

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Galimberti G, Soffritti G (2020). A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models. STATISTICS & PROBABILITY LETTERS, 157, 1-5 [10.1016/j.spl.2019.108630].

Availability:

This version is available at: <https://hdl.handle.net/11585/732057> since: 2020-02-24

Published:

DOI: <http://doi.org/10.1016/j.spl.2019.108630>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Giuliano Galimberti, Gabriele Soffritti. (2020). A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models. With Supplementary material for the paper. *Statistics & Probability Letters*, 157, 108630.

The final published version is available online at:

<https://doi.org/10.1016/j.spl.2019.108630>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models*

Giuliano Galimberti and Gabriele Soffritti[†]

Department of Statistical Sciences, University of Bologna, Italy

September 19, 2019

Abstract

This [letter](#) illustrates simple assumptions [for proving](#) consistency of the maximum likelihood estimator [under](#) multivariate Gaussian [and Student's \$t\$](#) linear cluster-weighted models, which [allow](#) density estimation, clustering and linear regression analysis with continuous random predictors in presence of unobserved heterogeneity.

Keywords: [Mixture model](#); Envelope function; Linear model; Regularity condition

1 Introduction

Cluster-weighted modelling represents a flexible framework for data analysis in which both supervised and unsupervised techniques are exploited. In this framework, the joint distribution of a given random vector is modelled by assuming that this vector is composed of an outcome (response, dependent variable) and its explanatory variables (covariates, predictors). Furthermore, a finite mixture is embedded into the model in order to account for the possible presence of unknown clusters of units. Thus, cluster-weighted models are capable of capturing both observed and unobserved sources of relevant information from a sample, and can be used for density estimation, clustering and/or regression analysis with random covariates in the presence of unobserved heterogeneity. The usefulness and effectiveness of such models are prominent when the sample observations come from several sub-populations, the effect of the covariates on

*Post-print accepted for publication in Statistics and Probability Letters.

[†]Corresponding author. Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy. E-mail address: gabriele.soffritti@unibo.it

the response changes with the sub-populations and the covariates are not under the control of the researcher.

Cluster-weighted models date back to the late 20th century. Gershensfeld (1997) introduces a model for continuous variables and a univariate response, based on Gaussian mixtures. Recently, the research on such models has been intense. In Ingrassia et al. (2012) and Ingrassia et al. (2014), models under both Gaussian and Student's t mixture distributions are studied and compared. The use of t distributions makes the resulting cluster-weighted models more versatile and robust against the possible presence of outliers both in the response and in the covariates. Punzo and Ingrassia (2013), Punzo and Ingrassia (2015) and Ingrassia et al. (2015) develop extensions for dealing with various types of responses. Models with non-linear relationships are described in Punzo (2014). Solutions suitable when there are many covariates can be found in Subedi et al. (2013), Subedi et al. (2015), respectively. Dang et al. (2017) focus on the situation of continuous covariates and a multivariate response through the use of Gaussian mixtures; the resulting model is able to account for correlation among responses. Furthermore, parsimonious specifications of this latter model are also introduced in Dang et al. (2017), where suitable constraints are imposed on the eigen-decomposition of the component-covariance matrices so as to mitigate the problem of a large number of model parameters when dealing with several variables. Robust methods for cluster analysis and regression analysis based on cluster-weighted models are due to Punzo and McNicholas (2017) and García-Escudero et al. (2017), respectively. A package which enables any researcher to fit cluster-weighted models for a univariate response under either the Student's t distribution or some exponential family distributions has been recently developed (Mazza et al., 2018). Methods for the analysis of multilevel data and mixed-support longitudinal data in the framework of cluster-weighted models are due to Berta et al. (2016) and Punzo et al. (2018), respectively.

Similar to any mixture model, cluster-weighted models are affected by some identification issues. As far as the parameter estimation is concerned, the EM algorithm in the maximum likelihood (ML) framework is generally employed (Dempster et al., 1977). The parameter estimation is performed for a fixed number of mixture components; thus, if this number is unknown, different models will have to be estimated and compared in order to detect the one that best fits to a given sample. This latter goal is achieved by resorting to information criteria, such as the Bayesian information criterion (Schwarz, 1978).

Consistency, asymptotic normality and asymptotic efficiency are large sample properties of an estimator useful for constructing approximate confidence intervals and testing hypotheses. Thus, conditions ensuring these properties for specific classes of estimators and/or specific classes of models have been extensively studied and defined in the literature. Frequently, the results on the above mentioned properties are hierarchically nested. Namely, the asymptotic normality results assume consistency and the asymptotic efficiency results assume asymptotic normality. In line with previous studies, this paper illustrates simple assumptions on the parameters of the multivariate Gaussian and Student's t linear cluster-weighted models and shows that such assumptions ensure the

consistency of the ML estimator under the examined models. The proofs of these consistency results exploit some general regularity conditions illustrated in Newey and McFadden (1994).

The paper is organised as follows. Section 2 contains some preliminary definitions concerning the classes of multivariate Gaussian and Student's t linear cluster-weighted models. The assumptions on the parameters of the Gaussian model class are described in Section 3, where a theorem and two lemmas used to prove the consistency of the ML estimator are also reported. Section 4 provides the same information for the Student's t model class. A Supplementary Material with the proofs of two theorems completes the treatment of the subject.

2 Linear cluster-weighted models

Let $\mathbf{Y} = (Y_1, \dots, Y_q)'$ be a $q \times 1$ random vector of absolutely continuous dependent variables and $\mathbf{X} = (X_1, \dots, X_p)'$ be a $p \times 1$ random vector of absolutely continuous random predictors. Furthermore, let $\text{vec}(\mathbf{A})$ be the column vector obtained by stacking the columns of matrix \mathbf{A} one underneath the other, and $\mathbf{v}(\mathbf{B})$ be the column vector obtained from $\text{vec}(\mathbf{B})$ by eliminating all supradiagonal elements of a symmetric matrix \mathbf{B} (thus, $\mathbf{v}(\mathbf{B})$ contains only the distinct elements of \mathbf{B}) (for more details see, e.g., Horn and Johnson, 1990).

Definition 1. *The $(p+q) \times 1$ random vector $(\mathbf{X}', \mathbf{Y}')'$ follows a Gaussian linear cluster-weighted model of order G if its probability density function (p.d.f.) has the form*

$$f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \sum_{g=1}^G \pi_g \phi_p(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \phi_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \boldsymbol{\Gamma}_g), \quad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}, \quad (1)$$

for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\phi_l(\cdot; \boldsymbol{\xi}, \boldsymbol{\Xi})$ denotes the p.d.f. of an l -dimensional Gaussian random vector with expected value $\boldsymbol{\xi}$ and positive definite covariance matrix $\boldsymbol{\Xi}$, $\pi_g > 0 \ \forall g$, $\sum_{g=1}^G \pi_g = 1$, $\boldsymbol{\gamma}_g = \boldsymbol{\lambda}_g + \mathbf{B}_g \mathbf{x}$, with \mathbf{B}_g denoting a $q \times p$ matrix with real elements $\forall g$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{G-1})'$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}'_1, \dots, \boldsymbol{\lambda}'_G)'$, $\mathbf{B} = (\text{vec}(\mathbf{B}_1)', \dots, \text{vec}(\mathbf{B}_G)')'$, $\boldsymbol{\Gamma} = (\mathbf{v}(\boldsymbol{\Gamma}_1)', \dots, \mathbf{v}(\boldsymbol{\Gamma}_G)')'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_G)'$, $\boldsymbol{\Sigma} = (\mathbf{v}(\boldsymbol{\Sigma}_1)', \dots, \mathbf{v}(\boldsymbol{\Sigma}_G)')'$, and $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\lambda}', \mathbf{B}', \boldsymbol{\Gamma}', \boldsymbol{\mu}', \boldsymbol{\Sigma}')'$.

Definition 2. *The $(p+q) \times 1$ random vector $(\mathbf{X}', \mathbf{Y}')'$ follows a Student's t linear cluster-weighted model of order G if its p.d.f. has the form*

$$f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi}) = \sum_{g=1}^G \pi_g h_p(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) h_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \boldsymbol{\Gamma}_g, \kappa_g), \quad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}, \quad (2)$$

for some $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, where $h_l(\cdot; \boldsymbol{\xi}, \boldsymbol{\Xi}, \omega)$ denotes the p.d.f. of an l -dimensional Student's t random vector with location parameter $\boldsymbol{\xi}$, positive definite scatter matrix $\boldsymbol{\Xi}$, and degrees of freedom ω (see equation (A) in the Supplementary Material), $\boldsymbol{\psi} = (\boldsymbol{\theta}', \boldsymbol{\nu}', \boldsymbol{\kappa}')$, $\boldsymbol{\nu} = (\nu_1, \dots, \nu_G)'$, $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_G)'$ and $\boldsymbol{\theta}$ is the vector introduced in Definition 1.

Definition 3. The class of Gaussian linear cluster-weighted models of order G is $\mathfrak{N}_G = \{f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}\}$, where $f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ is defined in equation (1) and $\boldsymbol{\lambda}_g \neq \boldsymbol{\lambda}_j \vee \text{vec}(\mathbf{B}_g) \neq \text{vec}(\mathbf{B}_j) \vee \mathbf{v}(\boldsymbol{\Sigma}_k) \neq \mathbf{v}(\boldsymbol{\Sigma}_j)$ for $g \neq j$.

Definition 4. The class of Student's t linear cluster-weighted models of order G is $\mathfrak{T}_G = \{f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}\}$, where $f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})$ is defined in equation (2) and $\nu_g \neq \nu_j \vee \kappa_g \neq \kappa_j \vee \boldsymbol{\lambda}_g \neq \boldsymbol{\lambda}_j \vee \text{vec}(\mathbf{B}_g) \neq \text{vec}(\mathbf{B}_j) \vee \mathbf{v}(\boldsymbol{\Sigma}_k) \neq \mathbf{v}(\boldsymbol{\Sigma}_j)$ for $g \neq j$.

Dang et al. (2017) provide a sufficient condition for the identifiability of the class \mathfrak{N} in $\Omega \times \mathbb{R}^q$, where $\mathfrak{N} = \{\mathfrak{N}_G, G \in \mathbb{N}\}$, and $\Omega \subseteq \mathbb{R}^p$ has probability equal to one according to the p -variate Gaussian distribution. Their condition is summarized as follows.

(C0) The mixture of regression models

$$\sum_{g=1}^G \alpha_g(\mathbf{x}) \phi_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \boldsymbol{\Gamma}_g), \quad \mathbf{y} \in \mathbb{R}^q,$$

is identifiable for each fixed $\mathbf{x} \in \Omega$, where $\alpha_1(\mathbf{x}), \dots, \alpha_G(\mathbf{x})$ are positive weights summing to one for each $\mathbf{x} \in \Omega$.

In addition, they prove that, when the condition (C0) holds true, the model class \mathfrak{N} is identifiable. As far as the Student's t model class is concerned, this condition can be easily modified so as to ensure the identifiability of the class \mathfrak{T} in $\Omega \times \mathbb{R}^q$, where $\mathfrak{T} = \{\mathfrak{T}_G, G \in \mathbb{N}\}$, and $\Omega \subseteq \mathbb{R}^p$ has probability equal to one according to the p -variate Student's t distribution. The modified condition is defined as follows.

(C0)* The mixture of regression models

$$\sum_{g=1}^G \alpha_g(\mathbf{x}) h_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \boldsymbol{\Gamma}_g, \kappa_g), \quad \mathbf{y} \in \mathbb{R}^q,$$

is identifiable for each fixed $\mathbf{x} \in \Omega$, where $\alpha_1(\mathbf{x}), \dots, \alpha_G(\mathbf{x})$ are positive weights summing to one for each $\mathbf{x} \in \Omega$.

The proof of the identifiability of the model class \mathfrak{T} under the condition (C0)* can be easily obtained from the proof given in Dang et al. (2017) by exploiting the identifiability of finite mixtures of multivariate Student's t distributions (Holzmann et al., 2006).

3 Consistency of the ML estimator under Gaussian models

Let $(\mathbf{x}'_1, \mathbf{y}'_1)', \dots, (\mathbf{x}'_I, \mathbf{y}'_I)'$ be I independent and identically distributed (i.i.d.) sample observations of $(\mathbf{X}', \mathbf{Y}')'$ under the model defined in equation (1), and

let $\hat{\boldsymbol{\theta}}_I$ be the ML estimator of $\boldsymbol{\theta}$ based on these observations. Details about the EM algorithm can be found in Dang et al. (2017). Furthermore, let the true p.d.f. of $(\mathbf{X}', \mathbf{Y}')'$ be denoted as $g(\mathbf{x}, \mathbf{y})$. The consistency of the ML estimator $\hat{\boldsymbol{\theta}}_I$ has been studied for the model class $\mathfrak{N}_G = \{f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}\}$, with $\bar{\boldsymbol{\Theta}}$ denoting a compact metric subspace of $\boldsymbol{\Theta}$ whose elements fulfil the following conditions:

- (C1) $\boldsymbol{\mu}_g \in \mathcal{A}(\epsilon, p) \forall g$, where $\mathcal{A}(\epsilon, r) = \{\mathbf{a} \in \mathbb{R}^r : \|\mathbf{a}\| \leq \epsilon\}$, $0 < \epsilon < \infty$ and $\|\cdot\|$ is the Euclidean norm;
- (C2) $\boldsymbol{\Sigma}_g \in \mathcal{D}_p(a, b) \forall g$, where $\mathcal{D}_p(a, b)$ denotes the set of the $p \times p$ positive definite matrices with eigenvalues in $[a, b]$, with $0 < a < b < \infty$;
- (C3) $\boldsymbol{\lambda}_g \in \mathcal{A}(\eta, q) \forall g$, with $0 < \eta < \infty$;
- (C4) $\mathbf{B}_g \in \mathcal{B}(\rho, q, p) \forall g$, where $\mathcal{B}(\rho, q, p) = \{\mathbf{B} \in \mathcal{M}_{q \times p} : |||\mathbf{B}||| \leq \rho\}$, with $\mathcal{M}_{q \times p}$ denoting the set of $q \times p$ matrices with real elements, $0 < \rho < \infty$ and $|||\cdot|||$ being the following matrix norm:

$$|||\mathbf{B}||| = \sup \{\|\mathbf{B}\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^p \text{ with } \|\mathbf{x}\| = 1\}, \forall \mathbf{B} \in \mathcal{M}_{q \times p}.$$

- (C5) $\boldsymbol{\Gamma}_g \in \mathcal{D}_q(c, d) \forall g$, with $0 < c < d < \infty$.

Furthermore, it is supposed that

- (C6) a unique model $M^0 \in \mathfrak{N}_{G^0}$ exists such that $g(\mathbf{x}, \mathbf{y}) = f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \check{\boldsymbol{\theta}}_{M^0})$ for some parameter value $\check{\boldsymbol{\theta}}_{M^0} \in \bar{\boldsymbol{\Theta}}$, where the order G^0 of model M^0 is known.

For the proof of the consistency of $\hat{\boldsymbol{\theta}}_I$ some preliminary results are required. They are summarized in Theorem 1 and Lemmas 1 and 2. Theorem 1 guarantees the existence and the g -integrability of an envelope function $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$ for the model class \mathfrak{N}_G . Lemma 1 ensures that $\mathbb{E}(\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta})])$ has a unique maximum at $\boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ denotes the true value of the model parameters. Lemma 2 allows to state that $\mathbb{E}(\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta})])$ is continuous, and $\frac{1}{T} \sum_{i=1}^T \ln[f_{\mathcal{N}}(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta})]$ uniformly converges in probability to $\mathbb{E}(\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta})])$. Finally, the consistency of $\hat{\boldsymbol{\theta}}_I$ is given by Corollary 1. Proofs of the two lemmas and the corollary are based on some general theorems that hold true for the extremum estimators of parametric models in the presence of i.i.d. random variables (Newey and McFadden, 1994). The regularity conditions (C1)-(C6) have been defined so as to ensure that the general theorems in Newey and McFadden (1994) can be applied to the ML estimator of the specific model class examined in this paper. [The proof of Theorem 1 is reported in the Supplementary Material.](#)

Theorem 1. *Given the model class $\mathfrak{N}_G = \{f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}\}$ and the conditions (C1)-(C6), there exists a function $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$, $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}$, such that*

$$|\ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})]| \leq e_{\mathcal{N}}(\mathbf{x}, \mathbf{y}) \quad \forall \boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}, \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}, \quad (3)$$

$$\int e_{\mathcal{N}}(\mathbf{x}, \mathbf{y}) g(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} < \infty. \quad (4)$$

Lemma 1. *If conditions (C0)-(C6) hold true and $\theta_0 \in \bar{\Theta}$, then $\mathbb{E}(\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \theta)])$ has a unique maximum at θ_0 .*

Proof. Condition (C0) ensures that θ_0 is identified. Under the conditions (C1)-(C6), from Theorem 1 it follows that

$$\mathbb{E} \{ |\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \theta)]| \} \leq \mathbb{E} [e_{\mathcal{N}}(\mathbf{X}, \mathbf{Y})] \quad \forall \theta \in \bar{\Theta}.$$

Finally, Lemma 2.2 of Newey and McFadden (1994) leads to the result given in Lemma 1. \square

Lemma 2. *If $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_I, \mathbf{y}_I)$ are i.i.d. sample observations of (\mathbf{X}, \mathbf{Y}) under the model defined in equation (1), $\bar{\Theta}$ is compact and the conditions (C1)-(C6) are fulfilled, then $\mathbb{E}(\ln[f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \theta)])$ is continuous and*

$$\sup_{\theta \in \bar{\Theta}} \left| \frac{1}{I} \sum_{i=1}^I \ln[f_{\mathcal{N}}(\mathbf{x}_i, \mathbf{y}_i; \theta)] - \mathbb{E} [\ln f_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}; \theta)] \right| \xrightarrow{p} 0. \quad (5)$$

Proof. The results given in Lemma 2 follow immediately from Theorem 1 and Lemma 2.4 of Newey and McFadden (1994). \square

Corollary 1. *Given conditions (C0)-(C6) and if $\bar{\Theta}$ is compact, then the following convergence in probability holds true:*

$$\hat{\theta}_I \xrightarrow{p} \theta_0. \quad (6)$$

Proof. The result (10) follows immediately from Theorem 1 and Theorem 2.1, Lemmas 2.2 and 2.4 of Newey and McFadden (1994). \square

4 Consistency of the ML estimator under Student's t models

Suppose that $(\mathbf{x}'_1, \mathbf{y}'_1)', \dots, (\mathbf{x}'_I, \mathbf{y}'_I)'$ are i.i.d. sample observations of $(\mathbf{X}', \mathbf{Y})'$ under the model defined in equation (2). Furthermore, let $\hat{\psi}_I$ be the ML estimator of ψ based on these observations. The consistency of $\hat{\psi}_I$ has been proven for the model class $\mathfrak{T}_G = \{f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \psi), \psi \in \bar{\Psi}\}$, where $\bar{\Psi}$ is a compact metric subspace of Ψ whose elements fulfil the conditions (C1)-(C5) illustrated in Section 3 and, in addition, the following conditions:

(C7) $\nu \in \mathcal{F}(\delta, \tau, G)$, $\kappa \in \mathcal{F}(\delta, \tau, G)$, where $\mathcal{F}(\delta, \tau, r) = \{\mathbf{a} = (a_1, \dots, a_k, \dots, a_r)' \in \mathbb{R}^r : 2 + \delta \leq a_k \leq \tau, k = 1, \dots, r\}$, $0 < \delta < \infty$, $2 + \delta < \tau < \infty$;

(C8) a unique model $M^0 \in \mathfrak{T}_{G^0}$ exists such that $g(\mathbf{x}, \mathbf{y}) = f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \check{\psi}_{M^0})$ for some parameter value $\check{\psi}_{M^0} \in \bar{\Psi}$, where the order G^0 of model M^0 is known.

The theoretical result concerning the consistency of $\hat{\psi}_I$, which is given in Corollary 2, is based on the preliminary results summarised in Theorem 2 and Lemmas 3-4. A proof of Theorem 2 can be found in the Supplementary Material. The proofs of Lemma 4 and Corollary 2 are similar to the ones provided in Section 3 for Lemma 2 and Corollary 1 and, thus, are omitted.

Theorem 2. *Given the model class $\mathfrak{T}_G = \{f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \psi), \psi \in \bar{\Psi}\}$ and the conditions (C1)-(C5) and (C7)-(C8), there exists a function $e_{\mathcal{T}}(\mathbf{x}, \mathbf{y})$, $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}$, such that*

$$|\ln[f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \psi)]| \leq e_{\mathcal{T}}(\mathbf{x}, \mathbf{y}) \quad \forall \psi \in \bar{\Psi}, \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}, \quad (7)$$

$$\int e_{\mathcal{T}}(\mathbf{x}, \mathbf{y}) g(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} < \infty. \quad (8)$$

Lemma 3. *If conditions (C0)*-(C5) and (C7)-(C8) hold true and $\psi_0 \in \bar{\Psi}$, then $\mathbb{E}(\ln[f_{\mathcal{T}}(\mathbf{X}, \mathbf{Y}; \psi)])$ has a unique maximum at ψ_0 .*

Proof. Condition (C0)* ensures that ψ_0 is identified. Under the conditions (C1)-(C5) and (C7)-(C8), from Theorem 2 it follows that

$$\mathbb{E}\{|\ln[f_{\mathcal{T}}(\mathbf{X}, \mathbf{Y}; \psi)]|\} \leq \mathbb{E}[e_{\mathcal{T}}(\mathbf{X}, \mathbf{Y})] \quad \forall \psi \in \bar{\Psi}.$$

Finally, Lemma 2.2 of Newey and McFadden (1994) leads to the result given in Lemma 1. \square

Lemma 4. *If $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_I, \mathbf{y}_I)$ are i.i.d. sample observations of (\mathbf{X}, \mathbf{Y}) under the model defined in equation (2), $\bar{\Psi}$ is compact and the conditions (C1)-(C5) and (C7)-(C8) are fulfilled, then $\mathbb{E}(\ln[f_{\mathcal{T}}(\mathbf{X}, \mathbf{Y}; \psi)])$ is continuous and*

$$\sup_{\psi \in \bar{\Psi}} \left| \frac{1}{I} \sum_{i=1}^I \ln[f_{\mathcal{T}}(\mathbf{x}_i, \mathbf{y}_i; \psi)] - \mathbb{E}[\ln f_{\mathcal{T}}(\mathbf{X}, \mathbf{Y}; \psi)] \right| \xrightarrow{p} 0. \quad (9)$$

Corollary 2. *Given the conditions (C0)*-(C5) and (C7)-(C8) and if $\bar{\Psi}$ is compact, then the following convergence in probability holds true:*

$$\hat{\psi}_I \xrightarrow{p} \psi_0. \quad (10)$$

References

- Berta, P., Ingrassia, S., Punzo, A. and Vittadini, G. (2016). Cluster-weighted multilevel models for the evaluation of hospitals. *Metron* **74** 275–292.
- Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S. and Browne, R. P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *J. Classif.* **34** 4–34.

- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39** 1–22.
- García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S. and Mayo-Iscar, A. (2017). Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Stat. Comput.* **27** 377–402.
- Gershensfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences* **808** 18–24.
- Holzmann, H., Munk, A. and Tilmann, G. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Stat.* **33** 753–763.
- Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press, Cambridge.
- Ingrassia, S., Minotti, S. C. and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *J. Classif.* **29** 363–401.
- Ingrassia, S., Minotti, S. C. and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Comput. Stat. Data Anal.* **71** 159–182.
- Ingrassia, S., Punzo, A., Vittadini, G. and Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *J. Classif.* **32** 85–113.
- Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics* **65** 701–709.
- Mazza, A., Punzo, A. and Ingrassia, S. (2018). flexCWM: a flexible framework for cluster-weighted models. *J. Stat. Softw.* **86**(2) 1–30.
- Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, Volume 4, Chapter 36, 2111–2245. Elsevier.
- Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. *Stat. Model.* **14** 257–291.
- Punzo, A. and Ingrassia, S. (2015). On the use of the generalized linear exponential cluster-weighted model to assess local linear independence in bivariate data. *QdS - Journal of Methodological and Applied Statistics* **15** 131–144.
- Punzo, A. and Ingrassia, S. (2015). Clustering bivariate mixed-type data via the cluster-weighted model. *Comp. Stat.* **31** 989–1013.
- Punzo, A. and McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *J. Classif.* **34** 249–293.

- Punzo, A., Ingrassia, S. and Maruotti, A. (2018). Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Stat. Med.* **37** 2797–2808.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
- Subedi, S., Punzo, A., Ingrassia, S. and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Adv. Data Anal. Classif.* **7** 5–40.
- Subedi, S., Punzo, A., Ingrassia, S. and McNicholas, P. D. (2015). Cluster-weighted t -factor analyzers for robust model-based clustering and dimension reduction. *Stat. Methods Appl.* **24** 623–649.

Supplementary material for the paper: A note on the consistency of the maximum likelihood estimator under multivariate Gaussian linear cluster-weighted models*

Giuliano Galimberti and Gabriele Soffritti[†]

Department of Statistical Sciences, University of Bologna, Italy

September 19, 2019

A Some results about Student's t distributions and the gamma function

The p.d.f. of an l -dimensional Student's t random vector \mathbf{Z} is given by

$$h_l(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Xi}, \omega) = \frac{\Gamma\left(\frac{\omega+l}{2}\right)}{\Gamma\left(\frac{\omega}{2}\right) \omega^{\frac{l}{2}} \pi^{\frac{l}{2}} |\boldsymbol{\Xi}|^{\frac{1}{2}}} \left(1 + \frac{1}{\omega} \|\mathbf{z} - \boldsymbol{\xi}\|_{\boldsymbol{\Xi}}^2\right)^{-\frac{\omega+l}{2}}, \quad \mathbf{z} \in \mathbb{R}^l, \quad (\text{A})$$

where $\boldsymbol{\xi}$ is a location vector, $\boldsymbol{\Xi}$ is a positive definite scatter matrix, ω represents the degrees of freedom and $\|\mathbf{z} - \boldsymbol{\xi}\|_{\boldsymbol{\Xi}}^2 = (\mathbf{z} - \boldsymbol{\xi})' \boldsymbol{\Xi}^{-1} (\mathbf{z} - \boldsymbol{\xi})$.

The expected value of \mathbf{Z} is given by $\boldsymbol{\xi}$, provided that $\omega > 1$. Furthermore, if $\omega > 2$, then the covariance matrix of \mathbf{Z} is equal to $\frac{\omega}{\omega-2} \boldsymbol{\Xi}$ (see, for example, Kots and Nadarajah, 2004). Finally, according to Rong et al. (2012),

$$\mathbb{E}[\mathbf{Z}'\mathbf{Z}] = \int \|\mathbf{z}\|^2 h_l(\mathbf{z}; \boldsymbol{\xi}, \boldsymbol{\Xi}, \omega) d\mathbf{z} = \|\boldsymbol{\xi}\|^2 + \frac{\omega}{\omega-2} \text{tr}(\boldsymbol{\Xi}). \quad (\text{B})$$

The p.d.f. in equation (A) involves the gamma function $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} dt$. It is easy to show that, if $2 + \delta \leq \omega \leq \tau$ ($0 < \delta < +\infty$, $2 + \delta < \tau < \infty$) and $l \geq 1$, then

$$0 < \Gamma\left(\frac{3+\delta}{2}\right) \leq \Gamma\left(\frac{\omega+l}{2}\right) \leq \Gamma\left(\frac{\tau+l}{2}\right) < +\infty, \quad (\text{C})$$

*The authors have no competing interests to declare.

[†]Corresponding author. Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy. E-mail address: gabriele.soffritti@unibo.it

$$0 < \Gamma(c^*) \leq \Gamma\left(\frac{\omega}{2}\right) \leq \max\left\{1, \Gamma\left(\frac{\tau}{2}\right)\right\} < +\infty, \quad (\text{D})$$

where $c^* \cong 1.46163$ (see, for example, Deming and Colcord, 1935; Davis, 1970).

Taking into account inequalities (C) and (D) and considering that, for any positive definite matrix Ξ and any $2 + \delta \leq \omega \leq \tau$,

$$\left(1 + \frac{1}{\omega} \|\mathbf{z} - \boldsymbol{\xi}\|_{\Xi}^2\right)^{-\frac{\omega+l}{2}} \leq 1 \quad \forall \mathbf{z}$$

and

$$-\frac{\omega+l}{2} \ln\left(1 + \frac{1}{\omega} \|\mathbf{z} - \boldsymbol{\xi}\|_{\Xi}^2\right) \geq -\frac{\omega+l}{2\omega} \|\mathbf{z} - \boldsymbol{\xi}\|_{\Xi}^2 \geq -\frac{\tau+l}{4+2\delta} \|\mathbf{z} - \boldsymbol{\xi}\|_{\Xi}^2 \quad \forall \mathbf{z},$$

it is easy to show that

$$\ln[h_l(\mathbf{z}; \boldsymbol{\xi}, \Xi, \omega)] \leq \ln \Gamma\left(\frac{\tau+l}{2}\right) - \ln \Gamma(c^*) - \frac{l}{2} \ln[\pi(2+\delta)] - \frac{1}{2} \ln |\Xi| \quad (\text{E})$$

and

$$\ln[h_l(\mathbf{z}; \boldsymbol{\xi}, \Xi, \omega)] \geq \ln \Gamma\left(\frac{3+\delta}{2}\right) - \max\left\{0, \ln \Gamma\left(\frac{\tau}{2}\right)\right\} - \frac{l}{2} \ln \tau \pi - \frac{1}{2} \ln |\Xi| - \frac{\tau+l}{4+2\delta} \|\mathbf{z} - \boldsymbol{\xi}\|_{\Xi}^2. \quad (\text{F})$$

B Proof of Theorem 1

The proof is composed of two parts: (i) the explicit expression of the envelope function $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$ is derived (first result); (ii) $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$ is shown to be a g -integrable function (second result). Both parts exploit arguments similar to the ones used in Maugis et al. (2007).

Let $\|\mathbf{x} - \boldsymbol{\mu}_g\|_{\Sigma_g}^2 = (\mathbf{x} - \boldsymbol{\mu}_g)' \Sigma_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)$, and $\|\mathbf{y} - \boldsymbol{\gamma}_g\|_{\Gamma_g}^2 = (\mathbf{y} - \boldsymbol{\gamma}_g)' \Gamma_g^{-1} (\mathbf{y} - \boldsymbol{\gamma}_g)$. Since Σ_g and Γ_g are positive definite, $\|\mathbf{x} - \boldsymbol{\mu}_g\|_{\Sigma_g}^2 \geq 0 \quad \forall \mathbf{x}$, and $\|\mathbf{y} - \boldsymbol{\gamma}_g\|_{\Gamma_g}^2 \geq 0 \quad \forall \mathbf{y}$. Furthermore, $|\Sigma_g|^{-\frac{1}{2}} \leq a^{-\frac{p}{2}}$ and $|\Gamma_g|^{-\frac{1}{2}} \leq c^{-\frac{q}{2}}$, where a and c denote the lower bound for the eigenvalues of Σ_g and Γ_g , respectively, $\forall g$ (see Maugis et al., 2007, lemma 3). Then,

$$\begin{aligned} \ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})] &= \ln \left[\sum_{g=1}^G \pi_g \phi_p(\mathbf{x}; \boldsymbol{\mu}_g, \Sigma_g) \phi_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \Gamma_g) \right] \\ &= \ln \left[\sum_{g=1}^G \pi_g |2\pi \Sigma_g|^{-\frac{1}{2}} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_g\|_{\Sigma_g}^2}{2}\right) |2\pi \Gamma_g|^{-\frac{1}{2}} \exp\left(-\frac{\|\mathbf{y} - \boldsymbol{\gamma}_g\|_{\Gamma_g}^2}{2}\right) \right] \\ &\leq \ln \left[\sum_{g=1}^G \pi_g (2\pi a)^{-\frac{p}{2}} (2\pi c)^{-\frac{q}{2}} \right] \end{aligned} \quad (\text{G})$$

$$\leq -\frac{p}{2} \ln(2\pi a) - \frac{q}{2} \ln(2\pi c), \quad (\text{H})$$

where the last inequality is obtained by exploiting the result $\sum_{k=1}^K \pi_k = 1$. Thus, $U_{\mathcal{N}} = -\frac{p}{2} \ln(2\pi a) - \frac{q}{2} \ln(2\pi c)$ is an upper bound of $\ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})] \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}, \quad \forall \boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}$.

The lower bound of this function can be obtained as follows. Using the concavity of the logarithm function it is possible to write:

$$\begin{aligned} \ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})] &\geq \sum_{g=1}^G \pi_g \ln \left[(2\pi)^{-\frac{p+q}{2}} |\boldsymbol{\Sigma}_g|^{-\frac{1}{2}} \exp \left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_g\|_{\boldsymbol{\Sigma}_g}^2}{2} \right) |\boldsymbol{\Gamma}_g|^{-\frac{1}{2}} \exp \left(-\frac{\|\mathbf{y} - \boldsymbol{\gamma}_g\|_{\boldsymbol{\Gamma}_g}^2}{2} \right) \right] \\ &= -\frac{p+q}{2} \ln(2\pi) - \frac{1}{2} \sum_{g=1}^G \pi_g \left[\ln(|\boldsymbol{\Sigma}_g|) + \|\mathbf{x} - \boldsymbol{\mu}_g\|_{\boldsymbol{\Sigma}_g}^2 + \ln(|\boldsymbol{\Gamma}_g|) + \|\mathbf{y} - \boldsymbol{\gamma}_g\|_{\boldsymbol{\Gamma}_g}^2 \right]. \quad (\text{I}) \end{aligned}$$

Furthermore, recalling that $\boldsymbol{\gamma}_g = \boldsymbol{\lambda}_g + \mathbf{B}_g \mathbf{x}$, it is possible to write

$$\|\mathbf{y} - \boldsymbol{\lambda}_g - \mathbf{B}_g \mathbf{x}\|_{\boldsymbol{\Gamma}_g}^2 \leq \frac{1}{c} \|\mathbf{y} - (\boldsymbol{\lambda}_g + \mathbf{B}_g \mathbf{x})\|^2 \quad (\text{J})$$

$$\leq \frac{2}{c} (\|\mathbf{y}\|^2 + \|\boldsymbol{\lambda}_g + \mathbf{B}_g \mathbf{x}\|^2) \quad (\text{K})$$

$$\leq \frac{2}{c} \|\mathbf{y}\|^2 + \frac{4}{c} (\|\boldsymbol{\lambda}_g\|^2 + \|\mathbf{B}_g \mathbf{x}\|^2) \quad (\text{L})$$

$$\leq \frac{2}{c} \|\mathbf{y}\|^2 + \frac{4}{c} \|\boldsymbol{\lambda}_g\|^2 + \frac{4}{c} \|\mathbf{B}_g\|^2 \|\mathbf{x}\|^2 \quad (\text{M})$$

$$\leq \frac{2}{c} \|\mathbf{y}\|^2 + \frac{4}{c} \eta^2 + \frac{4}{c} \rho^2 \|\mathbf{x}\|^2 \quad (\text{N})$$

where the inequality (J) is a consequence of the lemma 3 in Maugis et al. (2007); inequalities (K) and (L) are obtained from the parallelogram identity (see, e.g. Horn and Johnson, 1990, p. 263); inequalities (M) and (N) hold because of the conditions (C3)-(C4) on $\boldsymbol{\lambda}_g$ and \mathbf{B}_g .

In a similar way, by exploiting the condition (C1) on $\boldsymbol{\mu}_g$ it is possible to write

$$\begin{aligned} \|\mathbf{x} - \boldsymbol{\mu}_g\|_{\boldsymbol{\Sigma}_g}^2 &\leq \frac{1}{a} \|\mathbf{x} - \boldsymbol{\mu}_g\|^2 \\ &\leq \frac{2}{a} (\|\mathbf{x}\|^2 + \|\boldsymbol{\mu}_g\|^2) \\ &\leq \frac{2}{a} \|\mathbf{x}\|^2 + \frac{2}{a} \epsilon^2. \quad (\text{O}) \end{aligned}$$

Using lemma 3 in Maugis et al. (2007) and combining the results given in equations (I), (N) and (O) leads to the following lower bound for $\ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})]$:

$$L_{\mathcal{N}} = -\frac{p+q}{2} \ln(2\pi) - \frac{p \ln(b)}{2} - \frac{q \ln(d)}{2} - \frac{4}{c} \eta^2 - \frac{2}{a} \epsilon^2 - \frac{2}{c} \|\mathbf{y}\|^2 - \left(\frac{2}{a} + \frac{4}{c} \rho^2 \right) \|\mathbf{x}\|^2. \quad (\text{P})$$

Thus, $\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+q}$ and $\forall \boldsymbol{\theta} \in \bar{\boldsymbol{\Theta}}$, it results that $L_{\mathcal{N}} \leq \ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})] \leq U_{\mathcal{N}}$, $\forall f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \in \mathfrak{F}_G$. As a consequence of above, it is possible to write

$$|\ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})]| \leq C_1(a, b, c, d, \epsilon, \eta, p, q) + C_2(c) \|\mathbf{y}\|^2 + C_3(a, c, \rho) \|\mathbf{x}\|^2, \quad (\text{Q})$$

defining the envelope function $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$, where $C_1(\cdot)$, $C_2(\cdot)$ and $C_3(\cdot)$ are positive constants. This concludes the proof of the first result.

The g -integrability of the envelope function $e_{\mathcal{N}}(\mathbf{x}, \mathbf{y})$ can be proved by showing that $\int \|\mathbf{x}, \mathbf{y}\|^2 g(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} < \infty$. By exploiting the condition (C6), it is possible to write

$$\begin{aligned}
\int \|\mathbf{x}, \mathbf{y}\|^2 g(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} &= \int (\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2) f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \check{\boldsymbol{\theta}}_{M_0}) d\mathbf{x} d\mathbf{y} \\
&= \int (\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2) \sum_{g=1}^{G^0} \check{\pi}_g \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) \phi_q(\mathbf{y}|\mathbf{x}; \check{\boldsymbol{\gamma}}_g, \check{\boldsymbol{\Gamma}}_g) d\mathbf{y} d\mathbf{x} \\
&= \sum_{g=1}^{G^0} \check{\pi}_g \int \|\mathbf{x}\|^2 \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \\
&\quad + \sum_{g=1}^{G^0} \check{\pi}_g \int \|\mathbf{y}\|^2 \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) \phi_q(\mathbf{y}|\mathbf{x}; \check{\boldsymbol{\gamma}}_g, \check{\boldsymbol{\Gamma}}_g) d\mathbf{y} d\mathbf{x}. \tag{R}
\end{aligned}$$

As far as the first term in the right part of the equation (R) is concerned, using lemmas 3 and 4 in Maugis et al. (2007) it results that

$$\begin{aligned}
\sum_{g=1}^{G^0} \check{\pi}_g \int \|\mathbf{x}\|^2 \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} &\leq 2 \sum_{g=1}^{G^0} \check{\pi}_g \left[\|\check{\boldsymbol{\mu}}_g\|^2 + \text{tr}(\check{\boldsymbol{\Sigma}}_g) \right] \\
&\leq 2 \sum_{g=1}^{G^0} \check{\pi}_g (\epsilon^2 + bp) \\
&\leq 2 (\epsilon^2 + bp). \tag{S}
\end{aligned}$$

As far as the second term in the right part of the equation (R) is concerned, let the integral within such term be denoted as A_g . For this quantity it is possible to write:

$$\begin{aligned}
A_g &= \int \left[\|\mathbf{y}\|^2 \phi_q(\mathbf{y}|\mathbf{x}; \check{\boldsymbol{\gamma}}_g, \check{\boldsymbol{\Gamma}}_g) d\mathbf{y} \right] \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \\
&\leq 2 \int \left[\|\check{\boldsymbol{\gamma}}_g\|^2 + \text{tr}(\check{\boldsymbol{\Gamma}}_g) \right] \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \tag{T}
\end{aligned}$$

$$\leq 2 \int \left[2\|\check{\boldsymbol{\lambda}}_g\|^2 + 2\|\check{\mathbf{B}}_g \mathbf{x}\|^2 + dq \right] \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \tag{U}$$

$$\leq 2 \int \left[2\eta^2 + 2\rho^2 \|\mathbf{x}\|^2 + dq \right] \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \tag{V}$$

$$\leq 4\eta^2 + 2dq + 4\rho^2 \int \|\mathbf{x}\|^2 \phi_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g) d\mathbf{x} \tag{W}$$

$$\leq 4\eta^2 + 2dq + 8\epsilon^2 \rho^2 + 8bp\rho^2, \tag{X}$$

where inequalities (T) and (U) are obtained using lemmas 3 and 4 given in Maugis et al. (2007), and the final inequality exploits the result obtained in equation (S).

Combining this result with the equation (R) makes it possible to write

$$\int \|\mathbf{x}, \mathbf{y}\|^2 g(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \leq 2\epsilon^2(1 + 4\rho^2) + 2bp(1 + 4\rho^2) + 4\eta^2 + 2dq < \infty.$$

This concludes the proof of the second result.

C Proof of Theorem 2

The proof of Theorem 2 is essentially similar to the one of Theorem 1. In particular, the explicit expression of the envelope function $e_{\mathcal{T}}(\mathbf{x}, \mathbf{y})$ can be obtained by a straightforward modification of the first part of the proof of Theorem 1. This modification exploits inequalities (E) and (F). Namely, by applying inequality (E) to any $h_p(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ and $h_q(\mathbf{y}|\mathbf{x}; \boldsymbol{\gamma}_g, \boldsymbol{\Gamma}_g, \kappa_g)$ ($g = 1, \dots, G$) and following arguments similar to those leading to the upper bounds for $\ln[f_{\mathcal{N}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})]$ given by equation (H), the following upper bound for $\ln[f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})]$ can be obtained:

$$U_{\mathcal{T}} = \ln \Gamma \left(\frac{\tau + p}{2} \right) + \ln \Gamma \left(\frac{\tau + q}{2} \right) - 2 \ln \Gamma(c^*) - \frac{p}{2} \ln [\pi a(2 + \delta)] - \frac{q}{2} \ln [\pi c(2 + \delta)].$$

Similarly, by exploiting inequality (F) to adapt equation (I) and, consequently, equation (P), it is possible to show that

$$\begin{aligned} L_{\mathcal{T}} = & 2 \ln \Gamma \left(\frac{3 + \delta}{2} \right) - \max \left\{ 0, 2 \ln \Gamma \left(\frac{\tau}{2} \right) \right\} - \frac{p}{2} \ln(\tau \pi b) - \frac{q}{2} \ln(\tau \pi d) - \frac{4(\tau + q)}{c(4 + 2\delta)} \eta^2 + \\ & - \frac{2(\tau + p)}{a(4 + 2\delta)} \epsilon^2 - \frac{2(\tau + q)}{c(4 + 2\delta)} \|\mathbf{y}\|^2 - \left(\frac{2(\tau + p)}{a(4 + 2\delta)} + \frac{4(\tau + q)}{c(4 + 2\delta)} \rho^2 \right) \|\mathbf{x}\|^2 \end{aligned}$$

is a lower bound $\ln[f_{\mathcal{T}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})]$. Finally, $U_{\mathcal{T}}$ and $L_{\mathcal{T}}$ can be combined to derive an expression for $e_{\mathcal{T}}(\mathbf{x}, \mathbf{y})$, that has a structure similar to equation (Q), with non-negative constants $C_1(\cdot)$, $C_2(\cdot)$ and $C_3(\cdot)$ now depending also on τ and δ .

As for Theorem 1, the second part of the proof of Theorem 2 concerns the g -integrability of $e_{\mathcal{T}}(\mathbf{x}, \mathbf{y})$. This requires only minor changes in equations (R) to (X), by using equation (B) instead of lemma 4 in Maugis et al. (2007). In particular, as condition (C7) implies

$$\frac{\omega}{\omega - 2} \leq \frac{2 + \delta}{\delta} \quad \forall \omega \geq 2 + \delta,$$

inequality (S) can be replaced by

$$\sum_{g=1}^{G^0} \tilde{\pi}_g \int \|\mathbf{x}\|^2 h_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g, \check{\nu}_g) d\mathbf{x} \leq 2 \left(\epsilon^2 + \frac{2 + \delta}{\delta} bp \right) < +\infty,$$

while inequality (X) becomes

$$\begin{aligned} \int \left[\|\mathbf{y}\|^2 h_q(\mathbf{y}|\mathbf{x}; \check{\boldsymbol{\gamma}}_g, \check{\boldsymbol{\Gamma}}_g, \check{\kappa}_g) d\mathbf{y} \right] h_p(\mathbf{x}; \check{\boldsymbol{\mu}}_g, \check{\boldsymbol{\Sigma}}_g, \check{\nu}_g) d\mathbf{x} \leq & 4\eta^2 + \frac{(4 + 2\delta)dq}{\delta} + \\ & + 8\epsilon^2 \rho^2 + \frac{(16 + 8\delta)bp}{\delta} \rho^2 < +\infty. \end{aligned}$$

References

- Davis P. J. (1970). Gamma function and related functions. In: Abramowitz, M. and Stegun, I. (eds): *Handbook of Mathematical Functions*. Dover Publications, New York.
- Deming, W. E. and Colcord C. G. (1935). The minimum in the gamma function. *Nature*. **135**, 917.
- Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press, Cambridge.
- Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and their Applications*. Cambridge University Press, NewYork.
- Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2007). Variable selection for clustering with Gaussian mixture models. *Technical Report RR-6211*. Inria, France.
- Rong, J.-Y., Lu, Z.-F. and Liu X.-Q. (2012). On quadratic forms of multivariate t distribution with applications. *Communications in Statistics - Theory and Methods*. **41**, 300–308.