entropy

MDPI

Article

Model Selection in a Composite Likelihood Framework Based on Density Power Divergence

Elena Castilla [1,*], Nirian Martín [2], Leandro Pardo [1] and Konstantinos Zografos [3]

[1] Interdisciplinary Mathematics Institute and Department of Statistics and O.R. I, Complutense University of Madrid, 28040 Madrid, Spain; lpardo@mat.ucm.es

[2] Interdisciplinary Mathematics Institute and Department of Financial and Actuarial Economics & Statistics, Complutense University of Madrid, 28003 Madrid, Spain; nirian@estad.ucm.es

[3] Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece; kzograf@uoi.gr

* Correspondence: elecasti@ucm.es

Received: 22 January 2020; Accepted: 25 February 2020; Published: 27 February 2020

Abstract: This paper presents a model selection criterion in a composite likelihood framework based on density power divergence measures and in the composite minimum density power divergence estimators, which depends on an tuning parameter α. After introducing such a criterion, some asymptotic properties are established. We present a simulation study and two numerical examples in order to point out the robustness properties of the introduced model selection criterion.

Keywords: composite likelihood; composite minimum density power divergence estimators; model selection

1. Introduction

Composite likelihood inference is an important approach to deal with those real situations of large data sets or very complex models, in which classical likelihood methods are computationally difficult, or even, not possible to manage. Composite likelihood methods have been successfully used in many applications concerning, for example, genetics ([1]), generalized linear mixed models ([2]), spatial statistics ([3–5]), frailty models ([6]), multivariate survival analysis ([7,8]), etc.

Let us introduce the problem, adopting here the notation by [9]. Let $\{f(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, p \geq 1\}$ be a parametric identifiable family of distributions for an observation $\boldsymbol{y} = (y_1, ..., y_m)^T$, a realization of a random m-vector $\boldsymbol{Y}$. In this setting, the composite likelihood function based on K different marginal or conditional distributions has the form

$$\mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{k=1}^{K} \left(f_{A_k}(y_j, j \in A_k; \boldsymbol{\theta})\right)^{w_k}$$

and the corresponding composite log-density

$$\log \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{k=1}^{K} w_k \ell_{A_k}(\boldsymbol{\theta}, \boldsymbol{y}), \tag{1}$$

with $\ell_{A_k}(\boldsymbol{\theta}, \boldsymbol{y}) = \log f_{A_k}(y_j, j \in A_k; \boldsymbol{\theta})$, where $\{A_k\}_{k=1}^{K}$ is a family of sets of indices associated either with marginal or conditional distributions involving some y_j, $j \in \{1, ..., m\}$ and w_k, $k = 1, ..., K$ are non-negative and known weights. If the weights are all equal, then they can be ignored. In this case, all the statistical procedures give equivalent results. The composite maximum likelihood estimator (CMLE), $\widehat{\boldsymbol{\theta}}_c$, is obtained by maximizing, in respect to $\boldsymbol{\theta} \in \Theta$, the expression (1).

The CMLE is consistent and asymptotically normal and, based on it, we can establish hypothesis testing procedures in a similar way to the classical likelihood ratio test, Wald test or Rao's score test. A development of the asymptotic theory of the CMLE including its application to obtain the composite ratio statistics, Wald-type tests and Rao score tests in the context of composite likelihood can be seen in [10]. However, in [11–13] is shown that the CMLE and the derived testing procedures present an important lack of robustness. In this sense, [11–13] derived some new distance-based estimators and tests with good robustness behaviour without an important loss of efficiency. In this paper, we are going to consider the composite minimum density power divergence estimator (CMDPDE), introduced in [12], in order to present a model selection criterion in a composite likelihood framework.

Model selection criteria, for summarizing data evidence in favor of a model, is a very well studied subject in statistical literature, overall in the context of full likelihood. The construction of such criteria requires a measure of similarity between two models, which are typically described in terms of their distributions. This can be achieved if an unbiased estimator of the expected overall discrepancy is found, which measures the statistical distance between the true, but unknown model, and the entertained model. Therefore, the model with the smallest value of the criterion is the most preferable model. The use of divergence measures, in particular Kullback–Leibler divergence ([14]), to measure this discrepancy, is the main idea of some of the most known criteria: Akaike Information Criterion (AIC, [15,16]), the criterion proposed by Takeuchi (TIC, [17]) and other modifications of AIC [18]. DIC criterion, based on the density power divergence (DPD), was presented in [19] and, recently, [20] presented a local BHHJ power divergence information criterion following [21]. In the context of the composite likelihood there are some criteria based on Kullback–Leibler divergence, see for instance [22–24] and references therein. To the best of our knowledge only Kullback–Leibler divergence was used to develop model selection criteria in a composite likelihood framework. To fill this gap, our interest is now focused on DPD.

In this paper, we present a new information criterion for model selection in the framework of composite likelihood based on DPD measure. This divergence measure, introduced and studied in the case of complete likelihood by [25], has been considered previously in [12,13] in the context of composite likelihood. In these papers, a new estimator, the CMDPDE, was introduced and its robustness in relation to the CMLE as well as the robustness of some families of test statistics were studied, but the problem of model selection was not considered. This problem is considered in this paper. The criterion introduced in this paper will be called composite likelihood DIC criterion (CLDIC). The motivation of considering a criterion based on DPD instead of Kullback–Leibler divergence is due to the robustness of the procedures based on DPD in statistical inference, not only in the context of full likelihood [25,26], but also in the context of composite likelihood [12,13]. In Section 2, the CMDPDE is presented and some properties of this estimator are discussed. The new model selection criterion, CLDIC, based on CMDPDE is introduced in Section 3 and some of its asymptotic properties are studied. A simulation study is carried out in Section 4 and some numerical examples are presented in Section 5. Finally, some concluding remarks are presented in Section 6.

2. Composite Minimum Density Power Divergence Estimator

Given two probability density functions g and f, associated with two m-dimensional random variables respectively, the DPD ([25]) measures a statistical distance between g and f by

$$d_\alpha(g,f) = \int_{\mathbb{R}^m} \left\{ f(\boldsymbol{y})^{1+\alpha} - \left(1+\frac{1}{\alpha}\right) f(\boldsymbol{y})^\alpha g(\boldsymbol{y}) + \frac{1}{\alpha}\, g(\boldsymbol{y})^{1+\alpha} \right\} d\boldsymbol{y}, \tag{2}$$

for $\alpha > 0$, while for $\alpha = 0$ it is defined by

$$d_0(g,f) = \lim_{\alpha \to 0^+} d_\alpha(g,f) = d_{KL}(g,f),$$

where $d_{KL}(g, f)$ is the Kullback–Leibler divergence (see, for example, [26]). For $\alpha = 1$, the expression (2) leads to the L_2 distance $L_2(g, f) = \int_{\mathbb{R}^m} (f(\boldsymbol{y}) - g(\boldsymbol{y}))^2 \, d\boldsymbol{y}$. It is also interesting to note that (2) is a special case of the so-called Bregman divergence

$$\int_{\mathbb{R}^m} \left[T(g(\boldsymbol{y})) - T(f(\boldsymbol{y})) - \{g(\boldsymbol{y}) - f(\boldsymbol{y}\} T'(f(\boldsymbol{y})) \right] d\boldsymbol{y}. \tag{3}$$

If we consider $T(l) = \frac{1}{\alpha} l^{1+\alpha}$ in (3), we get $d_\alpha(g, f)$. The parameter α controls the trade-off between robustness and asymptotic efficiency of the parameter estimates which are the minimizers of this family of divergences. For more details about this family of divergence measures we refer to [27].

Let now $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be independent and identically distributed replications of $\boldsymbol{Y}$ which are characterized by the true but unknown distribution g. Taking into account that the true model g is unknown, suppose that $\Xi = \{f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, p \geq 1\}$ is a parametric identifiable family of candidate distributions to describe the observations $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$. Then, the DPD between the true model g and the composite likelihood function, $\mathcal{CL}(\boldsymbol{\theta}, \cdot)$, associated to the parametric model $f(\cdot; \boldsymbol{\theta})$ is defined as

$$d_\alpha(g(\cdot), \mathcal{CL}(\boldsymbol{\theta}, \cdot)) = \int_{\mathbb{R}^m} \left\{ \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^\alpha g(\boldsymbol{y}) + \frac{1}{\alpha} g(\boldsymbol{y})^{1+\alpha} \right\} d\boldsymbol{y}, \tag{4}$$

for $\alpha > 0$, while for $\alpha = 0$ we have $d_{KL}(g(\cdot), \mathcal{CL}(\boldsymbol{\theta}, \cdot))$, which is defined by

$$d_{KL}(g(\cdot), \mathcal{CL}(\boldsymbol{\theta}, \cdot)) = \int_{\mathbb{R}^m} g(\boldsymbol{y}) \log \frac{g(\boldsymbol{y})}{\mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})} d\boldsymbol{y}. \tag{5}$$

In Section 3, we are going to introduce and study the CLDIC criterion based on (4).

Let

$$\{M_k\}_{k \in \{1,...,\ell\}} \tag{6}$$

be a family of candidate models to govern the observations $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$. We shall assume that the true model is included in $\{M_k\}_{k \in \{1,...,\ell\}}$. For a specific $k = 1, \ldots, \ell$, the parametric model M_k is described by the composite likelihood function

$$\mathcal{CL}(\boldsymbol{\theta}, \cdot), \quad \boldsymbol{\theta} \in \Theta_k \subset \mathbb{R}^k.$$

In this setting, it is quite clear that the most suitable candidate model to describe the observations is the model that minimizes the DPD in (4). However, the unknown parameter $\boldsymbol{\theta}$ is included in it, so it is not possible to use directly this measure for the choice of the most suitable model. A way to overcome this problem is to plug-in, in (4), the unknown parameter $\boldsymbol{\theta}$ by an estimator which is desirable to obey some nice properties, like consistency and asymptotic normality. Based on this point, the CMDPDE, introduced in [12], can be used. This estimator is described in the sequel for the sake of completeness.

If we denote the kernel of (4) as

$$W_\alpha(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{1+\alpha} d\boldsymbol{y} - \left(1 + \frac{1}{\alpha}\right) \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^\alpha g(\boldsymbol{y}) d\boldsymbol{y}, \tag{7}$$

we can write

$$d_\alpha(g(\cdot), \mathcal{CL}(\boldsymbol{\theta}, \cdot)) = W_\alpha(\boldsymbol{\theta}) + \frac{1}{\alpha} \int_{\mathbb{R}^m} g(\boldsymbol{y})^{1+\alpha} d\boldsymbol{y}$$

and the term $\frac{1}{\alpha} \int_{\mathbb{R}^m} g(\boldsymbol{y})^{1+\alpha} d\boldsymbol{y}$ does not depend on $\boldsymbol{\theta}$ and could be ignored in (9). A natural estimator of $W_\alpha(\boldsymbol{\theta})$, given in (7), can be obtained by observing that the last integral in (7), can be expressed in the form $\int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^\alpha dG(\boldsymbol{y})$, for G the distribution function corresponding to g. Hence, if the

empirical distribution function of $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ will be exploited, this last integral is approximated by $\frac{1}{n}\sum_{i=1}^{n} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{Y}_i)^{\alpha}$, i.e.,

$$W_{n,\alpha}(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{\alpha+1} d\boldsymbol{y} - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^{n} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{Y}_i)^{\alpha}. \tag{8}$$

Definition 1. *The CMDPDE of* $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_c^{\alpha}$, *is defined, for* $\alpha > 0$, *by*

$$\widehat{\boldsymbol{\theta}}_c^{\alpha} = \arg\min_{\boldsymbol{\theta} \in \Theta} W_{n,\alpha}(\boldsymbol{\theta}). \tag{9}$$

We shall denote the score of the composite likelihood by

$$\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y}) = \frac{\partial log \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\theta}}. \tag{10}$$

Let $\boldsymbol{\theta}_0$ be the true value of the parameter $\boldsymbol{\theta}$. In [12], it was shown that the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_c^{\alpha}$ is given by

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_c^{\alpha} - \boldsymbol{\theta}_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{J}_{\alpha}(\boldsymbol{\theta}_0) \boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0)^{-1}\right),$$

being

$$\boldsymbol{H}_{\alpha}(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{\alpha+1} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y})^T d\boldsymbol{y} \tag{11}$$

and

$$\begin{aligned} \boldsymbol{J}_{\alpha}(\boldsymbol{\theta}) = & \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{2\alpha+1} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y})^T d\boldsymbol{y} \\ & - \int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{\alpha+1} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{y} \int_{\mathbb{R}^m} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y})^T \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})^{1+\alpha} d\boldsymbol{y}. \end{aligned} \tag{12}$$

Remark 1. *For* $\alpha = 0$ *we get the CMLE of* $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}}_c = \arg\min_{\boldsymbol{\theta} \in \Theta} \left(-\frac{1}{n} \sum_{i=1}^{n} \log \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y}_i)\right). \tag{13}$$

At the same time it is well-known that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{G}_*(\boldsymbol{\theta})^{-1}\right),$$

where $\boldsymbol{G}_*(\boldsymbol{\theta})$ *denotes the Godambe information matrix defined by* $\boldsymbol{G}_*(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta}) \boldsymbol{J}(\boldsymbol{\theta})^{-1} \boldsymbol{H}(\boldsymbol{\theta})$, *with* $\boldsymbol{H}(\boldsymbol{\theta})$ *being the sensitivity or Hessian matrix and* $\boldsymbol{J}(\boldsymbol{\theta})$ *being the variability matrix, defined, respectively, by*

$$\boldsymbol{H}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[-\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{Y})^T\right], \quad \boldsymbol{J}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{Y}) \boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{Y})^T\right].$$

3. A New Model Selection Criterion

In order to describe the CLDIC criterion we consider the model M_k given in (6). Following standard methodology (cf. [28], pp. 240), the most suitable candidate model to describe the data $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ is the model that minimizes the expected estimated DPD

$$E_{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n}\left[d_{\alpha}(g(\cdot), \mathcal{CL}(\widehat{\boldsymbol{\theta}}_c^{\alpha}, \cdot))\right], \tag{14}$$

subject to the assumption that the unknown model g is belonging to Ξ, i.e., the true model is included in $\{M_s\}_{s \in \{1,...,\ell\}}$ and taking into account that $\widehat{\boldsymbol{\theta}}_c^{\alpha}$, defined in (9), is a consistent and asymptotic normally

distributed estimator of $\boldsymbol{\theta}$. However, this expected value is still depending on the unknown parameter $\boldsymbol{\theta}$. So, as a criterion, it should be used an asymptotically unbiased estimator of (14), for $g \in \Xi$.

The most appropriate model to select is the model which minimizes the expected value

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_\alpha\left(\widehat{\boldsymbol{\theta}}_c^\alpha\right)\right].$$

This expected value is still depending on the unknown parameter $\boldsymbol{\theta}$. So, an asymptotically unbiased estimator of the above expected value could be the basis of a selection criterion, for $g \in \Xi$. In order to proceed with the derivation of such an asymptotically unbiased estimator of $E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_\alpha\left(\widehat{\boldsymbol{\theta}}_c^\alpha\right)\right]$. The empirical version of $W_\alpha(\boldsymbol{\theta})$, in (7), is $W_{n,\alpha}(\boldsymbol{\theta})$, given in (8), and plays a central role in the development of the model selection criterion on the basis of the next theorem which expresses the expected value $E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_\alpha\left(\widehat{\boldsymbol{\theta}}_c^\alpha\right)\right]$ by means of the respective expected value of $W_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^\alpha)$, in an asymptotically equivalent way.

Theorem 1. *If the true distribution g belongs to the parametric family Ξ and $\boldsymbol{\theta}_0$ denotes the true value of the parameter $\boldsymbol{\theta}$, then we have*

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\right] = E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_{n,\alpha}(\widehat{\boldsymbol{\theta}}_\alpha) + \frac{\alpha+1}{n} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\,\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right)\right] + o_p(1)$$

with $\boldsymbol{H}_\alpha(\boldsymbol{\theta})$ and $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ given in (11) and (12), respectively.

Based on the above theorem, the proof of which is presented in a full detail in the Appendix A, an asymptotic unbiased estimator of $E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\right]$ is given by

$$W_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^\alpha) + \frac{\alpha+1}{n} trace\left(\boldsymbol{J}_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\boldsymbol{H}_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)^{-1}\right).$$

This ascertainment is the basis and a strong motivation for the next definition which introduces the model selection criterion.

Definition 2. *Let $\{M_k\}_{k\in\{1,\ldots,\ell\}}$ be candidate models for the observations $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$. The selected model M^* verifies*

$$M^* = \min_{k\in\{1,\ldots,\ell\}} CLDIC_\alpha(M_k),$$

where

$$CLDIC_\alpha(M_k) = W_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^\alpha) + \frac{\alpha+1}{n} trace\left(\boldsymbol{J}_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\boldsymbol{H}_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)^{-1}\right),$$

$W_{n,\alpha}(\boldsymbol{\theta})$ was given in (8) and $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ and $\boldsymbol{H}_\alpha(\boldsymbol{\theta})$ were defined in (11) and (12), respectively.

The next remark summarizes the model selection criterion in the case $\alpha = 0$ and it therefore extends, in a sense, the pioneer and classic AIC.

Remark 2. *For $\alpha = 0$ we have,*

$$d_{KL}(g(\cdot), \mathcal{CL}(\boldsymbol{\theta},\cdot)) = W_0(\boldsymbol{\theta}) + \int_{\mathbb{R}^n} g(\boldsymbol{y}) \log g(\boldsymbol{y}) d\boldsymbol{y}$$

with $W_0(\boldsymbol{\theta}) = -\int_{\mathbb{R}^n} \log \mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y}$. Therefore, the most appropriate model which should be selected, is the model which minimizes the expected value

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_0(\widehat{\boldsymbol{\theta}}_c)\right], \tag{15}$$

where $\widehat{\boldsymbol{\theta}}_c$ *is the CMLE of* $\boldsymbol{\theta}_0$ *defined in (9).*

The expected value (15) is still depending on the unknown parameter $\boldsymbol{\theta}$*. A natural estimator of* $W_0(\widehat{\boldsymbol{\theta}}_c)$ *can be obtained by replacing the distribution function G, of g, by the empirical distribution function based on* $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$*,*

$$W_{n,0}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y}_i).$$

Based on it, we select the model M^* *that verifies*

$$M^* = \min_{k\in\{1,\ldots,,\ell\}} CLDIC_0\left(M_k\right),$$

with

$$CLDIC_0\left(M_k\right) = H_{n,0}(\widehat{\boldsymbol{\theta}}_c) + \frac{1}{n} trace\left(\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_c)\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_c)^{-1}\right),$$

where $\boldsymbol{J}(\widehat{\boldsymbol{\theta}}_c)$ *and* $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_c)$ *are defined in Remark 1. In a manner, quite similar to that of the previous theorem, it can be established that* $CLDIC_0(M_k)$ *is an asymptotic unbiased estimator of* $E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_0(\widehat{\boldsymbol{\theta}}_c)\right]$*.*

This would be the model selection criterion in a composite likelihood framework based on Kullback–Leibler divergence. We can observe that this criterion coincides with the criterion given in [22] as a generalization of the classical criterion of Akaike, which will be referred from now as Composite Akaike Information Criterion (CAIC).

4. Numerical Simulations

4.1. Scenario 1: Two-Component Mixed Model

We are starting with a simulation example, which is motivated and follows ideas from the paper [29] and the Example 4.1 in [20] which will compare the behaviour of the proposed criteria with the CAIC criterion, for $\alpha = 0$ (see Remark 2).

Consider the random vector $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$ from an unknown density g and let now $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be independent and identically distributed replications of $\boldsymbol{Y}$ which are described by the true but unknown distribution g. Taking into account that the true model g is unknown, suppose that $\{f(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, p \geq 1\}$ is a parametric identifiable family of candidate distributions to describe the observations $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$. Let also $\mathcal{CL}(\boldsymbol{\theta}, \boldsymbol{y})$ denotes the composite likelihood function associated to the parametric model $f(\cdot;\boldsymbol{\theta})$.

We consider the problem of choosing (on the basis of n independent and identically distributed replications $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ of $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)^T$) between a 4-variate normal distribution, $\mathcal{N}\left(\boldsymbol{\mu}^N, \boldsymbol{\Sigma}\right)$, with $\boldsymbol{\mu}^N = (\mu_1^N, \mu_2^N, \mu_3^N, \mu_4^N)^T$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & 2\rho & 2\rho \\ \rho & 1 & 2\rho & 2\rho \\ 2\rho & 2\rho & 1 & \rho \\ 2\rho & 2\rho & \rho & 1 \end{pmatrix},$$

and a 4-variate t-distribution with ν degrees of freedom, $t_\nu\left(\boldsymbol{\mu}^{t_\nu}, \boldsymbol{\Sigma}^*\right)$, with different location parameters $\boldsymbol{\mu}^{t_\nu} = (\mu_1^{t_\nu}, \mu_2^{t_\nu}, \mu_3^{t_\nu}, \mu_4^{t_\nu})^T$ and same variance-covariance matrix $\boldsymbol{\Sigma}$, and density,

$$C_m|\boldsymbol{\Sigma}^*|^{-1/2}\left[1 + \frac{1}{\nu}(\boldsymbol{y} - \boldsymbol{\mu}^{t_\nu})^T(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{t_\nu})\right]^{-(\nu+m)/2},$$

with $\boldsymbol{\Sigma}^* = \frac{\nu-2}{\nu}\boldsymbol{\Sigma}$, $C_m = (\pi\nu)^{-m/2}\frac{\Gamma[(\nu+m)/2]}{\Gamma(\nu/2)}$ and $m = 4$.

Consider the composite likelihood function,

$$\mathcal{CL}N(\rho, \boldsymbol{y}) = f_{A_1}^N(\boldsymbol{y};\rho) f_{A_2}^N(\boldsymbol{y};\rho),$$

with $f_{A_1}^N(\boldsymbol{y};\rho) = f_{12}^N(y_1, y_2; \mu_1^N, \mu_2^N; \rho)$ and $f_{A_2}^N(\boldsymbol{y};\rho) = f_{34}^N(y_3, y_4; \mu_3^N, \mu_4^N; \rho)$, where f_{12}^N and f_{34}^N are the densities of the marginals of $\boldsymbol{Y}$, i.e., bivariate normal distributions with mean vectors $(\mu_1^N, \mu_2^N)^T$ and $(\mu_3^N, \mu_4^N)^T$, respectively, and common variance-covariance matrix

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In a similar manner consider the composite likelihood

$$\mathcal{CL}t_\nu(\rho, \boldsymbol{y}) = f_{A_1}^{t_\nu}(\boldsymbol{y};\rho) f_{A_2}^{t_\nu}(\boldsymbol{y};\rho),$$

with $f_{A_1}^{t_\nu}(\boldsymbol{y};\rho) = f_{12}^{t_\nu}(y_1, y_2; \mu_1^{t_\nu}, \mu_2^{t_\nu}; \rho)$ and $f_{A_2}^{t_\nu}(\boldsymbol{y};\rho) = f_{34}^{t_\nu}(y_3, y_4; \mu_3^{t_\nu}, \mu_4^{t_\nu}; \rho)$, where $f_{12}^{t_\nu}$ and $f_{34}^{t_\nu}$ are the densities of the marginals of $\boldsymbol{Y}$, i.e., bivariate t-distributions with mean vectors $(\mu_1^{t_\nu}, \mu_2^{t_\nu})^T$ and $(\mu_3^{t_\nu}, \mu_4^{t_\nu})^T$, respectively, and common variance-covariance matrix

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Under this formulation, the simulation study follows in the next two scenarios.

4.1.1. Scenario 1a

Following Example 4.1 in [20], the steps of the simulation study are the following:

- Generate 1000 samples of size $n = 5, 7, 10, 20, 40, 50, 70, 100$ from a two component mixture of two 4-variate distributions, namely, a 4-variate normal and a 4-variate t-distribution,

$$h_\omega(\boldsymbol{y}) = \omega N\left(\boldsymbol{\mu}^N, \boldsymbol{\Sigma}\right) + (1-\omega) t_\nu\left(\boldsymbol{\mu}^{t_\nu}, \boldsymbol{\Sigma}^*\right), \quad 0 \le \omega \le 1,$$

  with $\boldsymbol{\mu}^N = (0, 0, 0.5, 0)$ and $\boldsymbol{\mu}^{t_\nu} = (3.2, 1.5, 0.5, 2)$, for $\omega = 0, 0.25, 0.45, 0.5, 0.55, 0.75, 1$, $\nu = 5, 10, 30$ degrees of freedom and with specific values of $\rho = -0.15, -0.10, 0.10$. As pointed out in [29], taking into account that $\boldsymbol{\Sigma}$ should be semi-positive definite, the following condition is imposed: $-\frac{1}{5} \le \rho \le \frac{1}{3}$.
- Estimate the common parameter ρ, separately in each model, by using the CMDPDE estimator for different values of the tuning parameter $\alpha = 0, 0.3$. The composite density which corresponds to the mixture $h_\omega(\boldsymbol{y})$ is defined by

$$\mathcal{CL}(\rho, \boldsymbol{y}) = \omega \mathcal{CL}N(\rho, \boldsymbol{y}) + (1-\omega)\mathcal{CL}t_\nu(\rho, \boldsymbol{y}), \quad 0 \le \omega \le 1,$$

  and it is used to obtain the CMDPDE estimator, $\widehat{\rho}$, of ρ.
- Define the mixture composite likelihood function

$$\mathcal{CL}(\widehat{\rho}, \boldsymbol{y}) = \omega \mathcal{CL}N(\widehat{\rho}, \boldsymbol{y}) + (1-\omega)\mathcal{CL}t_\nu(\widehat{\rho}, \boldsymbol{y}), \quad 0 \le \omega \le 1.$$

- Calculate $CLDIC_\alpha\left(M_k\right)$, the value of the model selection criterion considered in this paper, for the two candidate models, with

$$CLDIC_\alpha\left(M_k\right) = W_{n,\alpha}\left(\widehat{\rho}\right) + \frac{\alpha+1}{n} trace\left(\boldsymbol{J}_\alpha\left(\widehat{\rho}\right)\boldsymbol{H}_\alpha\left(\widehat{\rho}\right)^{-1}\right).$$

  An explanation of how to obtain this value for the both candidate models is given in Appendix B.
- Compute the times that the 4-variate normal model was selected.

Results are summarized in Table 1. Extreme values of $\omega = 0, 1$ represent the times that the 4-variate normal model was selected under the 4-variate t-distribution and 4-variate normal distribution,

respectively. This means that, for $\omega = 1$, the perfect discrimination will be achieved when 1000 of the 1000 simulated samples are correctly assigned, while for $\omega = 0$, the more near to 0, the better discrimination of the criterion. $\omega = 0.5$ means that each sample was generated both from the normal and t-distribution in the same proportion.

Table 1. Main results, Scenario 1a.

	$\alpha = 0$ (CAIC)							$\alpha = 0.3$						
ω	**0**	**0.25**	**0.45**	**0.5**	**0.55**	**0.75**	**1**	**0**	**0.25**	**0.45**	**0.5**	**0.55**	**0.75**	**1**
							$\nu = 5$, $\rho = -0.15$							
n = 5	0	1	269	499	713	996	1000	0	0	273	498	712	1000	1000
7	0	1	246	504	758	998	1000	0	1	220	511	738	999	1000
10	0	0	202	482	775	1000	1000	0	0	185	467	771	1000	1000
20	0	0	114	486	871	1000	1000	0	0	112	473	866	1000	1000
40	0	0	41	459	947	1000	1000	0	0	54	496	954	1000	1000
50	0	0	21	475	964	1000	1000	0	0	41	556	986	1000	1000
70	0	0	9	461	985	1000	1000	0	0	48	656	995	1000	1000
100	0	0	5	472	992	1000	1000	0	0	142	885	1000	1000	1000
							$\nu = 10$, $\rho = -0.15$							
5	0	3	222	445	688	996	1000	0	3	218	433	688	997	1000
7	0	1	191	439	720	1000	1000	0	0	179	431	690	999	1000
10	0	0	163	432	747	1000	1000	0	0	152	402	725	1000	1000
20	0	0	59	399	819	1000	1000	0	0	49	361	773	1000	1000
40	0	0	19	336	912	1000	1000	0	0	12	326	899	1000	1000
50	0	0	6	362	936	1000	1000	0	0	10	334	925	1000	1000
70	0	0	1	292	960	999	1000	0	0	2	356	973	1000	1000
100	0	0	0	301	983	1000	1000	0	0	1	531	992	1000	1000
							$\nu = 30$, $\rho = -0.15$							
5	0	4	237	423	677	997	1000	0	2	235	413	656	996	1000
7	0	0	155	394	689	1000	1000	0	0	141	379	677	999	1000
10	0	0	144	413	719	1000	1000	0	0	134	393	701	1000	1000
20	0	0	57	351	801	1000	1000	0	0	40	311	764	1000	1000
40	0	0	11	296	904	1000	1000	0	0	8	263	882	1000	1000
50	0	0	6	271	918	1000	1000	0	0	3	253	903	1000	1000
70	0	0	1	225	942	1000	1000	0	0	0	229	941	1000	1000
100	0	0	0	208	978	1000	1000	0	0	0	303	989	1000	1000
							$\nu = 10$, $\rho = -0.10$							
5	0	4	242	464	680	996	1000	0	3	238	459	682	999	1000
7	0	0	187	461	733	997	1000	0	0	199	457	731	998	1000
10	0	0	162	445	738	1000	1000	0	0	165	407	713	1000	1000
20	0	0	62	378	807	1000	1000	0	0	59	354	789	1000	1000
40	0	0	19	357	902	999	1000	0	0	14	333	895	1000	1000
50	0	0	6	325	932	1000	1000	0	0	8	325	931	1000	1000
70	0	0	2	305	954	1000	1000	0	0	6	367	967	1000	1000
100	0	0	0	307	979	1000	1000	0	0	2	507	993	1000	1000
							$\nu = 10$, $\rho = 0.10$							
5	0	11	268	459	669	991	1000	1	11	268	478	680	993	1000
7	0	1	211	456	720	999	1000	0	3	207	464	716	998	1000
10	0	0	168	423	704	1000	1000	0	0	162	403	702	1000	1000
20	0	0	86	360	789	1000	999	0	0	89	357	786	1000	1000
40	0	0	35	367	893	1000	1000	0	0	38	398	896	1000	1000
50	0	0	19	331	886	1000	1000	0	0	19	360	913	1000	1000
70	0	0	11	311	933	1000	1000	0	0	16	379	963	1000	1000
100	0	0	2	276	969	1000	1000	0	0	7	490	985	1000	1000

4.1.2. Scenario 1b

Same Scenario is evaluated under the more-closed means $\boldsymbol{\mu}^N = (0, 1.5, 0.5, -0.75)$ and $\boldsymbol{\mu}^{t_\nu} = (0, 1.5, 0.5, 2)$ for moderate-large sample sizes and $\alpha \in \{0, 0.2, 0.4\}$. Here $\nu = 5$ and $\rho = -0.15$. Results are shown in Table 2. In this case, the models under consideration are more similar, so it would be understandable that the CLDIC criterion did not discriminate in such as good way.

Table 2. Main results, Scenario 1b.

	$\alpha = 0$ (CAIC)				$\alpha = 0.2$				$\alpha = 0.4$			
	0	**0.25**	**0.75**	**1**	**0**	**0.25**	**0.75**	**1**	**0**	**0.25**	**0.75**	**1**
n = 40	0	0	39	731	0	0	537	961	0	0	580	949
50	0	0	24	732	0	0	859	990	0	0	944	994
60	0	0	14	772	0	0	999	1000	0	1	999	1000
70	0	0	9	734	0	0	999	1000	0	27	999	1000
80	0	0	5	770	0	1	1000	1000	0	326	1000	1000
90	0	0	4	782	0	23	1000	1000	2	794	1000	1000
100	0	0	4	802	0	173	1000	1000	26	978	1000	1000

4.2. Scenario 2: Three-Component Mixed Model

Now, we consider a mixed model composed on two 4-variate normal distributions and a 4-variate t-distribution with $\nu = 10$ degrees of freedom. The three distributions have common variance-covariance matrix, as in the previous scenario, with unknown $\rho = -0.15$ and different but known means $\boldsymbol{\mu}_1^N = (0, 0, 0.5, 0)$, $\boldsymbol{\mu}_2^N = (0, 1.5, 0.5, 0)$ and $\boldsymbol{\mu}^t = (0, 1.5, 0.5, 2)$. The model is defined by

$$\omega \mathcal{N}(\boldsymbol{\mu}_1^N, \boldsymbol{\Sigma}) + \lambda \mathcal{N}(\boldsymbol{\mu}_2^N, \boldsymbol{\Sigma}) + (1 - \omega - \lambda) t_{\nu=10}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^*), \quad 0 \leq \omega, \lambda, \omega + \lambda \leq 1,$$

with $\boldsymbol{\Sigma}$ being again a common variance-covariance matrix with unknown parameter ρ of the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & 2\rho & 2\rho \\ \rho & 1 & 2\rho & 2\rho \\ 2\rho & 2\rho & 1 & \rho \\ 2\rho & 2\rho & \rho & 1 \end{pmatrix}.$$

Following the same steps that in the first scenario, we generate 1000 samples of the three-component mixture for different sample sizes $n = 5, 7, 10, 20, 40, 50, 70, 100$ and different values of ω and λ. Then, we consider the problem of choosing among one of the two 4-variate normal distributions and the 4-variate t-distribution through the CLDIC criterion, for different values of the tuning parameter $\alpha = 0, 0.3, 0.5, 0.7$. See Table 3 for results. Here, the normal models are denoted by N1 and N2, respectively, while the 4-variate t-distribution is denoted by MT. The first three cases evaluate the selected model under these multivariate distributions. In the last two scenarios, a mixed model is considered as the true distribution.

4.3. Discussion of Results

In Scenario 1a, two well-differentiated multivariate models are considered. In this case CLDIC criterion works in a very efficient way, with an almost-perfect discrimination for extreme values of ω. The good behaviour is also observed for not so extreme values of ω, such as $\omega = 0.55$ or 0.45. We can not observe a significant difference in the choice of α.

In Scenario 1b we consider closer models, which affect the discrimination power of the CLDIC. However, in this case, we do observe great differences when considering different α. While the discrimination power of CLDIC for $\alpha = 0$ (CAIC) and $\omega = 1$ is around 75%, for $\alpha = 0.2$ or $\alpha = 0.4$

the behaviour is excellent. This happens also for large but not extreme values of ω, such as $\omega = 0.75$. However, a medium value of α turns into a worse discrimination for low values of ω.

Table 3. Main results, Scenario 2.

	$\alpha = 0$ (CAIC)			$\alpha = 0.3$			$\alpha = 0.5$			$\alpha = 0.7$		
Model *	**N1**	**N2**	**MT**	**N1**	**N2**	**MT**	**N1**	**N2**	**MT**	**N1**	**N2**	**MT**
True model: $\mathcal{N}(\boldsymbol{\mu}_1^N, \boldsymbol{\Sigma})$												
n = 5	957	24	19	950	16	34	939	23	38	936	28	36
7	970	19	11	966	13	24	961	13	26	950	22	28
10	993	3	4	986	4	10	979	6	15	971	6	23
20	1000	0	0	1000	0	0	998	0	2	997	0	3
40	1000	0	0	1000	0	0	1000	0	0	1000	0	0
50	1000	0	0	1000	0	0	1000	0	0	1000	0	0
70	1000	0	0	1000	0	0	1000	0	0	1000	0	0
100	1000	0	0	1000	0	0	1000	0	0	999	0	0
True model: $\mathcal{N}(\boldsymbol{\mu}_2^N, \boldsymbol{\Sigma})$												
5	29	638	333	34	610	356	38	639	323	50	646	304
7	15	622	363	13	589	398	17	599	384	28	627	345
10	6	610	384	5	540	455	5	540	455	11	586	403
20	1	612	387	1	518	481	1	472	527	1	527	472
40	0	566	434	0	650	350	0	590	410	0	614	386
50	0	561	439	0	804	196	0	797	203	0	835	165
70	0	584	416	0	987	13	0	994	6	0	998	2
100	0	520	480	0	1000	0	0	1000	0	0	1000	0
True model: $t_{\nu=10}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma})$												
5	2	15	983	1	6	993	1	8	991	3	15	982
7	0	3	997	0	1	999	2	2	996	0	4	996
10	0	1	999	0	2	998	0	2	998	0	3	997
20	0	0	1000	0	0	1000	0	0	1000	0	0	1000
40	0	0	1000	0	0	1000	0	0	1000	0	0	1000
50	0	0	1000	0	0	1000	0	0	1000	0	0	1000
70	0	0	1000	0	0	1000	0	0	1000	0	0	1000
100	0	0	1000	0	0	1000	0	4	996	0	296	704
True model: $0.7\mathcal{N}(\boldsymbol{\mu}_2^N, \boldsymbol{\Sigma}) + 0.3t_{\nu=10}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma})$												
5	6	384	610	6	375	619	4	401	595	11	452	537
7	1	331	668	1	294	705	1	317	682	1	373	626
10	1	261	738	1	218	781	1	253	746	1	306	693
20	0	109	891	0	101	899	0	107	893	0	141	859
40	0	26	974	0	126	874	0	122	878	0	166	834
50	0	13	987	0	311	689	0	345	655	0	445	555
70	0	6	994	0	948	52	0	982	18	0	994	6
100	0	2	998	0	1000	0	0	1000	0	0	999	1
True model: $\frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_1^N, \boldsymbol{\Sigma}) + \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_2^N, \boldsymbol{\Sigma}) + \frac{1}{3}t_{\nu=10}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma})$												
5	127	377	496	121	363	516	107	392	501	107	424	469
7	87	357	556	70	339	591	66	356	578	63	396	541
10	69	326	605	61	314	625	56	330	614	45	381	574
20	37	259	704	25	298	677	17	337	646	15	349	636
40	7	145	848	9	452	539	4	508	488	1	469	530
50	2	122	876	5	744	251	3	814	183	3	853	144
70	0	99	901	4	996	0	4	996	0	4	996	0
100	0	36	964	355	645	0	645	355	0	856	144	0

* Here the model candidates are expressed as N1, N2, MT to denote $\mathcal{N}(\boldsymbol{\mu}_1^N, \boldsymbol{\Sigma})$, $\mathcal{N}(\boldsymbol{\mu}_2^N, \boldsymbol{\Sigma})$ and $t_{10}(\boldsymbol{\mu}^t, \boldsymbol{\Sigma})$, respectively.

Scenario 2 deals with three different models, two multivariate normal and one multivariate t (N1, N2 and MT, respectively). The second normal distribution is closer to MT in terms of means. While CLDIC criterion discriminate well between N1 and N2 and between N1 and MT, it has difficulties in distinguishing N2 an MT distributions, overall for small samples sizes and $\alpha = 0$.

It seems, therefore, that when we have well-discriminated models, CLDIC criterion works very well, independently of the sample size and the tuning parameter α considered. Dealing with closer models leads, as expected, to worst results, overall for $\alpha = 0$ (CAIC).

Note that the behaviour of Wald-type and Rao tests based on CMDPDEs was studied in [12,13] through extensive simulation studies.

5. Numerical Examples

5.1. Choice of the Tuning Parameter

In the previous sections, we have seen that CLDIC criterion works generally very well, independently of α, but that some values present a better behaviour, overall when distinguishing similar models. In these situations, it appears that values close to 0.2 or 0.3 work well, while CAIC criterion presents a worse behaviour. A data-driven approach for the choice of the tuning parameter which would be helpful in practice. The approach of [30] was adapted In [13], for the choice of the optimum α in CMDPDEs. This approach consisted on minimizing the estimated mean squared error by means of a pilot estimator, $\boldsymbol{\theta}^P$. This approximation is given by

$$\widehat{MSE}_\alpha = (\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}^P)^T(\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}^P) + \frac{1}{n}\text{Trace}\left(\boldsymbol{H}_\alpha^{-1}(\widehat{\boldsymbol{\theta}}_c^\alpha)\boldsymbol{J}_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\boldsymbol{H}_\alpha^{-1}(\widehat{\boldsymbol{\theta}}_c^\alpha)\right), \tag{16}$$

where $\boldsymbol{H}_\alpha(\boldsymbol{\theta})$ and $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ are given in (11) and (12). The optimum α will be the one that minimizes expression (16). The choice of the pilot estimator is probably one of the major drawbacks of this approach, as it may lead to a choice of α too close to that used for the pilot estimator. A pilot estimator with $\alpha \approx 0.4$, was proposed in [13] after some simulations, in concordance with [30], where the initial choice of a pilot is suggested to be a robust one in order to obtain the best results in terms of robustness.

5.2. Iris Data

The Iris data (Fisher, [31]) includes 3 categories of 50 sample values each, where each category refers to a type of iris plant: *setosa*, *versicolor* and *virginica*. Each plant is categorized in its class and described by other 4 variables: (1) sepal length, (2) sepal width, (3) petal length and (4) petal width. This is one of the most known data sets for discriminant analysis. [32] proposed the use of a Gaussian finite mixture for modeling Iris data, in which each known class is modeled by a single Gaussian term with the same variance-covariance matrix. The resulting model is as follows

$$f(\boldsymbol{x}) = \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + \frac{1}{3}\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}), \tag{17}$$

with

$$\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14})^T, \quad \boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})^T, \quad \boldsymbol{\mu}_3 = (\mu_{31}, \mu_{32}, \mu_{33}, \mu_{34})^T$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}.$$

Exact values can be obtained by *MclustDA*() function of *mclust* package in R Software ([32]).

We propose a composite likelihood approach to modeling (17) where we suppose independence between the two first and two last variables. This is

$$f_{CL}(\boldsymbol{y}) = \frac{1}{3}CLN_1 + \frac{1}{3}CLN_2 + \frac{1}{3}CLN_3, \tag{18}$$

with

$$CLN_i = f^N_{A_{i1}}(\rho_{12}, \boldsymbol{y}) f^N_{A_{i2}}(\rho_{34}, \boldsymbol{y}),$$

where $f^N_{A_{i1}}(\rho_{12}, \boldsymbol{y}) = f^N_{A_{i1}}(\rho_{12}, \mu_{i1}, \mu_{i2}, \boldsymbol{\Sigma}_{A_1}, \boldsymbol{y})$ and $f^N_{A_{i2}}(\rho_{34}, \boldsymbol{y}) = f^N_{A_{i2}}(\rho_{34}, \mu_{i3}, \mu_{i4}, \boldsymbol{\Sigma}_{A_2}, \boldsymbol{y})$, $i = 1, 2, 3$ are bivariate normals with variance-covariance matrices

$$\boldsymbol{\Sigma}_{A_1} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{A_2} = \begin{pmatrix} \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{pmatrix}.$$

We are going to evaluate the behavior of the CLDIC criterion proposed in previous sections. After estimating parameters ρ_{12} and ρ_{34} in (18), we consider 10 different subsets of the IRIS data:

- SE subset: 50 first observations, corresponding to Setosa plants ($n = 50$).
- VE subset: 50 second observations, corresponding to Versicolor plants ($n = 50$).
- VI subset: 50 last observations, corresponding to Virginica plants ($n = 50$).
- SE(VE) subset: SE subset with 2 first observations of VE subset ($n = 52$).
  Equivalently: SE(VI), VE(SE), VE(VI), VI(SE) and VI(VE).
- VI(SE+VE) subset: VI subset with 2 first observations of SE and VE subsets ($n = 54$).

In Table 4, chosen models for each one of the subsets are obtained by the proposed CLDIC criterion. When a "pure" subset is considered, all the tuning parameters lead to optimal decisions, but when a "contaminated" subset is under consideration, only $\alpha = 0.2, 0.3$ have an optimal response in all the cases.

Table 4. Selected model in each of the subsets. Iris data.

α	SE	VE	VI	SE(VE)	SE(VI)	VE(SE)	VE(VI)	VI(SE)	VI(VE)	VI(SE+VE)
0 (CAIC)	CN1	CN2	CN3	CN1	CN1	CN1*	CN2	CN1*	CN3	CN3
0.2	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.3	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.4	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN1*	CN3	CN3
0.5	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN1*	CN3	CN3
0.8	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN1*	CN3	CN3
0.22	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3

We now apply the ad hoc approach presented in Section 5.1 for selecting the tuning parameter α in a composite likelihood framework. Applying this procedure to our data set though a grid search of length 100 and by means of a pilot estimator with $\alpha = 0.4$ leads to the optimal tuning parameter $\alpha = 0.22$, what is in concordance with the obtained results (see Table 5). We can see that the use of other pilot estimators would not affect very much to the final decission.

Table 5. Selected α for different pilot estimators, ad-hoc tuning parameter selection procedure. Iris and Wine data

	α_{pilot}	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Iris	α_{opt}	0.31	0.17	0.20	0.21	**0.22**	0.23	0.24	0.24	0.25	0.25	0.25
Wine	α_{opt}	0.45	0.46	0.47	0.49	**0.51**	0.53	0.55	0.56	0.56	0.56	0.57

5.3. Wine Data

We now work with Wine data ([33]), which contain a chemical analysis of 178 Italian wines from three different cultivars (Barolo, Grignolino, Barbera) yielded 13 measurements. In order to illustrate our criterion, we will work with only first four explanatory variables: Alcohol, Malic, Ash and Alkalinity. As in the previous section, we adjust a Gaussian mixture model with weights, in this case: 59/178 , 72/178 and 47/178 corresponding to Barolo, Grignolino and Barbera classes, respectively. We now consider these 10 different subsets of the Wine data:

- BO subset: 20 first observations of Barolo wines ($n = 20$).
- GR subset: 20 first observations of Grignolino wines ($n = 20$).
- BA subset: 20 first observations of Barbera wines ($n = 20$).
- BO(GR) subset: BO subset with 5 first observations of GR subset ($n = 25$).
  Equivalently: BO(BA), GR(BO), GR(BA), BA(BO) and BA(GR).
- BA(BO+GR) subset: BA subset with 3 first observations of BO and GR subsets ($n = 26$).

We can observe how, for medium values of α, the discrimination is perfect (see Table 6). Applying ad-hoc tuning parameter choice procedure we obtain $\alpha_{opt} \approx 0.51$, with a perfect discrimination again (Table 5).

Table 6. Selected model in each of the subsets. Wine data.

α	**BO**	**GR**	**BA**	**BO(GR)**	**BO(BA)**	**GR(BO)**	**GR(BA)**	**BA(BO)**	**BA(GR)**	**BA(BO+GR)**
0 (CAIC)	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN2*
0.2	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.3	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.4	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.5	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3
0.8	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN2*	CN2*	CN3
0.51	CN1	CN2	CN3	CN1	CN1	CN2	CN2	CN3	CN3	CN3

6. Conclusions and Future Research

In this paper, we have addressed the problem of model selection in the framework of composite likelihood methodology, on the basis of the DPD as a measure of the closeness of the composite density and the true model that drives the data. In this context, an information criterion is introduced and studied which is defined by means of composite minimum distance type estimators of the unknown parameters, well-known for having nice robustness properties. Thanks to a simulation study, we have shown that the proposed here model selection criterion works well in practice and mainly that the use of CMDPDE makes the criterion more robust than the criteria based on the classic CMLE and the Kullback–Leibler divergence, given in [22]. The analysis of two real data examples of the literature illustrate on how the model selection criterion, presented here, can be applied in practical cases. This paper is a part of a series of papers by the authors where composite likelihood ideas and methods are harmonically weaved with divergence theoretic methods in order to develop statistical inference (estimation and testing of hypotheses) and model selection criteria, as well. We envision future work in some directions. The development of change point methodology on the basis of composite density with CMDPDE and divergence measures would be maybe an appealing problem for a future research on the topic. However, all the information theoretic methods developed on the

basis of the composite likelihood depend on the choice of the family of sets $\{A_k\}_{k=1}^K$, appeared in Formula (1). A question is raised at this point: how the information theoretic procedures developed on the basis of the composite likelihood are affected by this family of sets? It is an appealing problem which deserves also investigation in a future work.

Author Contributions: Conceptualization, E.C., N.M., L.P. and K.Z.; Methodology, E.C., N.M., L.P. and K.Z.; Software, E.C., N.M., L.P. and K.Z.; Validation, E.C., N.M., L.P. and K.Z.; Formal Analysis, E.C., N.M., L.P. and K.Z.; Investigation, E.C., N.M., L.P. and K.Z.; Resources, E.C., N.M., L.P. and K.Z.; Data Curation, E.C., N.M., L.P. and K.Z.; Writing—Original Draft Preparation, E.C., N.M., L.P. and K.Z.; Writing—Review & Editing, E.C., N.M., L.P. and K.Z.; Visualization, E.C., N.M., L.P. and K.Z.; Supervision, E.C., N.M., L.P. and K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by Grant PGC2018-095194-B-I00 and Grant FPU16/03104 from Ministerio de Ciencia, Innovación y Universidades (Spain). E. Castilla, N. Martín and L. Pardo are members of the Instituto de Matemática Interdisciplinar, Complutense University of Madrid.

Acknowledgments: The authors would like to thank the Editor and Reviewers for taking their precious time to make several valuable comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLE	Maximum likelihood estimator
CMLE	Composite maximum likelihood estimator
CLDIC	Composite likelihood DIC
DPD	Density power divergence
MDPDE	Minimum density power divergence estimator
CMDPDE	Composite minimum density power divergence estimator
AIC	Akaike Information Criterion
CAIC	Composite Akaike Information Criterion
TIC	Takeuchi Information Criterion

Appendix A. Proof of Theorem 1

Proof. A Taylor expansion of $W_\alpha(\boldsymbol{\theta})$ around the true parameter $\boldsymbol{\theta}_0$ and evaluated in $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_c^\alpha$, gives

$$W_\alpha\left(\widehat{\boldsymbol{\theta}}_c^\alpha\right) = W_\alpha(\boldsymbol{\theta}_0) + \left(\frac{\partial W_\alpha(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}_0\right) + \frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}_0\right)^T\left(\frac{\partial^2 W_\alpha(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left(\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}_0\right) + o\left(\left\|\left(\widehat{\boldsymbol{\theta}}_c^\alpha - \boldsymbol{\theta}_0\right)\right\|^2\right).$$

Now,

$$\begin{aligned}\frac{\partial W_\alpha(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \int_{\mathbb{R}^m}(1+\alpha)\,\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^\alpha\frac{\partial\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}}d\boldsymbol{y} - \left(1+\frac{1}{\alpha}\right)\alpha\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha-1}\frac{\partial\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}}g(\boldsymbol{y})d\boldsymbol{y}\\ &= (1+\alpha)\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{y} - (1+\alpha)\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^\alpha\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,g(\boldsymbol{y})d\boldsymbol{y}.\end{aligned}$$

It is clear that if the true distribution g belongs to the parameter family $f(.;\boldsymbol{\theta}), \boldsymbol{\theta}\in\Theta$ and $\boldsymbol{\theta}_0$ denotes the true value of the parameter $\boldsymbol{\theta}$, we get

$$\left(\frac{\partial W_\alpha(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}.$$

Now we are going to get

$$\begin{aligned}\frac{\partial^2 W_\alpha(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} &= (1+\alpha)\left\{\int_{\mathbb{R}^m}(1+\alpha)\,\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})^T d\boldsymbol{y}\right.\\ &-\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\left(-\frac{\partial^2\log\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)d\boldsymbol{y}\\ &\left.-\alpha\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})^T g(\boldsymbol{y})d\boldsymbol{y}+\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha}\left(-\frac{\partial^2\log\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)g(\boldsymbol{y})d\boldsymbol{y}\right\}.\end{aligned}$$

If the true distribution g belongs to the parameter family $f_{\boldsymbol{\theta}}(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\Theta$ and $\boldsymbol{\theta}_0$ denotes the true value of the parameter $\boldsymbol{\theta}$, verifies,

$$\begin{aligned}\left(\frac{\partial^2 W_\alpha(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= (1+\alpha)\int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta}_0,\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta}_0,\boldsymbol{y})\,\boldsymbol{u}(\boldsymbol{\theta}_0,\boldsymbol{y})^T d\boldsymbol{y}\\ &= (1+\alpha)\,\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0).\end{aligned}$$

Therefore,

$$nW_\alpha\left(\widehat{\boldsymbol{\theta}}_c^\alpha\right) = nW_\alpha(\boldsymbol{\theta}_0)+\frac{(1+\alpha)}{2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)^T\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)+no\left(\left\|\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)\right\|^2\right).$$

But

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)\xrightarrow[n\to\infty]{L}N\left(\mathbf{0},\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right),$$

and $no\left(\left\|\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)\right\|^2\right)=o(O_p(1))=o_p(1)$.

The asymptotic distribution of the quadratic form $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)^T\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)$, verifies

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)^T\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^\alpha-\boldsymbol{\theta}_0\right)\xrightarrow[n\to\infty]{\mathcal{L}}\sum_{r=1}^k\lambda_r Z_r^2$$

being λ_r, $r=1,...,k$, the eigenvalues of the matrix

$$\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)\,\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}=\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}$$

and Z_r are independent normal random variable of mean zero and variance 1. Therefore,

$$\begin{aligned}E_{\boldsymbol{Y}_1,...,\boldsymbol{Y}_n}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^a-\boldsymbol{\theta}_0\right)^T\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_c^a-\boldsymbol{\theta}_0\right)\right] &= \sum_{r=1}^k\lambda_r+o_p(1)\\ &= trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right)+o_p(1)\end{aligned}$$

and

$$E_{\boldsymbol{Y}_1,...,\boldsymbol{Y}_n}\left[nW_\alpha(\widehat{\boldsymbol{\theta}}_c^\alpha)\right]=nW_\alpha(\boldsymbol{\theta}_0)+\frac{(1+\alpha)}{2}trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right)+o_p(1).$$

Now a Taylor expansion of $W_{n,\alpha}(\boldsymbol{\theta})$, around $\widehat{\boldsymbol{\theta}}_c^\alpha$ and evaluated at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ gives

$$W_{n,\alpha}(\boldsymbol{\theta}_0) = W_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha}) + \left(\frac{H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_c^{\alpha}} \left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right) + \frac{1}{2}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right)^T \left(\frac{\partial^2 W_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\partial\boldsymbol{\theta}^T\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_c^{\alpha}} \left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right) + o\left(\left\|\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right\|^2\right).$$

But

$$\frac{W_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = (\alpha+1)\int_{\mathbb{R}^m} \mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{y} - (\alpha+1)\frac{1}{n}\sum_{k=1}^{n}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}_k)^{\alpha}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_k)$$

therefore

$$\left(\frac{W_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_c^{\alpha}} \underset{n\to\infty}{\overset{P}{\to}} \mathbf{0}.$$

On the other hand

$$\frac{\partial^2 W_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T} = (1+\alpha)\left\{\int_{\mathbb{R}^m}(1+\alpha)\,\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})^T\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{y} + \int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\frac{\partial\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}^T}d\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n}\alpha\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}_i)^{\alpha}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i)^T\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i) - \frac{1}{n}\sum_{i=1}^{n}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}_i)^{\alpha}\frac{\partial\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}^T}\right\}.$$

But

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}_i)^{\alpha}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i)^T\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i) \underset{n\to\infty}{\overset{P}{\to}} \int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})^T\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})\,d\boldsymbol{y}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y}_i)^{\alpha}\frac{\partial\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}^T} \underset{n\to\infty}{\overset{P}{\to}} \int_{\mathbb{R}^m}\mathcal{CL}(\boldsymbol{\theta},\boldsymbol{y})^{\alpha+1}\frac{\partial\boldsymbol{u}(\boldsymbol{\theta},\boldsymbol{y})}{\partial\boldsymbol{\theta}^T}d\boldsymbol{y}.$$

Therefore

$$\left(\frac{\partial^2 H_{n,\alpha}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_c^{\alpha}} \underset{n\to\infty}{\overset{P}{\to}} (1+\alpha)\,\boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0).$$

We can now write

$$nW_{n,\alpha}(\boldsymbol{\theta}_0) = nW_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha}) + \frac{(1+\alpha)}{2}\sqrt{n}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right)^T \boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0)\sqrt{n}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right) + o_p(1).$$

It is clear that

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[\sqrt{n}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right)^T \boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0)\sqrt{n}\left(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_c^{\alpha}\right)\right] = \sum_{r=1}^{k}\lambda_r + o_p(1) = trace\left(\boldsymbol{J}_{\alpha}(\boldsymbol{\theta}_0)\boldsymbol{H}_{\alpha}(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1).$$

Then

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_{n,\alpha}(\boldsymbol{\theta}_0)\right] = E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha})\right] + \frac{(1+\alpha)}{2} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1)$$

and, on the other hand, it is clear that

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[W_{n,\alpha}(\boldsymbol{\theta}_0)\right] = W_\alpha(\boldsymbol{\theta}_0).$$

Therefore,

$$\begin{aligned}
E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_\alpha(\widehat{\boldsymbol{\theta}}_c^{\alpha})\right] &= nW_\alpha\left(\boldsymbol{\theta}_0\right) + \frac{(1+\alpha)}{2} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1) \\
&= E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_{n,\alpha}\left(\boldsymbol{\theta}_0\right)\right] + \frac{(1+\alpha)}{2} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1) \\
&= E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha})\right] + \frac{(1+\alpha)}{2} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) \\
&+ \frac{(1+\alpha)}{2} trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1) \\
&= E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha})\right] + (1+\alpha)\, trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right) + o_p(1).
\end{aligned}$$

Hence $nW_{n,\alpha}(\widehat{\boldsymbol{\theta}}_c^{\alpha}) + (1+\alpha)\, trace\left(\boldsymbol{J}_\alpha(\boldsymbol{\theta}_0)\boldsymbol{H}_\alpha(\boldsymbol{\theta}_0)^{-1}\right)$ is an asymptotic unbiased estimator of

$$E_{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n}\left[nW_\alpha(\widehat{\boldsymbol{\theta}}_c^{\alpha})\right].$$

□

Appendix B. Computation of the CLDIC in Section 4.1

We have to compute

$$CLDIC\left(M_k\right) = W_{n,\alpha}\left(\widehat{\rho}\right) + \frac{\alpha+1}{n}\frac{J_\alpha\left(\widehat{\rho}\right)}{H_\alpha\left(\widehat{\rho}\right)},$$

where

$$W_{n,\alpha}\left(\widehat{\rho}\right) = \int_{\mathbb{R}^4} \mathcal{CL}(\widehat{\rho},\boldsymbol{y})^{\alpha+1} d\boldsymbol{y} - (1-\alpha^{-1})\frac{1}{n}\sum_{i=1}^{n}\mathcal{CL}(\widehat{\rho},\boldsymbol{y}_i)^{\alpha}, \tag{A1}$$

$$J_\alpha(\widehat{\rho}) = \int_{\mathbb{R}^4} \mathcal{CL}(\widehat{\rho},\boldsymbol{y})^{2\alpha+1} u(\widehat{\rho},\boldsymbol{y})^2 d\boldsymbol{y} - \left(\int_{\mathbb{R}^4} \mathcal{CL}(\widehat{\rho},\boldsymbol{y})^{\alpha+1} u(\widehat{\rho},\boldsymbol{y}) d\boldsymbol{y}\right)^2, \tag{A2}$$

$$H_\alpha(\widehat{\rho}) = -\int_{\mathbb{R}^4} \mathcal{CL}(\widehat{\rho},\boldsymbol{y})^{\alpha+1} u(\widehat{\rho},\boldsymbol{y})^2 d\boldsymbol{y}, \tag{A3}$$

for our candidate models, namely, composite normal and composite 4-variate t-distribution. As commented in Section 4.1, we consider a composite likelihood function based on the product of two bivariate distributions with common variance-covariance matrix. It is therefore, necessary in this example, to obtain values (A1), (A2) and (A3) for both composite normal and composite t-distributions. However, as stated in [10], while the sensitivity and variability matrices can be sometimes be evaluated explicitly, it is more usual to use empirical estimates. Following this comment, in the current example, we compute Equations (A1), (A2) and (A3) empirically through the sample data using

$$\widehat{W}_{n,\alpha}(\widehat{\rho}) = \sum_{i=1}^{n} \mathcal{CL}(\widehat{\rho}, \boldsymbol{y}_i)^{\alpha+1} - (1-\alpha^{-1})\frac{1}{n}\sum_{i=1}^{n} \mathcal{CL}(\widehat{\rho}, \boldsymbol{y}_i)^{\alpha},$$
$$\widehat{J}_{\alpha}(\widehat{\rho}) = \sum_{i=1}^{n} \mathcal{CL}(\widehat{\rho}, \boldsymbol{y}_i)^{2\alpha+1} u(\widehat{\rho}, \boldsymbol{y}_i)^2 - \left(\sum_{i=1}^{n} \mathcal{CL}(\widehat{\rho}, \boldsymbol{y}_i)^{\alpha+1} u(\widehat{\rho}, \boldsymbol{y}_i)\right)^2$$
$$\widehat{H}_{\alpha}(\widehat{\rho}) = -\sum_{i=1}^{n} \mathcal{CL}(\widehat{\rho}, \boldsymbol{y}_i)^{\alpha+1} u(\widehat{\rho}, \boldsymbol{y}_i)^2.$$

Now, we obtain the score of the composite likelihood $u(\widehat{\rho}, \boldsymbol{y}_i)$ explicitly for both cases. By equation (A.5) in [12],

$$u^{N}(\widehat{\rho}, \boldsymbol{y}_i) = \frac{\widehat{\rho}}{1-\widehat{\rho}^2}\left[2 + \frac{1}{\widehat{\rho}}(t_{1i}t_{2i} + t_{3i}t_{4i}) \right.$$
$$\left. -\frac{1}{1-\widehat{\rho}^2}\left(t_{1i}^2 - 2\widehat{\rho}t_{1i}t_{2i} + t_{2i}^2\right) - \frac{1}{1-\widehat{\rho}^2}\left(t_{3i}^2 - 2\widehat{\rho}t_{3i}t_{4i} + t_{4i}^2\right)\right],$$

with $t_{ji} = y_{ji} - \mu_j$, $j = 1, \ldots, 4$. On the other hand, we want to compute $u^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i)$.

$$\begin{aligned} u^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i) &= \frac{\partial \mathcal{CL}^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i)}{\partial \widehat{\rho}} = \frac{\partial \log \mathcal{CL}^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i)}{\partial \widehat{\rho}} = \frac{1}{\mathcal{CL}^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i)}\frac{\partial \mathcal{CL}^{t_\nu}(\widehat{\rho}, \boldsymbol{y}_i)}{\partial \widehat{\rho}} \\ &= \frac{1}{f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho}) f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}\left[\frac{\partial}{\partial \widehat{\rho}} f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho}) f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\right] \\ &= \frac{1}{f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho}) f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}\left[\left(\frac{\partial}{\partial \widehat{\rho}} f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\right) f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho}) + f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\left(\frac{\partial}{\partial \widehat{\rho}} f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\right)\right] \\ &= \frac{1}{f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}\left(\frac{\partial}{\partial \widehat{\rho}} f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\right) + \frac{1}{f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}\left(\frac{\partial}{\partial \widehat{\rho}} f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\right). \end{aligned}$$

Now, it can be shown that

$$\frac{\partial f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}{\partial \widehat{\rho}} = f_{12}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\frac{\nu\left[(\nu-2)\widehat{\rho}^3 - t_{1i}t_{2i}\nu\widehat{\rho}^2 + \left((t_{1i}^2 + t_{2i}^2 - 1)\nu + t_{2i}^2 + t_{1i}^2 + 2\right)\widehat{\rho} - t_{1i}t_{2i}\nu - 2t_{1i}t_{2i}\right]}{(1-\widehat{\rho}^2)\left[(\nu-2)\widehat{\rho}^2 + 2t_{1i}t_{2i}\widehat{\rho} - \nu - t_{1i}^2 - t_{2i}^2 + 2\right]}$$

and

$$\frac{\partial f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})}{\partial \widehat{\rho}} = f_{34}^{t_\nu}(\boldsymbol{y}_i;\widehat{\rho})\frac{\nu\left[(\nu-2)\widehat{\rho}^3 - t_{3i}t_{4i}\nu\widehat{\rho}^2 + \left((t_{3i}^2 + t_{4i}^2 - 1)\nu + t_{4i}^2 + t_{3i}^2 + 2\right)\widehat{\rho} - t_{1i}t_{4i}\nu - 2t_{3i}t_{4i}\right]}{(1-\widehat{\rho}^2)\left[(\nu-2)\widehat{\rho}^2 + 2t_{3i}t_{4i}\widehat{\rho} - \nu - t_{3i}^2 - t_{4i}^2 + 2\right]}.$$

References

1. Fearnhead, P.; Donnelly, P. Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2002**, *64*, 657–680. [CrossRef]
2. Renard, ; D.; Molenberghs, G.; Geys, H. A pairwise likelihood approach to estimation in multilevel probit models. *J. Comput. Stat. Data Anal.* **2004**, *44*, 649–667. [CrossRef]
3. Hjort, N.L.; Omre, H. Topics in spatial statistics. *Scand. J. Stat.* **1994**, *21*, 289–357.
4. Heagerty, P.J.; Lele, S.R. A composite likelihood approach to binary spatial data. *J. Am. Stat. Assoc.* **1998**, *93*, 1099–1111. [CrossRef]
5. Varin, C.; Host, G.; Skare, O. Pairwise likelihood inference in spatial generalized linear mixed models. *Comput. Stat. Data Anal.* **2005**, *49*, 1173–1191 [CrossRef]
6. Henderson, R.; Shimakura, S. A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **2003**, *90*, 355–366. [CrossRef]
7. Parner, E.T. A composite likelihood approach to multivariate survival data. *Scand. J. Stat.* **2001**, *28*, 295–302. [CrossRef]
8. Li, Y.; Lin, X. Semiparametric Normal Transformation Models for Spatially Correlated Survival Data. *J. Am. Stat. Assoc.* **2006**, *101*, 593–603. [CrossRef]
9. Joe, H.; Reid, N.; Somg, P.X.; Firth, D.; Varin, C. Composite Likelihood Methods. Report on the Workshop on Composite Likelihood. 2012. Available online: http://www.birs.ca/events/2012/5-day-workshops/12w5046 (accessed on 23 July 2019).
10. Varin, C.; Reid, N.; Firth, D. An overview of composite likelihood methods. *Statist. Sin.* **2011**, *21*, 5–42.
11. Martín, N.; Pardo, L.; Zografos, K. On divergence tests for composite hypotheses under composite likelihood. *Stat. Pap.* **2019**, *60*, 1883–1919. [CrossRef]
12. Castilla, E.; Martin, N.; Pardo, L.; Zografos, K. Composite Likelihood Methods Based on Minimum Density Power Divergence Estimator. *Entropy* **2018**, *20*, 18. [CrossRef]
13. Castilla, E.; Martin, N.; Pardo, L.; Zografos, K. Composite likelihood methods: Rao-type tests based on composite minimum density power divergence estimator. *Stat. Pap.* **2019**. [CrossRef]
14. Kullback, S. *Information Theory and Statistics*; Wiley: New York, NY, USA, 1959.
15. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*; Petrov, B.N., Csaki, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281 .
16. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
17. Takeuchi, K. Distribution of information statistics and criteria for adequacy of models. *Math. Sci.* **1976**, *153*, 12–18. (In Japanese)
18. Murari, A.; Peluso, E.; Cianfrani, F.; Gaudio, P.; Lungaroni, M. On the Use of Entropy to Improve Model Selection Criteria. *Entropy* **2019**, *21*, 394. [CrossRef]
19. Mattheou, K.; Lee, S.; Karagrigoriou, A. A model selection criterion based on the BHHJ measure of divergence. *J. Stat. Plan. Inference* **2009**, *139*, 228–235. [CrossRef]
20. Avlogiaris, G.; Micheas, A.; Zografos, K. A criterion for local model selection. *Shankhya* **2019**, *81*, 406–444. [CrossRef]
21. Avlogiaris, G.; Micheas, A.; Zografos, K. On local divergences between two probability measures. *Metrika* **2016**, *79*, 303–333. [CrossRef]
22. Varin, C.; Vidoni, P. A note on composite likelihood inference and model selection. *Biometrika* **2005**, *92*, 519–528. [CrossRef]
23. Gao, X.; Song, P.X.K. Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Am. Stat. Assoc.* **2010**, *105*, 1531–1540. [CrossRef]
24. Ng, C.T.; Joe, H. Model comparison with composite likelihood information criteria. *Bernoulli* **2014**, *20*, 1738–1764. [CrossRef]
25. Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **1998**, *85*, 549–559. [CrossRef]
26. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall CRC Press: Boca Raton, FL, USA, 2006.

27. Basu, A.; Shioya, H.; Park, C. *Statistical Inference. The Minimum Distance Approach*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011.
28. Burham, K.P.; Anderson, D.R. *Model Selection and Multinomial Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.
29. Xu, X., Reid, N. On the robustness of maximum composite estimate. *J. Stat. Plan. Inference* **2011**, *141*, 3047–3054. [CrossRef]
30. Warwick, J.; Jones, M.C. Choosing a robustness tuning parameter. *J. Stat. Comput. Simul.* **2005**, *75*, 581–588. [CrossRef]
31. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugenics.* **1936**, *7*, 179–188. [CrossRef]
32. Fraley, A.; Raftery, E.; Murphy, T.B.; Scrucca, L. *MCLUST Version 4 for R: Normal Mixture Modeling for Model-based Clustering, Classification, and Density Estimation*; Technical Report 597; Department of Statistics, University of Washington: Seattle, WA, USA, 2012.
33. Forina, M.; Lanteri, S.; Armanino, C.; Leardi, R. PARVUS: An *Extendable Package of Programs for Data Exploration, Classification, and Correlation*; Institute of Pharmaceutical and Food Analysis Technologies: Genoa, Italy, 1998.