}, \nu)$, then models (3.5) and (3.7) generate the same posterior probabilities.*

Proof. If the component marginal densities do not depend on Ω_g, then the posterior probabilities for the linear t CWM in (3.5) can be written as

$$\begin{aligned}\tau_{ng} &= \frac{\pi_g h_t\left(y_n|\mathbf{x}_n;\boldsymbol{\xi}_g,\zeta_g\right)\cancel{h_{t_d}\left(\mathbf{x}_n;\theta,\nu\right)}}{\sum_{j=1}^{G}\pi_j h_t\left(y_n|\mathbf{x}_n;\boldsymbol{\xi}_j,\zeta_j\right)\cancel{h_{t_d}\left(\mathbf{x}_n;\theta,\nu\right)}} \\ &= \frac{\pi_g h_t\left(y_n|\mathbf{x}_n;\boldsymbol{\xi}_g,\zeta_g\right)}{\sum_{j=1}^{G}\pi_j h_t\left(y_n|\mathbf{x}_n;\boldsymbol{\xi}_j,\zeta_j\right)},\end{aligned} \tag{3.10}$$

which correspond to the posterior probabilities for model (3.7). □

Let us obsere that the results in propositions 7 and 8 can be easily extended to the general CWM in (3.2).

3.3 The family of linear CWMs

In this section we will introduce the novel family of mixture models obtained from linear t CWM. More specifically in (3.5) we will consider the following conditions:

- the component conditional densities h_t have the same parameters for all Ω_g
- the component marginal densities h_{t_d} have the same parameters for all Ω_g
- the degrees of freedom ζ_g tend to infinity for each Ω_g
- the degrees of freedom ν_g tend to infinity for each Ω_g

By combining such constraints, we obtain twelve parsimonious and easily interpretable linear CWMs that are appropriate for describing various practical situations; they are schematically presented in Table 3.1 along with the number of parameters characterizing each component of the CW decomposition. For instance, if $\nu_g, \zeta_g \to \infty$ for each

Ω_g, we are assuming a normal distribution for the component conditional and marginal densities; furthermore, we can assume different linear models (in terms of β_g and σ_g^2) in each cluster while keeping the density of $\mathbf{X}$ equal between clusters. From a notational viewpoint, this leads to a linear CWM that we have simply denote as *NN-EV*: the first two letters represent the distribution of $\mathbf{X}|\Omega_g$ and $Y|\mathbf{X}, \Omega_g$ ($N \equiv$ Normal and $t \equiv t$), respectively, while the second two denote the distribution constraint between clusters ($E \equiv$ Equal and $V \equiv$ Variable) for $\mathbf{X}|\Omega_g$ and $Y|\mathbf{X}, \Omega_g$, respectively.

In principle there are sixteen models arising form the combination of the aforementioned constraints; nevertheless, four of them - those which shoulld be denoted as EE - do not make sense. Indeed, they lead to a single cluster regardless of the value of G. Finally, we remark that when $G = 1$, $VV \equiv VE \equiv EV$ regardless of the chosen distribution.

3.4 Estimation via the EM algorithm

In this section we will describe the estimation of the parameters for all linear CWMs in Table 3.1 using the EM algorithm. In the EM framework, the generic observation $(y_n, \boldsymbol{x}_n')'$ is viewed as being incomplete; its complete counterpart is given by $(y_n, \boldsymbol{x}_n', \boldsymbol{z}_n', u_n, w_n)'$, where $\boldsymbol{z}_n$ is the component-label vector in which $z_{ng} = 1$ if $(y_n, \boldsymbol{x}_n')'$ comes from the gth component and $z_{ng} = 0$ otherwise. In other words, it is convenient to view the observation augmented by $\boldsymbol{z}_n$ as still being incomplete and introduce into the complete observation the additional missing values u_n and v_n, which are defined so that $z_{ng} = 1$. In particular, from the standard theory of the (multivariate) t distribution, N independent draws from $t(\mu(\boldsymbol{x}; \boldsymbol{\beta}_g), \boldsymbol{\sigma}_g^2, \zeta_g)$ and $t_d(\mu_g, \Sigma_g, \nu_g)$ can be

Model	$\mathbf{X}\|\Omega_g$		$Y\|\mathbf{x},\Omega_g$		Number of free parameters				
Identifier	Density	Constraint	Density	Constraint	$\mathbf{X}$		$Y\|\mathbf{x}$		weights
tt-VV	t	Variable	t	Variable	$G\left(d+\frac{d(d+1)}{2}+1\right)$	+	$G\left(d+3\right)$	+	$G-1$
tt-VE	t	Variable	t	Equal	$G\left(d+\frac{d(d+1)}{2}+1\right)$	+	$d+3$	+	$G-1$
tt-EV	t	Equal	t	Variable	$d+\frac{d(d+1)}{2}+1$	+	$G\left(d+3\right)$	+	$G-1$
NN-VV	Normal	Variable	Normal	Variable	$G\left(d+\frac{d(d+1)}{2}\right)$	+	$G\left(d+2\right)$	+	$G-1$
NN-VE	Normal	Variable	Normal	Equal	$G\left(d+\frac{d(d+1)}{2}\right)$	+	$d+2$	+	$G-1$
NN-EV	Normal	Equal	Normal	Variable	$d+\frac{d(d+1)}{2}$	+	$G\left(d+2\right)$	+	$G-1$
tN-VV	t	Variable	Normal	Variable	$G\left(d+\frac{d(d+1)}{2}+1\right)$	+	$G\left(d+2\right)$	+	$G-1$
tN-VE	t	Variable	Normal	Equal	$G\left(d+\frac{d(d+1)}{2}+1\right)$	+	$d+2$	+	$G-1$
tN-EV	t	Equal	Normal	Variable	$d+\frac{d(d+1)}{2}+1$	+	$G\left(d+2\right)$	+	$G-1$
Nt-VV	Normal	Variable	t	Variable	$G\left(d+\frac{d(d+1)}{2}\right)$	+	$G\left(d+3\right)$	+	$G-1$
Nt-VE	Normal	Variable	t	Equal	$G\left(d+\frac{d(d+1)}{2}\right)$	+	$d+3$	+	$G-1$
Nt-EV	Normal	Equal	t	Variable	$d+\frac{d(d+1)}{2}$	+	$G\left(d+3\right)$	+	$G-1$

Table 3.1: Overview of linear CWMs. In "model identifier", the first and second letters represent, respectively, the density of $\mathbf{X}|\Omega_g$ and $Y|\mathbf{x},\Omega_g$ (here N $\equiv$Normal), while the third and fourth letters indicate, respectively, if $h_{t_d}\left(\mathbf{x};\boldsymbol{\vartheta}_g,\nu_g\right)$ and $h_t\left(y|\mathbf{x};\boldsymbol{\xi}_g,\zeta_g\right)$ are assumed to be Equal$\equiv$E or Variable$\equiv$V between groups.

respectively described, by compounding as

$$Y_n|\boldsymbol{x}_n, v_n, z_{ng} = 1 \overset{ind.}{\sim} N\left(\mu(\boldsymbol{x}_n;\boldsymbol{\beta}_g), \frac{\sigma_g^2}{v_n}\right) \tag{3.11}$$

$$V_n|z_{ng} = 1 \overset{i.i.d.}{\sim} \text{Gamma}\left(\frac{\zeta_g}{2}, \frac{\zeta_g}{2}\right) \tag{3.12}$$

for $n = 1, \ldots, N$ and

$$\boldsymbol{X}_n|u_n, z_{ng} = 1 \overset{ind.}{\sim} N\left(\boldsymbol{\mu}_g, \frac{\boldsymbol{\Sigma}_g}{u_n}\right) \tag{3.13}$$

$$U_n|z_{ng} = 1 \overset{i.i.d.}{\sim} \text{Gamma}\left(\frac{\nu_g}{2}, \frac{\nu_g}{2}\right) \tag{3.14}$$

for $n = 1, \ldots, N$. Because of the conditional structure of the complete-data model given by distributions (3.11),(3.12),(3.13) and (3.14), the complete-data loglikelihhod can be decomposed as

$$l_c(\underset{\sim}{\boldsymbol{\psi}}) = l_{1c}(\underset{\sim}{\pi}) + l_{2c}(\underset{\sim}{\boldsymbol{\xi}}) + l_{3c}(\underset{\sim}{\boldsymbol{\zeta}}) + l_{4c}(\underset{\sim}{\theta}) + l_{5c}(\underset{\sim}{\nu}) \tag{3.15}$$

where

$$l_{1c}(\underset{\sim}{\pi}) = \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\ln\pi_g \tag{3.16}$$

$$l_{2c}(\underset{\sim}{\boldsymbol{\xi}}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\left\{-\ln(2\pi) + \ln v_n - \ln\sigma_g^2 - v_n\delta[y_n, \mu(\boldsymbol{x}_n;\boldsymbol{\beta}_g); \sigma_g^2]\right\} \tag{3.17}$$

$$l_{3c}(\underset{\sim}{\zeta}) = \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}\left[-\ln\Gamma\left(\frac{\zeta_g}{2}\right) + \frac{\zeta_g}{2}\ln\frac{\zeta_g}{2} + \frac{\zeta_g}{2}(\ln v_n - v_n) - \ln v_n\right] \tag{3.18}$$

$$l_{4c}(\underset{\sim}{\boldsymbol{\theta}}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}[-d\ln(2\pi) + d\ln u_n - \ln|\boldsymbol{\Sigma}_g| - u_n\delta(\boldsymbol{x}_n, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g)] \tag{3.19}$$

$$l_{5c}(\underset{\sim}{\nu}) = \sum_{n=1}^{N}\sum_{g=1}^{G} z_{ng}[-\ln\Gamma\left(\frac{\nu_g}{2}\right) + \frac{\nu_g}{2}\ln\frac{\nu_g}{2} + \frac{\nu_g}{2}(\ln u_n - u_n) - \ln u_n]. \tag{3.20}$$

3.4.1 E-step

The E-step, on the $(k+1)$th iteration, requires the calculation of

$$Q\left(\underset{\sim}{\psi};\underset{\sim}{\psi}^{(k)}\right) = E_{\underset{\sim}{\psi}^{(k)}}\left[l_c\left(\underset{\sim}{\psi}\right)\left|\left(y_1,\mathbf{x}_1'\right)',\ldots,\left(y_n,\mathbf{x}_n'\right)'\right.\right]. \tag{3.21}$$

In order to do this, we need to calculate $E_{\underset{\sim}{\psi}^{(k)}}\left(Z_{ng}\left|y_n,\mathbf{x}_n\right.\right)$, $E_{\underset{\sim}{\psi}^{(k)}}\left(V_n\left|y_n,\mathbf{x}_n,\boldsymbol{z}_n\right.\right)$, $E_{\underset{\sim}{\psi}^{(k)}}\left(\widetilde{V}_n\left|y_n,\mathbf{x}_n,\boldsymbol{z}_n\right.\right)$, $E_{\underset{\sim}{\psi}^{(k)}}\left(U_n\left|\mathbf{x}_n,\boldsymbol{z}_n\right.\right)$, and $E_{\underset{\sim}{\psi}^{(k)}}\left(\widetilde{U}_n\left|\mathbf{x}_n,\boldsymbol{z}_n\right.\right)$, for $n=1,\ldots,N$ and $g=1,\ldots,G$, where $\widetilde{U}_n=\ln U_n$ and $\widetilde{V}_n=\ln V_n$.

It follows that

$$\begin{aligned} E_{\underset{\sim}{\psi}^{(k)}}\left(Z_{ng}\left|y_n,\mathbf{x}_n\right.\right) &= \tau_{ng}^{(k)} \\ &= \frac{\pi_g^{(k)} h_t\left(y_n|\mathbf{x}_n;\boldsymbol{\xi}_g^{(k)},\zeta_g^{(k)}\right) h_{t_d}\left(\mathbf{x}_n;\boldsymbol{\vartheta}_g^{(k)},\nu_g^{(k)}\right)}{p\left(y_n,\mathbf{x}_n;\underset{\sim}{\psi}^{(k)}\right)}, \end{aligned} \tag{3.22}$$

$$\begin{aligned} E_{\underset{\sim}{\psi}^{(k)}}\left(V_n\left|y_n,\mathbf{x}_n,z_{ng}=1\right.\right) &= v_{ng}^{(k)} \\ &= \frac{\zeta_g^{(k)}+1}{\zeta_g^{(k)}+\delta\left[y_n,\mu\left(\mathbf{x}_n;\boldsymbol{\beta}_g^{(k)}\right);\sigma_g^{2(r)}\right]} \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} E_{\underset{\sim}{\psi}^{(k)}}\left(U_n\left|y_n,\mathbf{x}_n,z_{ng}=1\right.\right) &= u_{ng}^{(k)} \\ &= \frac{\nu_g^{(k)}+d}{\nu_g^{(k)}+\delta\left(\mathbf{x}_n,\boldsymbol{\mu}_g^{(k)};\boldsymbol{\Sigma}_g^{(k)}\right)}, \end{aligned} \tag{3.24}$$

where the expectations are affected using the current fit $\underset{\sim}{\psi}^{(k)}$ for $\underset{\sim}{\psi}$ ($n=1,\ldots,N$ and $g=1,\ldots,G$). Regarding the last two expectations, from the standard theory on the gamma distribution, we have that

$$\begin{aligned} E_{\underset{\sim}{\psi}^{(k)}}\left(\widetilde{V}_n\left|y_n,\mathbf{x}_n,z_{ng}=1\right.\right) &= \widetilde{v}_{ng}^{(k)} \\ &= \ln v_{ng}^{(k)} + \psi\left(\frac{\zeta_g^{(k)}+1}{2}\right) - \ln\left(\frac{\zeta_g^{(k)}+1}{2}\right) \end{aligned} \tag{3.25}$$

and

$$E_{\underset{\sim}{\psi}^{(k)}}\left(\widetilde{U}_n\left|\mathbf{x}_n, z_{ng}=1\right.\right) = \widetilde{u}_{ng}^{(k)}$$
$$= \ln u_{ng}^{(k)} + \psi\left(\frac{\nu_g^{(k)}+d}{2}\right) - \ln\left(\frac{\nu_g^{(k)}+d}{2}\right), \quad (3.26)$$

where $\psi(s) = \left[\partial\Gamma(s)/\partial s\right]/\Gamma(s)$ is the Digamma function. Using the results from (3.22) to (3.25) to calculate (3.21), we have that

$$Q\left(\underset{\sim}{\psi};\underset{\sim}{\psi}^{(k)}\right) = Q_1\left(\underset{\sim}{\pi};\underset{\sim}{\psi}^{(k)}\right)+Q_2\left(\underset{\sim}{\boldsymbol{\xi}};\underset{\sim}{\psi}^{(k)}\right)+Q_3\left(\underset{\sim}{\zeta};\underset{\sim}{\psi}^{(k)}\right)+Q_4\left(\underset{\sim}{\boldsymbol{\vartheta}};\underset{\sim}{\psi}^{(k)}\right)+Q_5\left(\underset{\sim}{\nu};\underset{\sim}{\psi}^{(k)}\right), \quad (3.27)$$

where

$$Q_1\left(\underset{\sim}{\pi};\underset{\sim}{\psi}^{(k)}\right) = \sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}\ln\pi_g, \quad (3.28)$$

$$Q_2\left(\underset{\sim}{\boldsymbol{\xi}};\underset{\sim}{\psi}^{(k)}\right) = \sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}Q_{2n}\left(\boldsymbol{\xi}_g;\underset{\sim}{\psi}^{(k)}\right), \quad (3.29)$$

$$Q_3\left(\underset{\sim}{\zeta};\underset{\sim}{\psi}^{(k)}\right) = \sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}Q_{3n}\left(\zeta_g;\underset{\sim}{\psi}^{(k)}\right), \quad (3.30)$$

$$Q_4\left(\underset{\sim}{\boldsymbol{\vartheta}};\underset{\sim}{\psi}^{(k)}\right) = \sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}Q_{4n}\left(\boldsymbol{\vartheta}_g;\underset{\sim}{\psi}^{(k)}\right) \quad (3.31)$$

and

$$Q_5\left(\underset{\sim}{\nu};\underset{\sim}{\psi}^{(k)}\right) = \sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}Q_{5n}\left(\nu_g;\underset{\sim}{\psi}^{(k)}\right), \quad (3.32)$$

with

$$Q_{2n}\left(\boldsymbol{\xi}_g;\underset{\sim}{\psi}^{(k)}\right) = \frac{1}{2}\left\{-\ln(2\pi)+\widetilde{v}_{ng}^{(k)}-\ln\sigma_g^2-v_{ng}\delta\left[y_n,\mu\left(\mathbf{x}_n;\boldsymbol{\beta}_g\right);\sigma_g^2\right]\right\} \quad (3.33)$$

and

$$Q_{4n}\left(\boldsymbol{\vartheta}_g;\underset{\sim}{\psi}^{(k)}\right) = \frac{1}{2}\left[-d\ln(2\pi)+d\widetilde{u}_{ng}^{(k)}-\ln|\boldsymbol{\Sigma}_g|-u_{ng}\delta\left(\mathbf{x}_n,\boldsymbol{\mu}_g;\boldsymbol{\Sigma}_g\right)\right], \quad (3.34)$$

and where, on ignoring terms not involving ζ_g and ν_g, respectively,

$$Q_{3n}\left(\zeta_g;\underset{\sim}{\psi}^{(k)}\right) = -\ln\Gamma\left(\frac{\zeta_g}{2}\right) + \frac{\zeta_g}{2}\ln\frac{\zeta_g}{2} + \frac{\zeta_g}{2}\left[\widetilde{v}_{ng}^{(k)} - \ln v_{ng}^{(k)} + \sum_{n=1}^{N}\left(\ln v_{ng}^{(k)} - v_{ng}^{(k)}\right)\right] \tag{3.35}$$

and

$$Q_{5n}\left(\nu_g;\underset{\sim}{\psi}^{(k)}\right) = -\ln\Gamma\left(\frac{\nu_g}{2}\right) + \frac{\nu_g}{2}\ln\frac{\nu_g}{2} + \frac{\nu_g}{2}\left[\widetilde{u}_{ng}^{(k)} - \ln u_{ng}^{(k)} + \sum_{n=1}^{N}\left(\ln u_{ng}^{(k)} - u_{ng}^{(k)}\right)\right]. \tag{3.36}$$

3.4.2 M-step

On the M-step, at the $(k+1)$th iteration, it follows from (3.27) that $\underset{\sim}{\pi}^{(k+1)}$, $\underset{\sim}{\boldsymbol{\xi}}^{(k+1)}$, $\underset{\sim}{\zeta}^{(k+1)}$, $\underset{\sim}{\boldsymbol{\vartheta}}^{(k+1)}$, and $\underset{\sim}{\nu}^{(k+1)}$ can be computed independently of each other, by separate consideration of (3.28), (3.29), (3.30), (3.31), and (3.32), respectively. The solutions for $\pi_g^{(k+1)}$, $\boldsymbol{\xi}_g^{(k+1)}$, and $\boldsymbol{\vartheta}_g^{(k+1)}$ exist in closed form. Only the updates $\zeta_g^{(k+1)}$ and $\nu_g^{(k+1)}$ need to be computed iteratively. Regarding the mixture weights, maximization of $Q_1\left(\underset{\sim}{\pi};\underset{\sim}{\psi}^{(k)}\right)$ in (3.28) with respect to $\underset{\sim}{\pi}$, subject to the constraints on those parameters, is obtained by maximizing the augmented function

$$\sum_{n=1}^{N}\sum_{g=1}^{G}\tau_{ng}^{(k)}\ln\pi_g - \lambda\left(\sum_{g=1}^{G}\pi_g - 1\right), \tag{3.37}$$

where λ is a Lagrangian multiplier.

Setting the derivative of equation (3.37) with respect to π_g equal to zero and solving for pi_g yields

$$\pi_g^{(k+1)} = \sum_{n=1}^{N}\tau_{ng}^{(k)}\Big/n, \tag{3.38}$$

The updated estimates of $\boldsymbol{\vartheta}_g$, $g = 1,\ldots,G$, result

$$\boldsymbol{\mu}_g^{(k+1)} = \sum_{n=1}^{N}\tau_{ng}^{(k)}u_{ng}^{(k)}\mathbf{x}_n\Big/\sum_{n=1}^{N}\tau_{ng}^{(k)}u_{ng}^{(k)} \tag{3.39}$$

and

$$\boldsymbol{\Sigma}_g^{(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)} \left(\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}\right)' \Big/ \sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)}, \tag{3.40}$$

where, as motivated for example in Shoham (2002) the true denominator $\sum_n \tau_{ng}^{(k)}$ of (3.40) has been changed to yield a significantly faster convergence for the EM algorithm. Regarding the updated estimates of $\boldsymbol{\xi}_g$, $g = 1, \ldots, G$, maximization of (3.29), after some algebra, yields

$$\begin{aligned} \boldsymbol{\beta}_{1g}^{(k+1)} = & \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n \sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} \right)^{-1} \cdot \\ & \cdot \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} y_n \sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} \right), \end{aligned} \tag{3.41}$$

$$\beta_{0g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} - \boldsymbol{\beta}_{1g}^{(k+1)'} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}} \tag{3.42}$$

and

$$\sigma_g^{2(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)} \left[y_n - \left(\beta_{0g}^{(k+1)} + \boldsymbol{\beta}_{1g}^{(k+1)'} \mathbf{x}_n \right) \right]^2 \Big/ \sum_{n=1}^{N} \tau_{ng}^{(k)} v_{ng}^{(k)}, \tag{3.43}$$

where the denominator of (3.43) has been modified in line with what was explained for equation (3.40).

As said before, because we are acting in the most general case in which the degrees of freedom ζ_g and ν_g are inferred from the data, we need to numerically solve the equations

$$\sum_{n=1}^{N} \frac{\partial}{\partial \zeta_g} Q_{3n} \left(\zeta_g; \underset{\sim}{\boldsymbol{\psi}}^{(k)} \right) = 0 \tag{3.44}$$

and

$$\sum_{n=1}^{N} \frac{\partial}{\partial \nu_g} Q_{5n}\left(\nu_g; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = 0, \tag{3.45}$$

which correspond to finding $\zeta_g^{(k+1)}$ and $\nu_g^{(k+1)}$ as the respective solutions of

$$-\psi\left(\frac{\zeta_g}{2}\right) + \ln\frac{\zeta_g}{2} + 1 + \frac{1}{N_g^{(k)}}\sum_{n=1}^{N}\tau_{ng}^{(k)}\left(\ln v_{ng}^{(k)} - v_{ng}^{(k)}\right) + \psi\left(\frac{\zeta_g^{(k)}+1}{2}\right) - \ln\left(\frac{\zeta_g^{(k)}+1}{2}\right) = 0 \tag{3.46}$$

and

$$-\psi\left(\frac{\nu_g}{2}\right) + \ln\frac{\nu_g}{2} + 1 + \frac{1}{N_g^{(k)}}\sum_{n=1}^{N}\tau_{ng}^{(k)}\left(\ln u_{ng}^{(k)} - u_{ng}^{(k)}\right) + \psi\left(\frac{\nu_g^{(k)}+d}{2}\right) - \ln\left(\frac{\nu_g^{(k)}+d}{2}\right) = 0, \tag{3.47}$$

where $N_g^{(k)} = \sum_n \tau_{ng}^{(k)}$, $g = 1, \ldots, G$.

3.4.3 EM-constraints for parsimonious models

In the following we describe how to impose constraints on the EM algorithm, described above for the most general model tt-VV, to obtain parameter estimates for all the other models in Table 3.1. To this end, the itemization given at the beginning of Section 3.3 will be considered as a benchmark scheme.

Common t for the component marginal densities

When we constrain all the groups to have a common t distribution for $\mathbf{X}$, we have $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_G = \boldsymbol{\mu}$, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$, and $\nu_1 = \cdots = \nu_G = \nu$. Thus, in the $(k+1)$th iteration of the EM algorithm, equations (3.24) and (3.26) must be replaced by

$$u_n^{(k)} = \frac{\nu^{(k)} + d}{\nu^{(k)} + \delta\left(\mathbf{x}_n, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)}\right)} \tag{3.48}$$

and

$$\widetilde{u}_n^{(k)} = \ln u_n^{(k)} + \psi\left(\frac{\nu^{(k)}+d}{2}\right) - \ln\left(\frac{\nu^{(k)}+d}{2}\right), \tag{3.49}$$

respectively.

Furthermore, noting that $\sum_g \tau_{ng} = 1$, equations (3.31) and (3.32) can be rewritten as

$$Q_4\left(\boldsymbol{\vartheta}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \sum_{n=1}^{N} Q_{4n}\left(\boldsymbol{\vartheta}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) \tag{3.50}$$

and

$$Q_5\left(\nu; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \sum_{n=1}^{N} Q_{5n}\left(\nu; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right), \tag{3.51}$$

respectively, where

$$Q_{4n}\left(\boldsymbol{\vartheta}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \frac{1}{2}\left[-d\ln(2\pi) + d\widetilde{u}_n^{(k)} - \ln|\boldsymbol{\Sigma}| - u_n\delta\left(\mathbf{x}_n, \boldsymbol{\mu}; \boldsymbol{\Sigma}\right)\right] \tag{3.52}$$

and

$$Q_{5n}\left(\nu; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = -\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\ln\frac{\nu}{2} + \frac{\nu}{2}\left[\widetilde{u}_n^{(k)} - \ln u_n^{(k)} + \sum_{n=1}^{N}\left(\ln u_n^{(k)} - u_n^{(k)}\right)\right]. \tag{3.53}$$

Maximization of (3.50), with respect to $\boldsymbol{\vartheta}$, leads to

$$\boldsymbol{\mu}^{(k+1)} = \sum_{n=1}^{N} u_n^{(k)}\mathbf{x}_n \Big/ \sum_{n=1}^{N} u_n^{(k)} \tag{3.54}$$

and

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{n=1}^{N} u_n^{(k)}\left(\mathbf{x}_n - \boldsymbol{\mu}^{(k+1)}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}^{(k+1)}\right)' \Big/ \sum_{n=1}^{N} u_n^{(k)}. \tag{3.55}$$

For the updating of ν, we need to numerically solve the equation

$$\sum_{n=1}^{N} \frac{\partial}{\partial\nu} Q_{5n}\left(\nu; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = 0, \tag{3.56}$$

which corresponds to finding $\nu^{(k+1)}$ as the solution of

$$-\psi\left(\frac{\nu}{2}\right) + \ln\frac{\nu}{2} + 1 + \sum_{n=1}^{N}\left(\ln u_n^{(k)} - u_n^{(k)}\right) + \psi\left(\frac{\nu^{(k)}+d}{2}\right) - \ln\left(\frac{\nu^{(k)}+d}{2}\right) = 0. \tag{3.57}$$

Common t for the component conditional densities

Similarly, when we constrain all the groups to have a common t distribution for $Y|\mathbf{x}$, we have $\boldsymbol{\beta}_{11} = \cdots = \boldsymbol{\beta}_{1G} = \boldsymbol{\beta}_1$, $\beta_{01} = \cdots = \beta_{0G} = \beta_0$, $\sigma_1^2 = \cdots = \sigma_G^2 = \sigma^2$, and $\zeta_1 = \cdots = \zeta_G = \zeta$. Thus, in the $(k+1)$th iteration of the EM algorithm, equations (3.23) and (3.25) must be replaced by

$$v_n^{(k)} = \frac{\zeta^{(k)} + 1}{\zeta_g^{(k)} + \delta\left[y_n, \mu\left(\mathbf{x}_n; \boldsymbol{\beta}^{(k)}\right); \sigma^{2(r)}\right]} \tag{3.58}$$

and

$$\widetilde{v}_n^{(k)} = \ln v_n^{(k)} + \psi\left(\frac{\zeta^{(k)} + 1}{2}\right) - \ln\left(\frac{\zeta^{(k)} + 1}{2}\right), \tag{3.59}$$

respectively.

Also, equations (3.29) and (3.30) can be rewritten as

$$Q_2\left(\boldsymbol{\xi}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \sum_{n=1}^{N} Q_{2n}\left(\boldsymbol{\xi}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) \tag{3.60}$$

and

$$Q_3\left(\zeta; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \sum_{n=1}^{N} Q_{3n}\left(\zeta; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right), \tag{3.61}$$

respectively, where

$$Q_{2n}\left(\boldsymbol{\xi}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = \frac{1}{2}\left\{-\ln(2\pi) + \widetilde{v}_n^{(k)} - \ln\sigma^2 - v_n\delta\left[y_n, \mu\left(\mathbf{x}_n; \boldsymbol{\beta}\right); \sigma^2\right]\right\} \tag{3.62}$$

and

$$Q_{3n}\left(\zeta; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right) = -\ln\Gamma\left(\frac{\zeta}{2}\right) + \frac{\zeta}{2}\ln\frac{\zeta}{2} + \frac{\zeta}{2}\left[\widetilde{v}_n^{(k)} - \ln v_n^{(k)} + \sum_{n=1}^{N}\left(\ln v_n^{(k)} - v_n^{(k)}\right)\right]. \tag{3.63}$$

Maximization of (3.60), with respect to $\boldsymbol{\xi}$, leads to the updates

$$\begin{aligned}\boldsymbol{\beta}_1^{(k+1)} &= \left(\frac{\sum_{n=1}^{N} v_n^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^{N} v_n^{(k)}} - \frac{\sum_{n=1}^{N} v_n^{(k)} \mathbf{x}_n \sum_{n=1}^{N} v_n^{(k)} \mathbf{x}_n'}{\sum_{n=1}^{N} v_n^{(k)} \sum_{n=1}^{N} v_n^{(k)}} \right)^{-1} \cdot \\ &\cdot \left(\frac{\sum_{n=1}^{N} v_n^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^{N} v_n^{(k)}} - \frac{\sum_{n=1}^{N} v_n^{(k)} y_n \sum_{n=1}^{N} v_n^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} v_n^{(k)} \sum_{n=1}^{N} v_n^{(k)}} \right), \end{aligned} \tag{3.64}$$

$$\beta_0^{(k+1)} = \frac{\sum_{n=1}^{N} v_n^{(k)} y_n}{\sum_{n=1}^{N} v_n^{(k)}} - \boldsymbol{\beta}_1^{(k+1)'} \frac{\sum_{n=1}^{N} v_n^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} v_n^{(k)}}$$

and

$$\sigma^{2(k+1)} = \sum_{n=1}^{N} v_n^{(k)} \left[y_n - \left(\beta_0^{(k+1)} + \boldsymbol{\beta}_1^{(k+1)'} \mathbf{x}_n \right) \right]^2 \Big/ \sum_{n=1}^{N} v_n^{(k)}.$$

For the updating of ζ, we need to numerically solve the equation

$$\sum_{n=1}^{N} \frac{\partial}{\partial \nu} Q_{3n} \left(\zeta; \underset{\sim}{\boldsymbol{\psi}}^{(k)} \right) = 0, \tag{3.65}$$

which corresponds to finding $\zeta^{(k+1)}$ as the solution of

$$-\psi\left(\frac{\zeta}{2}\right) + \ln\frac{\zeta}{2} + 1 + \sum_{n=1}^{N} \left(\ln v_n^{(k)} - v_n^{(k)} \right) + \psi\left(\frac{\zeta^{(k)}+1}{2}\right) - \ln\left(\frac{\zeta^{(k)}+1}{2}\right) = 0. \tag{3.66}$$

Normal component marginal densities

The normal case for the component distributions of $\mathbf{X}$ can be obtained, as stated previously, as a limiting case when $\nu_g \to \infty$, $g = 1, \ldots, G$. Then, in (3.24), $u_{ng}^{(k)} \to 1$. Substituting this value into (3.39) and (3.40), we obtain

$$\boldsymbol{\mu}_g^{(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \Big/ \sum_{n=1}^{N} \tau_{ng}^{(k)}$$

and

$$\boldsymbol{\Sigma}_g^{(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \left(\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}\right)' \Big/ \sum_{n=1}^{N} \tau_{ng}^{(k)}.$$

Naturally, in this case, we do not compute the additional M-step maximizing $Q_5\left(\underset{\sim}{\nu}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right)$ in (3.32). Accordingly, for the sub-case $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_G = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$, in equation (3.48) we have $u_n^{(k)} \to 1$ and the updated estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ become

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{n=1}^{N} \mathbf{x}_n$$

and

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}\right)',$$

which do not depend on the EM-iterations.

Normal component conditional densities

The normal case for the component distributions of $Y|\mathbf{X}$ can be obtained as a limiting case when $\zeta_g \to \infty$, $g = 1, \ldots, G$. Then, in (3.23), $v_{ng}^{(k)} \to 1$.

Substituting this value into (3.41) and (3.42), we obtain

$$\begin{aligned}\boldsymbol{\beta}_{1g}^{(k+1)} = & \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}\right)^{-1} \cdot \\ & \cdot \left(\frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}\right), \end{aligned} \tag{3.67}$$

$$\beta_{0g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} y_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} - \boldsymbol{\beta}_{1g}^{(k+1)'} \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{(k)}} \tag{3.68}$$

and

$$\sigma_g^{2(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} \left[y_n - \left(\beta_{0g}^{(k+1)} + \boldsymbol{\beta}_{1g}^{(k+1)'} \mathbf{x}_n \right) \right]^2 \Big/ \sum_{n=1}^{N} \tau_{ng}^{(k)}. \tag{3.69}$$

We again do not compute the additional M-step maximizing $Q_3\left(\underset{\sim}{\zeta}; \underset{\sim}{\boldsymbol{\psi}}^{(k)}\right)$ in (3.30). Accordingly, for the sub-case $\boldsymbol{\beta}_{11} = \cdots = \boldsymbol{\beta}_{1G} = \boldsymbol{\beta}_1$, $\beta_{01} = \cdots = \beta_{0G} = \beta_0$, and $\sigma_1^2 = \cdots = \sigma_G^2 = \sigma^2$, in equation (3.58) we have $v_n^{(k)} \to 1$ and the updated estimates of $\boldsymbol{\beta}_1$, β_0, and σ^2 become

$$\boldsymbol{\beta}_1 = \left(\frac{1}{n} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n' - \frac{1}{n^2} \sum_{n=1}^{N} \mathbf{x}_n \sum_{n=1}^{N} \mathbf{x}_n' \right)^{-1} \left(\frac{1}{n} \sum_{n=1}^{N} y_n \mathbf{x}_n - \frac{1}{n^2} \sum_{n=1}^{N} y_n \sum_{n=1}^{N} \mathbf{x}_n \right), \tag{3.70}$$

$$\beta_0 = \frac{1}{n} \sum_{n=1}^{N} y_n - \frac{1}{n} \boldsymbol{\beta}_1' \sum_{n=1}^{N} \mathbf{x}_n \tag{3.71}$$

and

$$\sigma^2 = \frac{1}{n} \sum_{n=1}^{N} \left[y_n - (\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_n) \right]^2, \tag{3.72}$$

which do not depend on the EM-iterations.

Chapter 4

An R package for Cluster Weighted Modeling

4.1 Introduction

In order to implement the procedure described in chapter 3 we implemented an R package that makes the optimal clustering of data among the models specified in Table 3.1.

4.1.1 The CWM function

The core of the package consists of the function `cwm` which returns an object of class cwm. This function, besides the data, permits to specify the models and the number of groups where the optimal result should be found. It permits also to specify the criterion used to find the optimal clustering of the data. From this point of view, two criteria can be chosen: BIC or ICL. Finally the function `cwm` permits to specify the maximum number of iteration before stopping the algorithm described previously in chapter 3.

The function `cwm` uses an internal function `.MS` which makes the initialization of the algorithm and the parameter fitting for each selected model.

The initialization process is implemented by using a hierarchical scheme which will be described below, see 4.1.1.

After the initialization of each model with this scheme, the texttt.MS tries to fit the parameters of each model through the internal function `.tCWM`. Given the the initial clustering of the observations, this function first makes an approximate estimation of the parameters for each group by using the function `mst.mle`. More specifically we assume that the observations of each group have a multivariate student-t distribution and then we use the function `mst.mle` to evaluate the parameters of this distribution from the observations that initially belong to the group. Then, starting from this initial estimates, the function `.tCWM` applies the algorithm of chapter 3 until it converges. A numerical search for the estimates of the degrees of freedom was carried out using the `uniroot` command in the `stats` package. This command is based on the Fortran subroutine `zeroin` described by Brent (2002). In order to expedite convergence, the range of values for ν_g, ζ_g, ν, and ζ was restricted to $(2, 200]$. Previous work in the context of model-based clustering (Andrews and McNicholas (2011)) and some experiments whose results are not reported here suggest that these restrictions do not hamper classification performance and show that the upper limit of 200 does not thwart the recovery of an underlying normal structure.

Finally `.tCWM` evaluates the BIC or the ICL coefficient for the considered model with the parameters fitted in the previous step. These coefficients, together with the fitted parameters, are returned to the function `.MS`. The function `.tCWM` also returns an estimate of classification of the observations for the fitted model.

Once the function `.MS` has obtained from `.tCWM` the values of the BIC or of the ICL

for the set of models and the set of groups specified in the call to the function `cwm`, it chooses the optimal solution and returns it to this function together with the relative fitted parameters and the classification of the observations.

Model initialization

As for the initialization, it is well known that the choice of starting values represents an important issue in the EM algorithm. The standard initialization consists of selecting a value for $\underset{\sim}{\psi}^{(0)}$. An alternative approach, more natural in the authors' opinion, is to specify a value for $\boldsymbol{z}_n^{(0)}$, $n = 1, \ldots, N$ (McLachlan and Peel (2000)). Within this approach, and due to the structure of our family of linear CWMs, we propose a random-hierarchical initialization procedure that helps in obtaining the natural ranking among the likelihoods.

For a fixed G, we start by considering NN-VE and NN-EV, because the former is nested in all of the VE-models, the latter is nested in all of the EV models, and both are nested in all of the VV-models. For NN-VE and NN-EV only, a random initialization is repeated 10 times, from different random positions, and the solution maximizing the likelihood among these 10 runs is selected. Note that, as underlined by Andrews *et al.* (2011), mixtures based on the multivariate t distribution are more sensitive to bad starting values than their Gaussian counterparts. Thus, by considering random initialization only for the above models of type NN, we prevent the possible failure of the algorithm due to poor starting values for models of type Nt, tN, and tt. In each run, the N vectors $\boldsymbol{z}_n^{(0)}$ are randomly drawn from a multinomial distribution with probabilities $(1/G, \ldots, 1/G)$. Once the EM-estimates $\widehat{\tau}_{ng}^{NN\text{-VE}}$ and $\widehat{\tau}_{ng}^{NN\text{-EV}}$ of the posterior probabilities have been obtained for these models, we can compute the maximum *a posteriori* (MAP) classification, say $\text{MAP}\left(\widehat{\tau}_{ng}^{NN\text{-VE}}\right) = \widehat{z}_{ng}^{NN\text{-VE}}$ and $\text{MAP}\left(\widehat{\tau}_{ng}^{NN\text{-EV}}\right) = \widehat{z}_{ng}^{NN\text{-EV}}$,

where

$$\text{MAP}\left(\widehat{\tau}_{ng}\right) = \widehat{z}_{ng} = \begin{cases} 1 & \text{if } \max_j \{\widehat{\tau}_{nj}\} \text{ occurs in component } g \\ 0 & \text{otherwise.} \end{cases}$$

Then, the hierarchical initialization procedure proceeds according to the scheme in Figure 4.1, where each arrow is directed from the model used for initialization to the model to be estimated. Thus, $\widehat{z}_{ng}^{NN\text{-VE}}$ is used to initialize the EM of both tN-VE and Nt-VE,

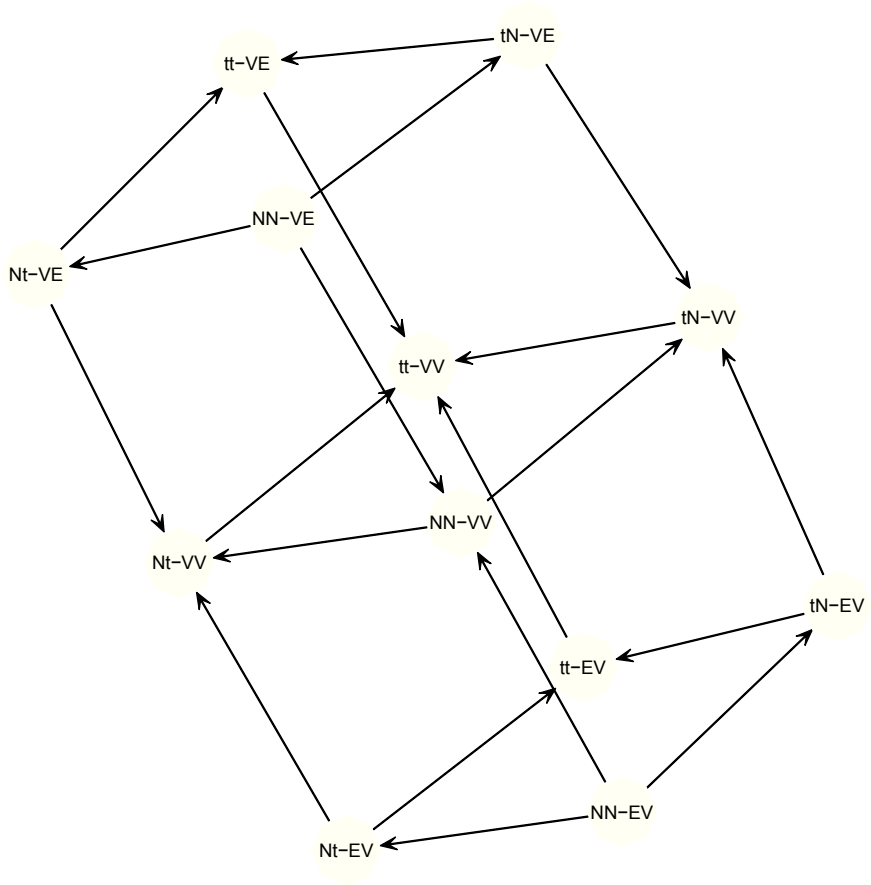

Figure 4.1: Relationships among the models in the hierarchical initialization strategy. Arrows are oriented from the model used for initialization to the model to be estimated.

obtaining $\widehat{z}_{ng}^{tN\text{-VE}}$ and $\widehat{z}_{ng}^{Nt\text{-VE}}$, respectively, while $\widehat{z}_{ng}^{NN\text{-EV}}$ is used to initialize the EM of

both tN-EV and Nt-EV, leading to $\widehat{z}_{ng}^{tN\text{-EV}}$ and $\widehat{z}_{ng}^{Nt\text{-EV}}$, respectively. Also, following the same principle, the model between NN-VE and NN-EV leading to the maximum likelihood is used to initialize the EM for NN-VV. Without going into further details on this hierarchical procedure, in the last step the model between Nt-VV, tN-VV, tt-VE, and tt-EV leading to the maximum likelihood is used to initialize the EM of tt-VV.

4.1.2 Output interface

The object of class cwm returned by the function `cwm`, contains information about the optimal model found for the specified options.

This information can be retrieved through the summary function. More specifically this function returns the label of the optimal model, the optimal number of groups, the criterion used to find the optimal model and the value of its related parameter, and the number of observations associated with each group.

The summary function has also the following options which permit to show further information about the optimal model:

- if the `parameters` option is set to `TRUE` the parameters of the optimal model will be shown
- if the `classification` option is set to `TRUE` the group of each observation will be shown
- if the `designMatrix` option is set to `TRUE` the values of the BIC or ICL parameters for each model specified in the options of the cwm function will be shown
- if the `posterior` option is set to `TRUE` posterior probabilities of each observation in the optimal model will be shown

- if the `indicator` option is set to `TRUE` an indicator matrix of the group of each observation in the optimal model will be shown

The package cwm contains also a plot function for objects of class cwm, which plots the values of the BIC or of the ICL for all the models considered during the search for the optimal model.
Finally the package has also the function `cwmModelNames` which gives a small explanation of the models presented in Table 3.1

In the following sections we will show a detailed explanation of the behaviour of each function.

R documentation

of 'cwm'

---

cwm *Linear Cluster-Weighted Models*

Description

Select the optimal model, in a family of linear cluster-weighted models, according to a (chosen by the user) likelihood-based selection criteria. The EM-algorithm is used to obtain maximum likelihood estimates of the parameters for the models.

Usage

```
cwm(X, Y, G = NULL, modelNames = NULL, method = "BIC", iter.max = 500)
```

Arguments

X	A numeric design matrix (or data frame) of covariates. Rows correspond to observations and columns correspond to variables.
Y	Vector of (unidimensional) observations for the response variable. Its length must coincide with the number of rows of X.

G	An integer vector specifying the numbers of clusters among which the choice of the optimal model is to be done. The default is G=1:3.
modelNames	A vector of character strings indicating the models to be fitted in the EM phase of clustering. The help file for cwmModelNames describes the available models. The default is: c("NN-VE","NN-EV","NN-VV","tN-VE","tN-EV","tN-VV","Nt-VE", "Nt-EV","Nt-VV","tt-VE","tt-EV","tt-VV")
method	Adopted model selection criterion. Possible choices are "BIC" or "ICL".
iter.max	Maximum number of iterations for each fitted model.

Value

An object of class "cwm" providing the optimal model estimation (according to the selected method).

The details of the output components are as follows:

call	The matched call
n	The number of observations in the data.
d	The dimension of the data.
X	The matrix of covariates data.
Y	The vector of responses.
loglik	The loglikelihoods of the entire set of models.
BIC	All BIC values.
ICL	All BIC values.

`method` The method used to select the best model.

`G` The optimal number of components.

`model` A character string denoting the selected model.

`bestLoglik` The loglikelihood corresponding to the optimal model.

`method.value` The value of the prarameter used to select the best model

`parameters` A list with the following components:

- `pi` A vector whose *k*th component is the mixing proportion for the *k*th component of the cluster weighted model.
- `MX` The mean of the covariates variables for each component in case of normal distribution of the covariates; otherwise it is the MX paramater of the t-student distribution of the covariates.
- `SX` The covariance matrix of the covariates for each component in case of normal distribution of the covariates; otherwise it is the SX parameter of the t-student distribution of the covariates.
- `nu` The degrees of freedom of the covariate variables (`Inf` for the normal case)
- `SY` The covariance matrix of the conditionale responce for each component in case of normal distribution of the response; otherwise it is the SY parameter or the t-student distribution of the repsonse.
- `zeta` The degrees of freedom of the response variable (Inf for the normal case)
- `B` The regression coefficients for each component
- `B0` The intercept for each component

group	The classification of each observation.
z	An n x g indicator matrix of 0 and 1 of the group of each observation.
tau	A matrix whose *[i,k]*th entry is the probability that observation *i* in the test data belongs to the *k*th group.

Author(s)

G. Incarbone, A. Punzo, S. Ingrassia.

References

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*. 29 (3), 363-401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2012). Model-based clustering via linear cluster-weighted models. arXiv.org e-print 1206.3974, available at: http://arxiv.org/abs/1206.3974.

Examples

```
library(Flury)
data(m.twins)
Y <- m.twins[,5]
X <- m.twins[,c(2,3,4,6,7)]
res<-cwm(X=X,Y=Y,G=1:2,modelNames=c("NN-EV","NN-VV","tN-EV","tN-VV","Nt-VE"),
method="ICL")
```

R documentation

of ‘cwmModelNames’

---

cwmModelNames *Names of the Linear Cluster-Weighted Models*

Description

Description of model names (modelNames) used in the *CWM* package.

Usage

```
cwmModelNames(model)
```

Arguments

model	A string specifying the model.

Details

The following models are available:

"NN-VE" normal (N) distribution for the component density of X, normal (N) distribution for the component density of Y|x, variable (V) component densities

of X between clusters, and equal (E) component regression models between clusters.

"NN-EV" normal (N) distribution for the component density of X, normal (N) distribution for the component density of Y|x, equal (E) component densities of X between clusters, and variable (V) component regression models between clusters.

"NN-VV" normal (N) distribution for the component density of X, normal (N) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and variable (V) component regression models between clusters.

"tN-VE" t (t) distribution for the component density of X, normal (N) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and equal (E) component regression models between clusters.

"tN-EV" t (t) distribution for the component density of X, normal (N) distribution for the component density of Y|x, equal (E) component densities of X between clusters, and variable (V) component regression models between clusters.

"tN-VV" t (t) distribution for the component density of X, normal (N) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and variable (V) component regression models between clusters.

"Nt-VE" normal (N) distribution for the component density of X, normal (N) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and equal (E) component regression models between clusters.

"Nt-EV" normal (N) distribution for the component density of X, t (t) distribution for the component density of Y|x, equal (E) component densities of X between clusters, and variable (V) component regression models between clusters.

"Nt-VV" normal (N) distribution for the component density of X, t (t) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and variable (V) component regression models between clusters.

"tt-VE" t (t) distribution for the component density of X, t (t) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and equal (E) component regression models between clusters.

"tt-EV" t (t) distribution for the component density of X, t (t) distribution for the component density of Y|x, equal (E) component densities of X between clusters, and variable (V) component regression models between clusters.

"tt-VV" t (t) distribution for the component density of X, t (t) distribution for the component density of Y|x, variable (V) component densities of X between clusters, and variable (V) component regression models between clusters.

Value

`model`	A character string indicating the model (as in input).
`type`	The description of the indicated model (see details).

Author(s)

G. Incarbone, A. Punzo, S. Ingrassia.

References

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*. 29 (3), 363-401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2012). Model-based clustering via linear cluster-weighted models. arXiv.org e-print 1206.3974, available at: http://arxiv.org/abs/1206.3974.

See Also

cwm

Examples

```
cwmModelNames("NN-VE")
cwmModelNames("NN-EV")
cwmModelNames("Nt-EV")
```

R documentation

of 'plot.cwm'

---

plot.cwm *Plot of BIC or ICL*

Description

This function plots the BIC or the ICL versus the number of considered groups and for all the considered linear cluster-weighted models.

Usage

```
plot.cwm(object, G = NULL, modelNames = NULL, symbols = NULL,
colors = NULL, xlab = NULL, ylim = NULL, legendArgs = list(x = "bottomright",
ncol = 2, cex = 1), ...)
```

Arguments

object	An object of class cwm resulting from a call to cwm.
G	One or more numbers of components corresponding to models fitted in object. The default is to plot the BIC or the ICL for all of the numbers of components fitted.

`modelNames`	One or more model names corresponding to models fitted in `object`. The default is to plot the BIC or the ICL for all of the models fitted.
`symbols`	Either an integer or character vector assigning a plotting symbol to each model.
`colors`	Either an integer or character vector assigning a plotting symbol to each model.
`xlab`	Optional label for the horizontal axis of the plot.
`ylim`	Optional limits for the vertical axis of the plot.
`legendArgs`	Arguments to pass to the `legend` function. Set to `NULL` for no legend.
`...`	Other graphics parameters.

Value

A plot of the BIC or the ICL values for the models specified in the `modelNames` argument.

Author(s)

G. Incarbone, A. Punzo, S. Ingrassia.

References

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*. 29 (3), 363-401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2012). Model-based clustering via linear cluster-weighted models. arXiv.org e-print 1206.3974, available at: http://arxiv.org/abs/1206.3974.

Examples

```
library(Flury)
data(m.twins)
Y <- m.twins[,5]            # response variable
X <- m.twins[,c(2,3,4,6,7)] # covariates
res<-cwm(X=X,Y=Y,G=1:2,modelNames=c("NN-EV","NN-VV","tN-EV","tN-VV","Nt-VE"),
method="ICL")
summary(res)
```

R documentation

of 'summary.cwm'

---

summary.cwm *Summarizing Cluster-Weighted Model Fits*

Description

Summary method for class "CWM".

Usage

```
summary.cwm(object, parameters = FALSE, classification = FALSE,
designMatrix = FALSE, posterior = FALSE, indicator = FALSE...)
```

Arguments

object	An object of class "CWM" resulting from a call to cwm.
parameters	Logical; if TRUE, the parameters of the selected linear cluster-weighted model are printed.
classification	Logical; if TRUE, the maximum a posteriori (MAP) classification of the observations is printed.

`designMatrix` Logical; if `TRUE`, the BIC or the ICL for the specified linear cluster weighted models and the numbers of clusters are printed.

`posterior` if `TRUE`, the posterior probabilities of group are printed.

`indicator` if `TRUE`, the MAP classification is printed using an indication matrix.

Author(s)

G. Incarbone, A. Punzo, S. Ingrassia.

References

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*. 29 (3), 363-401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2012). Model-based clustering via linear cluster-weighted models. arXiv.org e-print 1206.3974, available at: http://arxiv.org/abs/1206.3974.

Examples

```
library(Flury)
data(m.twins)
Y <- m.twins[,5]            # response variable
X <- m.twins[,c(2,3,4,6,7)] # covariates
res<-cwm(X=X,Y=Y,G=1:2,modelNames=c("NN-EV","NN-VV","tN-EV","tN-VV","Nt-VE"),
method="ICL")
summary(res)
```

```
summary(res,parameters=TRUE, classification=TRUE, designMatrix=TRUE,
posterior=TRUE, indicator = TRUE)
```

Bibliography

Anderson, J. (1972). Separate sample logistic discrimination. *Biometrika*, **59**(1), 19–35.

Andrews, J. and McNicholas, P. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**(3), 361–373.

Andrews, J., McNicholas, P., and Subedi, S. (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics & Data Analysis*, **55**(1), 520–529.

Andrews, R., Ansari, A., and Currim, I. (2002). Hierarchical bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research*, pages 87–98.

Baek, J. and McLachlan, G. (2011). Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics*, **27**(9), 1269–1276.

Bernardo, J. and Girón, F. (1992). Robust sequential prediction from non-random samples: the election night forecasting case. *Bayesian Statistics*, **4**, 61–77.

Brent, R. (2002). *Algorithms for minimization without derivatives*. Dover Publications.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *Knowledge and Data Engineering, IEEE Transactions on*, **8**(2), 195–210.

Cleveland, W. and Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

Dayton, C. and Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, **83**(401), 173–178.

DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, **5**(2), 249–282.

Dickey, J. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *The Annals of Mathematical Statistics*, **38**(2), 511–518.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer.

Gershenfeld, N. (1997). Non linear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(..), 18–24.

Gershenfeld, N. (1998). *The Nature of Mathematical Modeling*. Cambridge university press.

Gershenfeld, N., Schoner, B., and Metois, E. (1999). Cluster-weighted modelling for time-series analysis. *Nature*, **397**(6717), 329–332.

Heckerman, D. and Wellman, M. (1995). Bayesian networks. *Communications of the ACM*, **38**(3), 27–30.

Ingrassia, S., Minotti, S., and Incarbone, G. (2012a). An em algorithm for the student-t cluster-weighted modeling. In *Challenges at the Interface of Data Analysis, Com-*

puter Science, and Optimization: Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V., Karlsruhe, July 21-23, 2010, page 13. Springer.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012b). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.

Jordan, M. *et al.* (1995). Why the logistic function? a tutorial discussion on probabilities and neural networks.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, **6**(2), 181–214.

Jordan, M. I. (1999). *Learning in graphical models*. MIT Press Cambridge, MA, USA.

Kan, R. and Zhou, G. (2003). Modeling non-normality using multivariate t: Implications for asset pricing. Technical report, Citeseer.

Kent, J., Bibby, J., and Mardia, K. (1979). Multivariate analysis. *London, Aca*.

Lange, K., Little, R., and Taylor, J. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, **84**(408), 881–896.

Liu, C. and Rubin, D. (1995). Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, **5**(1), 19–39.

McLachlan, G. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. *Advances in pattern recognition*, pages 658–666.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*, volume 299. Wiley-Interscience.

Minotti, S. and Vittadini, G. (2010). Local multilevel modeling for comparison of istitutional performances. *Data Analysis and Classification: from the exploratory to the confirmatory approach*, pages 289–298.

Minotti S.C., V. G., S. (2011). Some notes on the applicability of cluster-weighted modeling in effectiveness studies. In *New Perspectives in Statistical Modeling and Data Analysis: Proceedings of the 7th Conference of the Classification and Data Analysis Group of the Italian Statistical Society, Catania, September 9-11, 2009*, volume 7, page 57. Springer.

Nadarajah, S. and Kotz, S. (2005). Mathematical properties of the multivariate t distribution. *Acta Applicandae Mathematicae*, **89**(1), 53–84.

Neal, R. (1995). *Bayesian learning for neural networks*. Ph.D. thesis, University of Toronto.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Pinheiro, J., Liu, C., and Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, **10**(2), 249–276.

Schöner, B. and Gershenfeld, N. (2001). Cluster weighted modeling: Probabilistic time series prediction, characterization, and synthesis. In A. Mees, editor, *Nonlinear Dynamics and Statistics*, Boston. Birkhauser.

Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, **35**(5), 1127–1142.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Wedel, M. and DeSarbo, W. (2002). Market segment derivation and profiling via a finite mixture model framework. *Marketing Letters*, **13**(1), 17–25.

Zellner, A. (1976). Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms. *Journal of the American Statistical Association*, **71**(354), 400–405.