Sankhyā : The Indian Journal of Statistics
2019, Volume 81-A, Part 2, pp. 406-444

A Criterion for Local Model Selection

G. Avlogiaris and K. Zografos
University of Ioannina, Ioannina, Greece
A. C. Micheas
University of Missouri, Columbia, USA

Abstract

In this paper, we introduce a class of local divergences between two probability distributions and illustrate its usefulness in model selection. Explicit expressions of the proposed local divergences are derived when the underlying distributions are members of the exponential family of distributions or they are described by multivariate normal models. In addition, a local model selection criterion, termed the local divergence information criterion ($LDiv.IC$), is proposed. Simulations and applications are presented in order to study and exemplify the performance of the proposed criterion.

AMS (2000) *subject classification.* Primary 62B10; Secondary 62F99.
Keywords and phrases. Model selection, AIC, Local divergence information criterion, Local model selection criterion, Local expected overall discrepancy, Local BHHJ power divergence, Mixture models, Point process theory

1 Introduction

Model selection criteria provide asystematic and rigorous method that allows statisticians to choose the most appropriate model from a collection of possible models used to describe the data. The construction of such criteria requires the creation of a measure of similarity between two entertained 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 smaller the value of the criterion is, the more preferable the model is.

A rich class of similarity measures can be created using ϕ-divergence measures (Csiszár, 1963, 1967). Several cases of these measures have been utilized in the creation of model selection criteria. In particular, the well-known Kullback and Leibler (1951) measure of divergence was used by Akaike (1973) in order to develop the Akaike information criterion (AIC). Since Akaike's pioneering work, there has been a vast literature on the construction of model selection criteria. We refer to Schwarz (1978), Konishi and Kitagawa (1996), Spiegelhalter et al. (2002), Seghouane and Bekara (2004), Cavanaugh (2004), Bengtsson and Cavanaugh (2006), Shang and Cavanaugh (2008), Shang (2008), Mattheou et al. (2009), Toma and Broniatowski (2011) and Toma (2014), and the references therein for the development and illustration of many classic model selection criteria. An alternative approach to these classic methods was presented in Claeskens and Hjort (2003), where the authors allowed different methods to be selected for different parameters of interest. Finally, the book by Claeskens and Hjort (2008) and the references therein provides an exhaustive discussion of model selection criteria.

The standard approach to creating a criterion proceeds as follows; consider the measurable space $(\mathcal{X}, \mathcal{A})$ and let $\mathcal{F} = \{F_\theta\}$ be a parametric family of probability measures on $(\mathcal{X}, \mathcal{A})$, indexed by the parameter $\theta \in \Theta \subseteq R^k$, with $k \geq 1$. Let $\mathcal{G}$ be the class of probability measures G on $(\mathcal{X}, \mathcal{A})$, dominated by a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$, and let $g = \frac{dG}{d\mu}$ denote the Radon-Nikodym derivative of G with respect to μ. For $\theta \in \Theta$, let $F_\theta << \mu$ and $f_\theta = \frac{dF_\theta}{d\mu}$ denote the corresponding Radon-Nikodym derivative. Similarly, for a known $\omega \in \Theta^* \subseteq R^M$, let $\{H_\omega\}$ be a parametric family of probability measures on $(\mathcal{X}, \mathcal{A})$ with $H_\omega << \mu$ and denote by $h_\omega = \frac{dH_\omega}{d\mu}$ the respective Radon-Nikodym derivative. Following Avlogiaris et al. (2016a), the class of local ϕ-divergences between g and f_θ, driven by h_ω, is given by

$$D_\phi^\omega(g, f_\theta) = \int\limits_{\mathcal{X}} h_\omega(x) f_\theta(x) \phi\left(\frac{g(x)}{f_\theta(x)}\right) d\mu(x). \tag{1.1}$$

As defined, the local ϕ-divergence is a measure of similarity between a member of the family $\mathcal{G}$ and a member of the family $\mathcal{F}$, and it is driven by another measure (a kernel) that determines the weights and the area over which the measure is calculated. The distribution h_ω can be chosen in such a way as to smooth or exemplify certain features of the area over which the integral is computed, and it does not necessarily belong to the parametric family $\mathcal{F} = \{F_\theta\}$ (cf. Avlogiaris et al., 2016a).

In order to avoid degenerate cases in definition (1.1), we restrict our interest to real convex functions ϕ which are defined on the interval $[0, \infty)$ and belong to the class of convex functions

$$\begin{aligned}\Phi &= \left\{\phi : \phi \text{ is strictly convex at } 1, \phi(1) = \phi'(1) = 0, 0\phi\left(\frac{0}{0}\right) = 0, 0\phi\left(\frac{u}{0}\right)\right.\\ &\qquad \left.= u \lim_{v\to\infty} \frac{\phi(v)}{v}\right\}. \end{aligned} \tag{1.2}$$

Based on Avlogiaris et al. (2016a), the local ϕ-divergence, defined by (1.1) with $\phi \in \Phi$, satisfies the key property

$$D_\phi^\omega(g, f_\theta) \geq 0, \text{ with equality, if and only if } g = f_\theta, \tag{1.3}$$

regardless of h_ω, and therefore, $D_\phi^\omega(g, f_\theta)$ can be used as a measure of divergence between g and f_θ, in the area of their joint domain which is specified by h_ω. It should be noted at this point that the set Φ contains important cases of Csiszár's ϕ-divergences, like Kullback and Leibler (1951) divergence ($\phi(u) = u \log u - u + 1,\ u > 0$), Kagan (1963) divergence ($\phi(u) = (u-1)^2, u > 0$), Cressie and Read (1984) λ-power divergence $\left(\phi_\lambda(u) = \frac{u^{\lambda+1} - u - \lambda(u-1)}{\lambda(\lambda+1)}, \lambda \neq 0, -1\right)$, and many more (cf. Avlogiaris et al., 2016a).

A general measure of divergence between g and f_θ was introduced in Basu, Harris, Hjort, and Jones (BHHJ) (cf. Basu et al., 1998), defined by

$$D_a(g, f_\theta) = \int\limits_{\mathcal{X}} \left(f_\theta^{1+a}(x) - \left(1 + \frac{1}{a}\right) g(x) f_\theta^a(x) + \frac{1}{a} g^{1+a}(x)\right) d\mu(x), \tag{1.4}$$

where $a > 0$ is the index parameter. The limit of the BHHJ family, when $a \to 0$, is the Kullback–Leibler divergence and for $a = 1$ the divergence in (1.4) reduces to the square of the standard L_2 distance between g and f_θ. We focus on the BHHJ measures of divergence because their functional expression is helpful to our work. In particular, in view of the BHHJ power divergence defined in (1.4), its local version can be immediately obtained from (1.1) for a particular choice of the convex function $\phi_a \in \Phi$ given by

$$\phi_a(u) = u^{1+a} - \left(1 + \frac{1}{a}\right) u^a + \frac{1}{a},\ a > 0. \tag{1.5}$$

As a result, the form of the local BHHJ power divergence is

$$D_a^\omega(g, f_\theta) = \int\limits_{\mathcal{X}} h_\omega(x) \left(f_\theta^{1+a}(x) - \left(1 + \frac{1}{a}\right) g(x) f_\theta^a(x) + \frac{1}{a} g^{1+a}(x)\right) d\mu(x),\ a > 0. \tag{1.6}$$

In order to motivate the results and the necessity for a model selection criterion in a local setting, we consider the following example. Figure 1 illustrates the exponential distribution with parameter $\lambda = 1$ and the log-normal distribution with parameters $\mu = -\ 0.347$, $\eta = 0.833$; these distributions are analyzed and discussed in detail in terms of their local characteristics, in the simulation study of Section 4. This problem has a long history in the statistical literature (see for example Vuong and Wang, 1993; Jiménez-Gamero et al., 2011 and the references therein). It is clear from this Figure that the models under consideration are dissimilar in some subsets of their joint domain whereas they are very close in some other subsets or coincide in the right tail. Therefore, an experimenter might reach the wrong conclusions about the usefulness of a model depending on where they choose to focus their attention. Typically, the comparison takes place over the whole domain of observation (globally) and parsimonious models that would provide a good fit to the data locally (e.g., in the tails) are rejected. Therefore, it is appealing to develop a method that selects the best model, among various available candidate models, in some areas of $\mathcal{X}$.

The latter can be achieved by using a divergence measure between the true model g and a candidatemodel f_θ from a parametric family of models

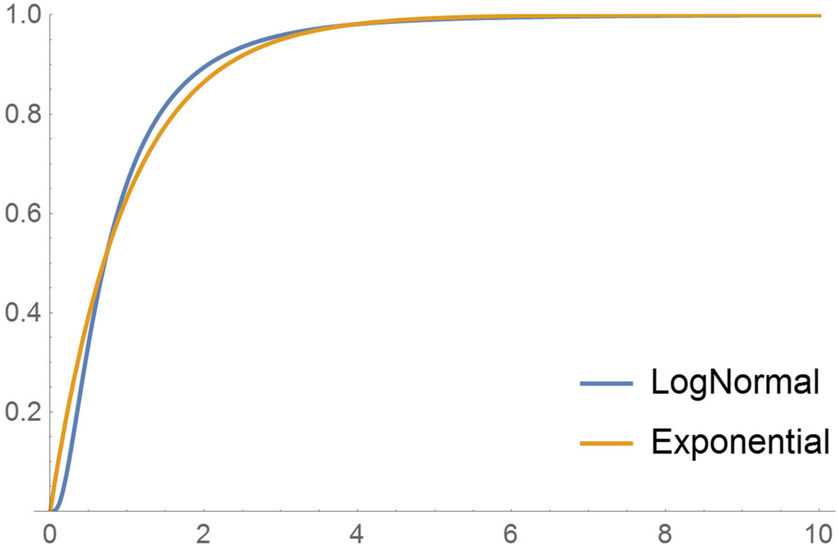

Figure 1: Cumulative distributions functions for log normal with $\mu = -0.347$ and $\eta = 0.833$ and exponential with $\lambda = 1$ distributions

$\mathcal{F}$, that focuses on a specific subset of $\mathcal{X}$. In particular, we need to choose an appropriate driving density h_ω that leads to the desired subset of $\mathcal{X}$. A suitable divergence measure that meets these requirements is the local measure of divergence (1.1), defined in Avlogiaris et al. (2016a) and further applied in order to develop tests of hypotheses in a local setting in Avlogiaris et al. (2016b). If the true distribution g belongs to the parametric family $\mathcal{F} = \{F_\theta\}$, and $\widehat{\theta}$ is a consistent and asymptotically normal estimator of the parameter θ, on the basis of the random sample $X_1, \ldots, X_n$ from $g \in \mathcal{F}$, then the nonnegative quantity $D_\phi^\omega(g, f_{\hat{\theta}})$ can be regarded as the overall discrepancy, which measures the distance between the true but unknown model g and a fitted model $f_{\hat{\theta}}$, in some areas of $\mathcal{X}$. Following Burnham and Anderson (2002), p. 363, among others, we must adopt the following criterion:

$$\text{“select the model } f_{\hat{\theta}} \text{ to minimize } E_g(D_\phi^\omega(g, f_{\hat{\theta}})).\text{”} \tag{1.7}$$

Since $E_g(D_\phi^\omega(g, f_{\hat{\theta}}))$ still depends on the unknown parameter θ, our aim in this paper is to find an unbiased estimator, say $\hat{E}_g(D_\phi^\omega(g, f_{\hat{\theta}}))$, of the expected overall discrepancy $E_g(D_\phi^\omega(g, f_{\hat{\theta}}))$, in some areas of $\mathcal{X}$. Clearly, if the value of $\hat{E}_g(D_\phi^\omega(g, f_{\hat{\theta}}))$ is small then the entertained model should be preferred locally, that is, the entertained model should be selected in the subset of $\mathcal{X}$, where the kernel density h_ω drives the mass of the integral $D_\phi^\omega(g, f_{\hat{\theta}})$ for specific values of $\omega \in \Theta^*$.

A comment is in order about the choice of the kernel density h_ω. Clearly, a normal kernel with some mean and variance is a standard choice, with the values chosen based on expert opinions of experimenters (when available), depending on where they want to focus their attention and the questions they need to answer. In the absence of such input, one way is to choose a rolling window across the domain as illustrated in our simulations and applications. In theory, any type of kernel may be chosen, including normal, uniform, or triangular, provided that h_ω is a proper density.

A more systematic way of selecting the kernel parameters can be constructed by considering the areas of the domain where the entertained model g is the least or most favorable to provide a good fit to the model. That is, given a suitable estimator $\widehat{g}$ of the true model g (e.g., a non-parametric density smoother), we find the values of the kernel parameters such that $\omega_{least} = argmax D_\phi^\omega(\widehat{g}, f_{\hat{\theta}})$ and $\omega_{most} = argmin D_\phi^\omega(\widehat{g}, f_{\hat{\theta}})$, and then report on the support of the kernels $h_{\omega_{least}}$ and $h_{\omega_{most}}$ as the areas of the domain where we achieve the best and the worst model fits. We will not present this approach here since in the cases we consider for our simulations or applications, we can simply inspect a scatter plot of the data and then choose the kernel parameters that focus on a specific window of the domain of observation. However, this approach can be particularly useful in the case of multivariate data with more than two variables, where visual inspection of the domain of observation is not feasible.

Based on the above discussion, the aim of this paper is twofold; first we introduce a measure of the local divergence between two probability distributions based on the BHHJ power divergence. On the other hand, we utilize this local measure of divergence in order to develop a local model selection criterion. In particular, the local BHHJ power divergences are introduced in Section 2, including the derivation of explicit forms of the local BHHJ power divergence between members of the exponential family of distributions. The case of multivariate normal distributions is also considered and we discuss further extensions to mixtures of normal components. In Section 3, local model selection criteria are developed using the local BHHJ power divergence. Simulations are presented in Section 4 in order to evaluate the performance of the proposed model selection criterion in a local setting. Section 5 presents two applications of the proposed methodology by analyzing real data sets. Some concluding remarks are given in Section 6. Section 7 provides the proofs of the theoretic results of the paper.

2 Local BHHJ Power Divergence and Its Explicit Form for the Exponential Family

The following proposition characterizes the lower bound of the local BHHJ power divergence and it is established easily using Theorem 1, parts (a) and (b) in Avlogiaris et al. (2016a).

Proposition 1. *The quantity $D_a^\omega(g, f_\theta)$ is a measure of divergence between the two models g and f_θ, in the sense that (i) $D_a^\omega(g, f_\theta) \geq 0$, for all g, f_θ and (ii) $D_a^\omega(g, f_\theta) = 0$ if and only if $g = f_\theta$ a.e., regardless of the choice of driving density h_ω.*

Explicit expressions for the local BHHJ divergence in (1.6) are derived in the rest of this section for the case of a parametric family $\mathcal{F} = \{F_\theta\}$ with members from the exponential family

$$f_C(x, \theta) = \exp\left\{\theta^t T(x) - C(\theta) + w(x)\right\}, \; x \in \mathcal{X}, \tag{2.1}$$

with natural parameters $\theta \in \Theta \subseteq R^k$ and $T(x) = (T_1(x), \ldots, T_k(x))^t, x \in \mathcal{X}$, where the superscript ^{t} is used to denote the transpose of a vector or a matrix.

For two members of this family, $f_C(x, \theta_i)$, $\theta_i \in \Theta \subseteq R^k$, $i = 1, 2$, and taking into account (1.6), the local BHHJ power divergence is defined by

$$D_a^{\omega}(\theta_1, \theta_2) = E_{\theta_2}\left(h_\omega(X) f_C^a(X, \theta_2)\right) - \left(1 + \frac{1}{a}\right) K_{a,\omega}(\theta_1, \theta_2) + \frac{1}{a} E_{\theta_1}\left(h_\omega(X) f_C^a(x, \theta_1)\right), \tag{2.2}$$

for $a > 0$, with

$$K_{a,\omega}(\theta_1, \theta_2) = \int\limits_{\mathcal{X}} h_\omega(x) f_C(x, \theta_1) f_C^a(x, \theta_2) d\mu(x), \tag{2.3}$$

$$E_{\theta_i}\left(h_\omega(X) f_C^a(X, \theta_i)\right) = \int\limits_{\mathcal{X}} h_\omega(x) f_C^{a+1}(x, \theta_i) d\mu(x), \tag{2.4}$$

and $\theta_i \in \Theta \subseteq R^k$, $i = 1, 2$. Note that equations (2.2)–(2.4) depend on the general form of the model $f_C(x, \theta)$, and therefore, one can replace $f_C(x, \theta)$ with any model other than the exponential family model of equation (2.1). In particular, in some of the applications that follow, we use a mixture of normal components as the entertained model, e.g., $f_C(x, \theta) = \sum_{j=1}^m p_j \phi_j(x; \mu_j, \Sigma_j)$, where $p = (p_1, p_2, \ldots, p_m)$ the component probabilities, with $\sum_{j=1}^m p_j = 1$, $p_j \geq 0$, $\mu_j \in R^d$ the j^{th} component mean and Σ_j the j^{th} component covariance matrix, $j = 1, 2, \ldots, m$. Then calculation of $D_a^{\omega}(\theta_1, \theta_2)$ is straightforward using Monte Carlo integration or quadrature rules. The theoretical development that follows is based on the model of equation (2.1) with explicit forms for the measure and the local model selection criterion. In addition, all quantities of interest under the mixture model will be obtained using the Monte Carlo integration approach.

Consider a kernel density h_ω defined on $\mathcal{X}$, that does not necessarily belong to the class of densities (2.1). In view of (2.2), straightforward calculations lead to the following result.

Proposition 2. *Let the kernel density h_ω be defined on $\mathcal{X}$ and consider two members $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$ of* (2.1)*. If $\theta_1+a\theta_2 \in \Theta$, for $a>0$, then the BHHJ local power divergence between $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$, driven by the density h_ω, is given by*

$$\begin{aligned} D_a^\omega(\theta_1,\theta_2) \quad = \quad & E_{\theta_2}\left(h_\omega(X) f_C^a(X,\theta_2)\right) - (1+a^{-1}) \\ & \times \exp\{M_{C,a}^{(1)}(\theta_1,\theta_2)\} E_{\theta_1+a\theta_2}(h_\omega(X)\exp\{aw(X)\}) \\ & + a^{-1} E_{\theta_1}\left(h_\omega(X) f_C^a(X,\theta_1)\right), \end{aligned} \tag{2.5}$$

with

$$M_{C,a}^{(1)}(\theta_1,\theta_2) = C(\theta_1+a\theta_2) - C(\theta_1) - aC(\theta_2), \tag{2.6}$$

where $E_{\theta_1+a\theta_2}\left(h_\omega(X)\exp\{aw(X)\}\right)$, $E_{\theta_i}\left(h_\omega(X) f_C^a(X,\theta_i)\right)$, $i=1,2$, are defined by (2.4)*.*

The proposition that follows presents the analytic expression for $D_a^\omega(\theta_1,\theta_2)$ when the kernel density h_ω belongs to the class of densities (2.1). The proof is omitted as it follows similar steps as those in Appendix B of Avlogiaris et al. (2016a).

Proposition 3. *Consider two members $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$ of* (2.1) *and assume that the kernel density $h_\omega(x) = f_C(x,\omega)$ is as in* (2.1)*. Then, subject to the assumptions $a\theta_i+\omega \in \Theta$, $i=1,2$ and $\theta_1+a\theta_2+\omega \in \Theta$, for $a>0$, the BHHJ local power divergence between $f_C(x,\theta_1)$ and $f_C(x,\theta_2)$, driven by $h_\omega(x)$, is given by*

$$\begin{aligned} D_a^\omega(\theta_1,\theta_2) \quad = \quad & \exp\{C(a\theta_2+\omega) - aC(\theta_2) - C(\omega)\} E_{a\theta_2+\omega}\left(\exp\{aw(X)\}\right) \\ & -(1+a^{-1})\exp\{M_{C,a}^{(2)}(\theta_1,\theta_2,\omega)\} E_{\theta_1+a\theta_2+\omega}(\exp\{(a+1)w(X)\}) \\ & + a^{-1}\exp\{C(a\theta_1+\omega) - aC(\theta_1) - C(\omega)\} E_{a\theta_1+\omega}\left(\exp\{aw(X)\}\right), \end{aligned} \tag{2.7}$$

where

$$M_{C,a}^{(2)}(\theta_1,\theta_2,\omega) = C(\theta_1+a\theta_2+\omega) - C(\theta_1) - aC(\theta_2) - C(\omega), \tag{2.8}$$

with

$$E_{a\theta_i+\omega}(\exp\{aw(X)\}) = \int_{\mathcal{X}} \exp\{aw(X)\} f_C(x, a\theta_i+\omega) d\mu(x), \;\; i=1,2, \tag{2.9}$$

and

$$E_{\theta_1+a\theta_2+\omega}(\exp\{(a+1)w(X)\}) = \int\limits_{\mathcal{X}} \exp\{(a+1)w(X) f_C(x,\theta_1+a\theta_2+\omega) d\mu(x). \tag{2.10}$$

Our interest is now focused on the explicit form of the local BHHJ power divergence, defined by (1.6), between two d-variate normal distributions, $N_d(\mu_1,\Sigma_1)$ and $N_d(\mu_2,\Sigma_2)$. Let the kernel density h_ω be the multivariate normal distribution $N_d(\mu,\Sigma)$ with mean vector $\mu \in R^d$ and covariance matrix Σ.

The densities of the d-variate normal models with mean vectors $\mu_i \in R^d$ and covariance matrices Σ_i, $i = 1, 2$, are given by

$$f(x;\mu_i,\Sigma_i) = (2\pi)^{-d/2}|\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)\right),\ i = 1, 2.$$

It is well known that the above d-variate normal distributions are included in the exponential family of distributions (2.1) with

$$\begin{aligned} \theta_i = (\theta_{i1},\theta_{i2}) &= \left(\Sigma_i^{-1}\mu_i, -\frac{1}{2}\Sigma_i^{-1}\right), \\ T(x) = (T_1(x), T_2(x)) &= \left(x, xx^t\right),\ w(x) = 0, \\ C(\theta_i) = \log\left((2\pi)^{d/2}|\Sigma_i|^{1/2}\right) &+ \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i = \log(2\pi)^{d/2} \\ &-\frac{1}{2}\log\left(|-2\theta_{i2}|\right) - \frac{1}{4}\theta_{i1}^t\theta_{i2}^{-1}\theta_{i1}, \end{aligned} \tag{2.11}$$

where $|\ .\ |$ is used to denote the determinant of a matrix. It should be noted that the inner product of $\alpha = (u, M)$ and $\beta = (v, N)$, which consists of two parts, a vector part u and v and a matrix part M and N, is defined by $\alpha^t\beta = u^t v + trace(M^t N)$ (cf. Nielsen and Nock (2011, p. 6)).

Proposition 4. *The local BHHJ power divergence, defined by* (1.6), *between two d-variate normal distributions* $N_d(\mu_1,\Sigma_1)$ *and* $N_d(\mu_2,\Sigma_2)$*, driven by a d-variate normal distribution* $N_d(\mu,\Sigma)$*, is given by*

$$
\begin{aligned}
&D_a^{(\mu,\Sigma)}((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)) = (2\pi)^{-ad/2}|\Sigma^{-1}+a\Sigma_2^{-1}|^{-1/2}|\Sigma|^{-1/2}|\Sigma_2|^{-a/2}\\
&\times\exp\left\{-\frac{1}{2}(\mu-\mu_2)^t(\Sigma+a\Sigma_2)^{-1}(\mu-\mu_2)\right\}\\
&-(1+a^{-1})(2\pi)^{-(a+1)d/2}|\Sigma|^{-\frac{1}{2}}|\Sigma_1|^{-\frac{1}{2}}|\Sigma_2|^{-\frac{a}{2}}\left|\Sigma_1^{-1}+a\Sigma_2^{-1}+\Sigma^{-1}\right|^{-\frac{1}{2}}\\
&\times\exp\left\{-\frac{1}{2}\left(\mu^t\Sigma^{-1}\mu+\mu_1^t\Sigma_1^{-1}\mu_1+a\mu_2^t\Sigma_2^{-1}\mu_2-B_1^tB_2B_1\right)\right\}\\
&+a^{-1}(2\pi)^{-ad/2}|\Sigma^{-1}+a\Sigma_1^{-1}|^{-1/2}|\Sigma|^{-1/2}|\Sigma_1|^{-a/2}\\
&\times\exp\left\{-\frac{1}{2}(\mu-\mu_1)^t(\Sigma+a\Sigma_1)^{-1}(\mu-\mu_1)\right\},
\end{aligned}
$$

with

$$
\begin{aligned}
B_1 &= \Sigma_1^{-1}\mu_1+a\Sigma_2^{-1}\mu_2+\Sigma^{-1}\mu,\\
B_2 &= \left(\Sigma_1^{-1}+a\Sigma_2^{-1}+\Sigma^{-1}\right)^{-1},
\end{aligned}
$$

for $a>0$.

The proof is omitted as it follows similar steps as those in Appendix C of Avlogiaris et al. (2016a).

3 Local Model Selection Criterion

In this section, we construct a local model selection criterion by means of the local BHHJ divergence defined by (1.6), applying the same methodology as the one used in the construction of the AIC (cf. Burnham and Anderson, 2002) and the Takeuchi information criterion (cf. Takeuchi, 1976). For the mathematical derivation of the local model selection information criterion ($LDiv.IC$), consider a random sample $X_1,\dots,X_n$ from some distribution G defined on a measurable space $(\mathcal{X},\mathcal{A})$, with true but unknown density function g, and a candidate model F_θ, with density function f_θ, from a parametric family of probability measures $\mathcal{F}=\{F_\theta\}$ (or densities $\{f_\theta\}$) in $(\mathcal{X},\mathcal{A})$, indexed by an unknown parameter $\theta\in\Theta\subseteq R^k$, with $k\geq 1$.

In order to construct the local criterion for a known $\omega\in\Theta^*$, consider a probability measure H_ω in $(\mathcal{X},\mathcal{A})$ with respective density h_ω and the quantity

$$
W_a^\omega(\theta)=\int\limits_{\mathcal{X}} h_\omega(x)f_\theta^{1+a}(x)d\mu(x)-(1+a^{-1})\int\limits_{\mathcal{X}} h_\omega(x)g(x)f_\theta^a(x)d\mu(x),\ a>0, \tag{3.1}
$$

which can be written as

$$W_a^{\omega}(\theta) = D_a^{\omega}(g, f_\theta) - \frac{1}{a}\int\limits_{\mathcal{X}} h_\omega(x) g^{1+a}(x) d\mu(x),\ a > 0, \tag{3.2}$$

where $D_a^{\omega}(g, f_\theta)$ is the local BHHJ power divergence, defined by (1.6). We notice that the term $\frac{1}{a}\int\limits_{\mathcal{X}} h_\omega(x) g^{1+a}(x)\ d\mu(x)$ remains constant regardless of the model f_θ used. Therefore, this quantity can be regarded as a measure of the local discrepancy between g and f_θ, which differs from the local BHHJ power divergence only up to a constant (see Burnham and Anderson, 2002, p. 364). Note that (3.1) can also be written as

$$W_a^{\omega}(\theta) = E_{f_\theta}(h_\omega(X) f_\theta^{\alpha}(X)) - (1 + a^{-1}) E_g(h_\omega(X) f_\theta^{\alpha}(X)),\ a > 0. \tag{3.3}$$

We assume that the true model g belongs to the parametric family $\{f_\theta\}$, and consider a consistent and asymptotically normal estimator of the parameter θ based on the sample $X_1, \ldots, X_n$ from the true model $g \in \{f_\theta\}$. This estimator possesses the necessary properties required for the derivation of the $LDiv.IC$. Such an estimator can be obtained either by maximizing the loglikelihood function or minimizing the BHHJ divergence (cf. Basu et al., 2011). In the latter case, the consistency, as well as, the asymptotic normality of the estimator are verified by Theorem 4.2 in Basu et al. (1998).

For a given $\omega \in \Theta^*$, we define the *weighted or local expected overall discrepancy* between g and f_θ by

$$\begin{aligned} E_g(W_a^{\omega}(\hat\theta)) &= E_g(W_a^{\omega}(\theta) | \theta = \hat\theta) \\ &= E_g\Big(E_{f_{\hat\theta}}(h_\omega(X) f_{\hat\theta}^{\alpha}(X))\Big) - (1 + a^{-1}) E_g\Big(E_g(h_\omega(X) f_{\hat\theta}^{a}(X))\Big), a > 0, \end{aligned} \tag{3.4}$$

where $\widehat{\theta}$ is any consistent and asymptotically normal estimator of the parameter θ. Our aim here is to construct an asymptotically unbiased estimator of the quantity $E_g(W_a^{\omega}(\hat\theta))$ with $g \in \{f_\theta\}$ (cf. Burnham and Anderson, 2002, p. 363).

We begin by obtaining some lemmas that are necessary in order to obtain the basic proposition of this section, which yields the desired estimator of $E_g(W_a^{\omega}(\hat\theta))$ with $g \in \{f_\theta\}$. Let $\nabla_\theta l_\theta(\cdot)$ denote the k-dimensional gradient vector of $l_\theta(\cdot)$ with respect to θ, and $\nabla_\theta^2 l_\theta(\cdot)$ the corresponding $k \times k$ Hessian Matrix, where $l_\theta(\cdot)$ is a scalar function. In what follows, in order to investigate the asymptotic behavior of the estimators, we consider the standard regularity assumptions of asymptotic statistics (cf. Pardo, 2006, p. 58; Avlogiaris et al., 2016b).

Lemma 1. *The gradient vector and the Hessian matrix of $W_\alpha^\omega(\theta)$ are given by*

$$\nabla_\theta W_a^\omega(\theta) = (a+1)\left(\int_{\mathcal{X}} h_\omega(x) u_\theta(x) f_\theta^{1+a}(x) d\mu(x) - E_g(h_\omega(X) u_\theta(X) f_\theta^a(X))\right), \tag{3.5}$$

and

$$\begin{aligned}\nabla_\theta^2 W_a^\omega(\theta) = (a+1)\Bigg\{&(a+1)\int_{\mathcal{X}} h_\omega(x) u_\theta(x) u_\theta^t(x) f_\theta^{1+a}(x) d\mu(x)\\ &-\int_{\mathcal{X}} h_\omega(x) \Xi_\theta(x) f_\theta^{1+a}(x) d\mu(x)\\ &+E_g(h_\omega(X)\Xi_\theta(X) f_\theta^a(X)) - E_g(a h_\omega(X) u_\theta(X) u_\theta^t(X) f_\theta^a(X))\Bigg\},\end{aligned} \tag{3.6}$$

where $u_\theta(x) = \nabla_\theta \log(f_\theta(x))$ *and* $\Xi_\theta(x) = -\nabla_\theta^2 \log(f_\theta(x))$.

The proof is similar to that of Lemma 2.1 of Mattheou et al. (2009) and will be omitted. An immediate consequence of Lemma 1 is as follows: If the true distribution g belongs to the parametric family $\{f_\theta\}$ and θ_0 represents the true value of the parameter θ, then the gradient vector and the Hessian matrix of the quantity $W_\alpha^\omega(\theta)$ are given by

$$[\nabla_\theta W_a^\omega(\theta)]_{\theta=\theta_0} = 0, \tag{3.7}$$

and

$$[\nabla_\theta^2 W_a^\omega(\theta)]_{\theta=\theta_0} = (a+1)\int_{\mathcal{X}} h_\omega(x) u_{\theta_0}(x) u_{\theta_0}^t(x) f_{\theta_0}^{1+a}(x) d\mu(x) = (a+1) J^\omega(\theta_0), \tag{3.8}$$

with

$$J^\omega(\theta) = \left(\int_{\mathcal{X}} h_\omega(x) f_\theta^{1+a}(x) \frac{\partial \log f_\theta(x)}{\partial \theta_i} \frac{\partial \log f_\theta(x)}{\partial \theta_j} d\mu(x)\right)_{i,j=1,\ldots,k}, \tag{3.9}$$

where $u_\theta(x) = \nabla_\theta \log(f_\theta(x))$.

The next lemma describes how the local expected overall discrepancy $E_g(W_a^\omega(\hat{\theta}))$ can be written via a Taylor expansion.

Lemma 2. *If the true distribution g belongs to the parametric family $\{f_\theta\}$ and θ_0 denotes the true value of the parameter θ, then the local expected overall discrepancy $E_g(W_a^\omega(\hat{\theta}))$ between g and f_θ is given by*

$$E_g(W_a^\omega(\hat{\theta})) = W_a^\omega(\theta_0) + \frac{\alpha+1}{2} E_g((\hat{\theta}-\theta_0)^t J^\omega(\theta_0)(\hat{\theta}-\theta_0)) + E_g\left(R_n\right), \ a > 0, \tag{3.10}$$

where $R_n = o\left(\|\hat{\theta}-\theta_0\|^2\right)$ and $J^\omega(\theta_0)$ as defined by (3.9).

The result is immediately obtained by applying a Taylor series expansion to the function $W_a^\omega(\theta)$ about the true parameter θ_0, taking $\theta = \hat{\theta}$ and using equations (3.7) and (3.8). As a consequence, we obtain

$$W_a^\omega(\hat{\theta}) = W_a^\omega(\theta_0) + \frac{\alpha+1}{2}(\hat{\theta}-\theta_0)^t J^\omega(\theta_0)(\hat{\theta}-\theta_0)) + o\left(\|\hat{\theta}-\theta_0\|^2\right), \ a > 0, \tag{3.11}$$

and therefore, (3.10) is established.

Now, for a given $\omega \in \Theta^*$ and $\theta \in \Theta$, an estimator $Q_a^\omega(\theta)$ of $W_a^\omega(\theta)$ is obtained by replacing $E_g(h_\omega(X) f_\theta^a(X))$ in the expression of $W_a^\omega(\theta)$ in (3.3) by its sample analog. Consequently, the estimator $Q_a^\omega(\theta)$ is given by

$$Q_a^\omega(\theta) = \int\limits_{\mathcal{X}} h_\omega(x) f_\theta^{1+a}(x) d\mu(x) - (1+a^{-1}) \frac{1}{n} \sum_{i=1}^{n} h_\omega(X_i) f_\theta^a(X_i), \ a > 0, \tag{3.12}$$

where $X_1, \ldots, X_n$ is a random sample from g. The gradient vector and the Hessian matrix for $Q_a^\omega(\theta)$ with respect to θ are given in the following Lemma. The proof follows similar steps as in the proof of Lemma 2.2 in Mattheou et al. (2009, p. 231) and will be omitted.

Lemma 3. *The gradient vector and the Hessian matrix of $Q_a^\omega(\theta)$, in (3.12), are given by*

$$\nabla_\theta Q_a^\omega(\theta) = (a+1)\left(\int\limits_{\mathcal{X}} h_\omega(x) u_\theta(x) f_\theta^{1+a}(x) d\mu(x) - \frac{1}{n}\sum_{i=1}^{n} h_\omega(X_i) u_\theta(X_i) f_\theta^a(X_i)\right), \tag{3.13}$$

and

$$\nabla^2_\theta Q^\omega_a(\theta) = (a+1)\Bigg\{(a+1)\int_{\mathcal{X}} h_\omega(x)u_\theta(x)u^t_\theta(x)f^{1+a}_\theta(x)d\mu(x)$$
$$-\int_{\mathcal{X}} h_\omega(x)\Xi_\theta(x)f^{1+a}_\theta(x)d\mu(x)$$
$$+\frac{1}{n}\sum_{i=1}^n h_\omega(X_i)\Xi_\theta(X_i)f^a_\theta(X_i)-\frac{1}{n}\sum_{i=1}^n ah_\omega(X_i)u_\theta(X_i)u^t_\theta(X_i)f^a_\theta(X_i)\Bigg\}, \tag{3.14}$$

where $u_\theta(x) = \nabla_\theta \log(f_\theta(x))$ *and* $\Xi_\theta(x) = -\nabla^2_\theta \log(f_\theta(x))$.

The following lemma provides the last component required in order to prove our main result.

Lemma 4. *Under the assumption that the true model g belongs to the parametric family $\{f_\theta\}$, the local expected overall discrepancy $E_g(W^\omega_a(\hat\theta))$ is given by*

$$E_g(W^\omega_a(\hat\theta)) = E_g\left(Q^\omega_a(\hat\theta) + (a+1)(\hat\theta-\theta_0)^t J^\omega(\theta_0)(\hat\theta-\theta_0) + R_n\right), \tag{3.15}$$

where $R_n = o\left(\|\hat\theta-\theta_0\|^2\right)$, θ_0 is the true parameter and $J^\omega(\theta_0)$ as in (3.9).

The proof of this Lemma is given in Section 7.1. The main result of this section is described below.

Proposition 5. *Under the assumption that the true model g belongs to the parametric family $\{f_\theta\}$, the following hold:*

(i) The local expected overall discrepancy $E_g(W^\omega_a(\hat\theta))$ defined in (3.4), multiplied by n, is given by

$$nE_g(W^\omega_a(\hat\theta)) = nQ^\omega_a(\hat\theta) + n(a+1)(\hat\theta-\theta_0)^t J^\omega(\theta_0)(\hat\theta-\theta_0), \tag{3.16}$$

where $\hat\theta$ is any consistent and asymptotically normal estimator and $Q^\omega_a(\hat\theta)$ is given by (3.12).

(ii) The expected value of the quadratic form $n(\hat\theta-\theta_0)^t J^\omega(\theta_0)(\hat\theta-\theta_0)$ is given by

$$E_g[n(\hat\theta-\theta_0)^t J^\omega(\theta_0)(\hat\theta-\theta_0)] = \sum_{i=1}^r \beta_i, \tag{3.17}$$

where $\hat{\theta}$ is as in Theorem 4.2 in Basu et al. (1998) and $\beta_1, \beta_2, \ldots, \beta_r$ denote the non-zero eigenvalues of the matrix

$$J^{\omega}(\theta_0)AVar(\theta_0),$$

where θ_0 is the true value of the parameter θ and

$$r = rank(AVar(\theta_0)J^{\omega}(\theta_0)AVar(\theta_0)).$$

The quantity $J^{\omega}(\theta_0)$ is defined by (3.9), and

$$AVar(\theta_0) = J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0),$$

denotes the variance-covariance matrix of $\sqrt{n}(\hat{\theta}-\theta_0)$, where $J(\theta_0)$ and $K(\theta_0)$ are given by

$$J(\theta_0) = \int u_{\theta_0}(x)u_{\theta_0}^t(x)f_{\theta_0}^{1+a}(x)d\mu(x),$$

and

$$\begin{aligned} K(\theta_0) \quad &= \quad \int u_{\theta_0}(x)u_{\theta_0}^t(x)f_{\theta_0}^{1+2a}(x)d\mu(x) \\ &\quad - \int u_{\theta_0}(x)f_{\theta_0}^{1+a}(x)d\mu(x) \int u_{\theta_0}^t(x)f_{\theta_0}^{1+a}(x)d\mu(x). \end{aligned}$$

The proof of this Proposition is given in Section 7.2.

Remark 1. (Procedure for local model selection). Taking into account the above discussion, the local divergence information criterion $LDiv.IC$ is defined by

$$L_{a,n}^{\omega}(\hat{\theta}, \beta_1 \ldots \beta_r) = nQ_a^{\omega}(\hat{\theta}) + (a+1)\sum_{i=1}^{r}\beta_i. \tag{3.18}$$

and the general approach to local model selection is as follows: assume that the true distribution g belongs to the parametric family $\{f_\theta\}$, and $\hat{\theta}$ is a consistent and asymptotically normal estimator of the parameter θ, on the basis of the random sample $X_1, \ldots, X_n$ from $g \in \{f_\theta\}$. Then for a given $\omega \in \Theta^*$ and $a > 0$, the value $L_{a,n}^{\omega}(\hat{\theta}, \beta_1 \ldots \beta_r)$, where $Q_a^{\omega}(\hat{\theta})$ is defined in (3.12) and $\beta_1, \beta_2, \ldots, \beta_r$ are as in Proposition 5, provides us with a criterion to choose the most appropriate model from a collection $\{f_\theta\}$ of possible models, in some area of $\mathcal{X}$, which is specified by the density h_ω.

In particular, we measure the value of $LDiv.IC$ between two entertained models $f^1_{\theta_1}$ and $f^2_{\theta_2}$ from the true model g, locally, we compute $L^{\omega}_{a,n}(\hat{\theta}_1, \beta_1 \dots \beta_r)$ and $L^{\omega}_{a,n}(\hat{\theta}_2, a_1 \dots a_s)$ on the basis of a random sample $X_1, \dots, X_n$, and if

$$L^{\omega}_{a,n}(\hat{\theta}_1, \beta_1 \dots \beta_r) < L^{\omega}_{a,n}(\hat{\theta}_2, a_1 \dots a_s), \tag{3.19}$$

then, $f_{\hat{\theta}_1}$ is a more appropriate model than $f_{\hat{\theta}_2}$. This procedure is illustrated in the application section.

Remark 2. When $\hat{\theta}$ is taken to be the MLE, the variance-covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ is the inverse of the Fisher information matrix $I_F^{-1}(\theta_0)$ and $\beta_1, \beta_2, \dots, \beta_r$ denote the non-zero eigenvalues of the matrix

$$J^{\omega}(\theta_0) I_F(\theta_0)^{-1},$$

where

$$r = rank(I_F(\theta_0)^{-1} J^{\omega}(\theta_0) I_F(\theta_0)^{-1}).$$

What distinguishes the MLE from the BHHJ estimator is the fact that it is computationally much faster and much more precise (cf. Mattheou et al., 2009, p. 233). This feature will become apparent in the simulation section.

Remark 3. The limit of the divergence in (1.6) when $a \to 0$ is the local Kullback-Leibler divergence (cf. Avlogiaris et al., 2016a) given by

$$D^{\omega}_0(g, f_\theta) = \lim_{a \to 0} D^{\omega}_a(g, f_\theta) = E_{f_\theta}(h_\omega(X)) - E_g(h_\omega(X)) + \int\limits_{\mathcal{X}} h_\omega(x) g(x) \log \frac{g(x)}{f_\theta(x)} d\mu(x). \tag{3.20}$$

The proof of (3.20) is given in Section 7.3. Therefore, when $a \to 0$, we have

$$W^{\omega}_a(\theta) = E_{f_\theta}(h_\omega(X)) - E_g(h_\omega(X) \log(f_\theta(X))), \tag{3.21}$$

and

$$Q^{\omega}_a(\theta) = E_{f_\theta}(h_\omega(X) - \frac{1}{n} \sum_{i=1}^{n} h_\omega(X_i) \log(f_\theta(X_i)). \tag{3.22}$$

In this case, the $LDiv.IC$ reduces to what we aptly call local Akaike criterion ($LAIC$), given by

$$LAIC = nQ^{\omega}_a(\hat{\theta}) + \sum_{i=1}^{r} \beta_i. \tag{3.23}$$

Remark 4. The matrices $J^{\omega}(\theta_0)$ and $I_F(\theta_0)$ can be estimated by replacing the true but unknown parameter θ_0 with an estimator $\hat{\theta}$ based on the observed data as follows:

$$I_F(\hat{\theta}) = \left(\int_{\mathcal{X}} f_\theta(x) \, \frac{\partial \log f_\theta(x_i)}{\partial \theta_i} \frac{\partial \log f_\theta(x_i)}{\partial \theta_j} d\mu(x) \bigg|_{\theta=\hat{\theta}} \right)_{i,j=1,\dots,k}, \tag{3.24}$$

and

$$J^{\omega}(\hat{\theta}) = \left(\int_{\mathcal{X}} h_\omega(x) f_\theta^{1+a}(x) \, \frac{\partial \log f_\theta(x)}{\partial \theta_i} \frac{\partial \log f_\theta(x)}{\partial \theta_j} d\mu(x) \bigg|_{\theta=\hat{\theta}} \right)_{i,j=1,\dots,k}. \tag{3.25}$$

When the integrals above are intractable, i.e., not available in closed form, then we can employ the empirical observed information matrix to approximate (3.24) (e.g., McLachlan and Peal, 2000, section 2.15.3) and use a similar approach for (3.25). Alternatively, one can approximate both integrals using Monte Carlo integration based on realizations from the estimated model $f_{\hat{\theta}}$.

4 Simulation Study

4.1. Univariate Case In order to evaluate the performance of the proposed local divergence information criterion, we performed a Monte Carlo simulation study using the *LDiv.IC*. We consider the problem of choosing between an exponential model, with density

$$f(x;\lambda) = f_\lambda(x) = \lambda \exp(-\lambda x), \;\; x > 0, \;\; \lambda > 0,$$

and a log-normal model with density

$$k(x;\mu,\eta) = k_{\mu,\eta}(x) = \frac{1}{x\eta\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\eta^2} \right), \;\; x > 0, \;\; \mu \in \mathcal{R}, \;\; \eta > 0,$$

locally. We are interested in investigating any differences between the two models, by focusing on a specific area of the domain of observations where the two models might differ. This problem has a long history in the statistical literature, when the comparison is made globally, i.e., over the whole domain of the distributions. See Vuong and Wang (1993), Jiménez-Gamero et al. (2011), and the references therein. Figure 1 illustrates the exponential distribution with parameter $\lambda = 1$ and the log-normal distribution with parameter $\mu = -0.347$, $\eta = 0.833$. The simulation study based on the *LDiv.IC* has the following characteristics:

- Generate 1000 samples of size $n = 100, 250, 500, 1000$ from
$$h(x;t) = tf(x;1) + (1-t)k(x;-0.347, 0.833),$$
for $t = 0.5, 1, 0, 0.75, 0.25$.
- Estimate the parameters λ, μ, and η of the exponential distributions and the log-normal distribution. The estimators of the parameters are obtained either by minimization of the BHHJ measure defined by (1.4) (and in this case, the parameter η is considered known and equal to 0.833, see Table 1), or by maximization of the loglikelihood function (see Table 2). If the BHHJ method is used, there is no closed form for the estimators of the parameters; instead, they are computed by solving the equations
$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta^a(X_i) - \int\limits_{\mathcal{X}} u_\theta(x) f_\theta^{1+a}(x)dx = 0,$$
numerically (cf. Basu et al., 1998), where $u_\theta(x) = \nabla_\theta \log(f_\theta(x))$, for $a = 0.1$.
- Calculate the value of the $LDiv.IC$ for $a = 0.1$.
- Compute the percentage of times that the exponential was selected and the log normal was selected for several values of the parameter $\omega = (\mu_\omega, \sigma_\omega^2)$ of the truncated normal kernel. In particular, we choose $\mu_\omega = 0.6, 1, 1.5, 2, 3$ and use a constant standard deviation $\sigma = 0.1$.

The behavior of the BHHJ measures is as follows (see Basu et al., 1998): parameter a controls the trade-off between robustness and asymptotic efficiency of the parameter estimates. Choices of a near 0 tend to yield robust estimators while retaining efficiency close to that of maximum likelihood. The latter is the case for $a = 0$, i.e., when the BHHJ reduces to the Kullback-Leibler divergence. Therefore, we choose $a = 0.1$ in our simulations and applications, so that our local estimators are close to MLE efficiency and somewhat robust, and therefore, the resulting $LDiv.IC$ criterion is robust to the parameter estimators. However, any of the standard methods can be used in order to select a appropriately. For example, the minimum contrast method (see, Diggle, 2013, p. 132) suggests minimization of a discrepancy measure, e.g., select a such that the quantity $Q(a) = \int\limits_{\Omega}(D_a^\omega(g, f_\theta) - \hat{D}^\omega(g, f_\theta))^2 d\omega$, is minimized with respect to $a > 0$, with ω the parameters of the kernel density, $\omega \in \Omega$, and $\hat{D}^\omega(g, f_\theta)$ denotes a local divergence between g and f_θ, that could

Table 1: Displaying the percentages of selecting an Exponential or LogNormal candidate models for several sample sizes and kernels using BHHJ estimators for λ and μ

n	100		250		500		1000	
Candidate models	Exp	LogN	Exp	LogN	Exp	LogN	Exp	LogN
Driving kernel: truncated normal ($\mu, \sigma = 0.1$)								
(a)	True model: $0.5Exp(1) + 0.5LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	53.1	46.9	48.5	51.5	50.2	49.8	50.3	49.7
1	49.5	50.5	48.3	51.7	39	61	40	60
1.5	58.8	41.2	58.8	41.2	64.5	35.5	67.2	32.8
2	53.4	46.6	56.1	43.9	58.7	41.3	66.4	33.6
3	48	52	51.5	48.5	54.7	45.3	60.7	39.3
(b)	True model: *Exp*(*1*)							
$\mu = 0.6$	84.2	15.8	95.9	4.1	99.1	0.9	99.9	0.1
1	57	43	66.4	33.6	72.4	27.6	80	20
1.5	64.6	35.4	72.7	27.3	78.	21.1	87.2	12.8
2	67.3	32.7	78.3	21.7	85.1	14.9	94.3	5.7
3	61.4	38.6	66.8	33.2	79.9	20.1	89.5	10.5
(c)	True model: $LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	16.4	83.6	6	94	1	99	0	100
1	36.6	63.4	28.3	71.7	19.9	80.1	7.7	92.3
1.5	55	45	49.5	50.5	44	56	42.2	57.8
2	42.6	57.4	34.8	65.2	28.3	71.7	21.2	78.8
3	28.4	71.6	27.7	72.3	26.6	73.4	22.4	77.6
(d)	True model: $0.75Exp(1) + 0.25LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	68.1	31.9	80.6	19.4	90.6	9.4	95.5	4.5
1	54.3	45.7	53.5	46.5	60.6	39.4	59.9	40.1
1.5	60.8	39.2	66.7	33.3	73.6	26.4	80.8	19.2
2	58.2	41.8	67.4	32.6	74.7	25.3	86.5	13.5
3	52.2	47.8	60.7	39.3	69.2	30.8	79.3	20.7
(e)	True model: $0.25Exp(1) + 0.75LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	28.7	71.3	21.8	78.2	13.8	86.2	4.5	95.5
1	44.3	55.7	36.4	63.6	30.1	69.9	22.5	77.5
1.5	56.1	43.9	54.2	45.8	54.1	45.9	51.5	48.5
2	43.1	56.9	47	53	46.1	53.9	40.3	59.7
3	37.1	62.9	41.8	58.2	41	59	41.3	58.7

The parameter η is considered known and equal to 0.833. The driving kernel is a truncated normal with varying means μ and fixed $\sigma = 0.1$. The true models that generate the data are displayed for each subsection of the table (i.e., cases a–e). See the text for the interpretation of the results

be a parametricor non-parametric estimate, e.g., the local Kulback-Leibler distance of Avlogiaris et al. (2016a). Clearly, the minimization is performed via numerical methods.

The results of the simulations conducted are displayed in Tables 1 (BHHJ estimators) and 2 (MLE), whereas Fig. 2 illustrates the true models used to

Table 2: Displaying the percentages of selecting an Exponential or LogNormal candidate models for several sample sizes and kernels using MLE for all parameters of the candidate models

n	100		250		500		1000	
Candidate models	Exp	LogN	Exp	LogN	Exp	LogN	Exp	LogN
Driving kernel: truncated normal $(\mu, \sigma = 0.1)$								
(a)	True model: $0.5Exp(1) + 0.5LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	39.7	60.3	24.4	75.6	12.1	87.9	4.2	95.8
1	67.1	32.9	74	26	77.8	22.2	87	13
1.5	64.2	35.8	65.4	34.6	70.6	29.4	75.5	24.5
2	52.1	47.9	54.1	45.9	55.1	44.9	57.4	42.6
3	47.1	52.9	43.9	56.1	43.8	56.2	37.9	62.1
(b)	True model: $Exp(1)$							
$\mu = 0.6$	73.4	26.6	79.6	20.4	79.6	20.4	87.6	12.4
1	65.7	34.3	78.7	21.3	78.7	21.3	85.1	14.9
1.5	68.3	31.7	83.6	16.4	83.6	16.4	90.9	9.1
2	65.2	34.8	82.1	17.9	82.1	17.9	89.5	10.5
3	56.1	43.9	61.5	38.5	61.5	38.5	69	31
(c)	True model: $LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	11.9	88.1	2.4	97.6	0.4	99.6	0	100
1	37.1	62.9	22	78	11.5	88.5	3.6	96.4
1.5	61.8	38.2	57	43	48.9	51.1	44.2	55.8
2	45	55	33.6	66.4	32.4	67.6	19.3	80.7
3	39.9	60.1	32.4	67.6	23.5	76.5	14.2	85.8
(d)	True model: $0.75Exp(1) + 0.25LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	44.5	55.5	50.1	49.9	61.1	38.9	41.1	58.9
1	83.2	16.8	74.1	25.9	72.2	27.8	89.5	10.5
1.5	78.6	21.4	71	29	66.5	33.5	85	15
2	70.1	29.9	59.9	40.1	58.5	41.5	78.2	21.8
3	50.9	49.1	51.7	48.3	49.2	50.8	55.6	44.4
(e)	True model: $0.25Exp(1) + 0.75LogNormal(-\ 0.347, 0.833)$							
$\mu = 0.6$	25.4	74.6	8.3	91.7	2.2	97.8	0.1	99.9
1	56.5	43.5	46	54	35.6	64.4	26.3	73.7
1.5	57.7	42.3	58.7	41.3	63.3	36.7	67.5	32.5
2	46.9	53.1	43.1	56.9	42.7	57.3	37.9	62.1
3	44.6	55.4	37.8	62.2	36.2	63.8	24.4	75.6

The driving kernel is a truncated normal with varying means μ and fixed $\sigma = 0.1$. The true models that generate the data are displayed for each subsection of the table (i.e., cases a–e). See the text for the interpretation of the results

generate the samples. First we note that as the sample size increases, the *LDiv.IC* is able to efficiently select the true model locally (see Tables 1 and 2, cases b and c). This is easily accomplished in areas of the domain (driven by the kernel) where the candidate models are clearly different (e.g., cases b and c of Tables 1 and 2, for $\mu = 0.6, 1, 1.5, 2$, and 3). For example,

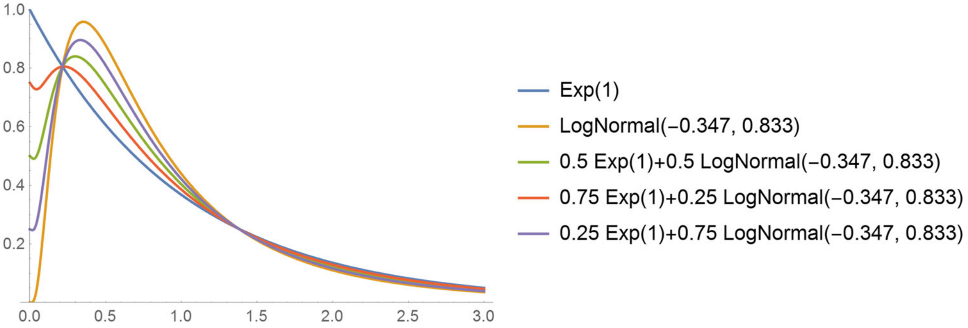

Figure 2: Plots of several densities used in the simulation study (see Tables 1 and 2)

for a kernel that focuses attention about $\mu = 0.6$, the true model is selected 99.9 or 100% of the time ($n = 1000$).

In the case where the true model is a mixture of the entertained $Exp(1)$ and $LogNormal(-0.347, 0.833)$ models (see Tables 1 and 2, cases a, d, and e), global selection criteria reject the usefulness of both models. However, focusing on specific areas of the domain, the $LDiv.IC$ is able to identify the true model even if the data arises from the mixture model (e.g., cases d and e in Table 1, for $\mu = 0.6$ and $n = 1000$, or case e in Table 2, for $\mu = 0.6$ and $n = 1000$). In addition, there are intervals where a certain model is fit best, which is different compared with the model that is selected in the whole domain according to the global selection criteria.

4.2. Multivariate Case: a Point Process Example Consider a region $\mathcal{W} \subset R^2$ and suppose that we observe n points $\{x_i\}_{i=1}^{n}$. To model this collection of points as a point process, we consider the observed points as arising from a model with two sources of randomness; the number of points n, and conditional on knowing n, the points are randomly generated over the region $\mathcal{W}$. The observed points (events) are then called a point pattern and are treated as a realization from a point process N over the window $\mathcal{W}$.

Similarly to the first moment for random variables, point processes can be characterized by their corresponding intensity function $\lambda(x)$, $x \in \mathcal{W}$, where $\lambda(x)dx$ assumes the interpretation of the probability of observing a point in an infinitesimal disc (sphere in R^d) centered at x, of volume dx. Of course the first moment does not uniquely determine the point process in general, unless certain conditions are met (e.g., Cressie, 1993).

More precisely, letting $N(\mathcal{W})$ denote the number of points over $\mathcal{W}$, we assume that $N(\mathcal{W}) \backsim$ Poisson $(\Lambda(\mathcal{W}))$, where $\Lambda(\mathcal{W})$ denotes the first-order measure of the point process, i.e., $E(N(\mathcal{W})) = \Lambda(\mathcal{W})$. In addition, we assume that counts over non-overlapping areas are independent random variables. These two assumptions lead to the most important point process model, known as the inhomogenous Poisson point process ($IPPP$) N with intensity measure Λ, and if Λ is dominated by Lebesgue measure, there exists a measurable function $\lambda(x|\theta)$ such that

$$\Lambda_\theta(\mathcal{W}) = \int_{\mathcal{W}} \lambda(x|\theta)dx, \tag{4.1}$$

for some parameter vector θ. Then the joint distribution of the events observed over some window $\mathcal{W}$, given that $N(\mathcal{W}) = n$, is given by

$$f(x_1, \ldots, x_n|\theta, N(\mathcal{W}) = n) = \prod_{i=1}^{n} \frac{\lambda(x_i|\theta)}{\Lambda_\theta(\mathcal{W})}, \tag{4.2}$$

for $x_i \in \mathcal{W}$, $i = 1, \ldots, n$. The function $\lambda(x|\theta)$ is called the intensity function and it uniquely identifies the point process distribution.

Our goal is to estimate θ in a robust way while we vary our focus over the window of observation $\mathcal{W}$. Following Micheas (2014), we choose to model the intensity function using a multiple of a proper density $\lambda_\theta(x)$, i.e.,

$$\lambda(x|\theta, \xi) = \xi\lambda_\theta(x), \tag{4.3}$$

which is not a density in general, with $\xi > 0$.

In order to evaluate the performance of the proposed local divergence information criterion, we performed a Monte Carlo simulation study using the $LDiv.IC$. Figure 3 illustrates $n = 1000$ events observed in the window $\mathcal{W} = (0, 1) \times (0, 1)$, which arise as a realization of a nonhomogeneous Poisson point process with $\xi = 1025$ and $\lambda_\theta(x)$ a mixture model of three normal components. More precisely,

$$\lambda_\theta(x) = 0.4N((0.25, 0.25), \Sigma) + 0.3N((0.75, 0.25), \Sigma) + 0.3N((0.5, 0.75), \Sigma),$$

with $\Sigma = \begin{pmatrix} 0.005 & 0 \\ 0 & 0.005 \end{pmatrix}$. We consider the problem of choosing between three candidate models, locally. As candidate models, we consider the bivariate normal components of $\lambda_\theta(x)$, i.e.,

$$f(x; \mu_i, \Sigma_i) = (2\pi)^{-1}|\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_i)^t\Sigma_i^{-1}(x - \mu_i)\right),$$

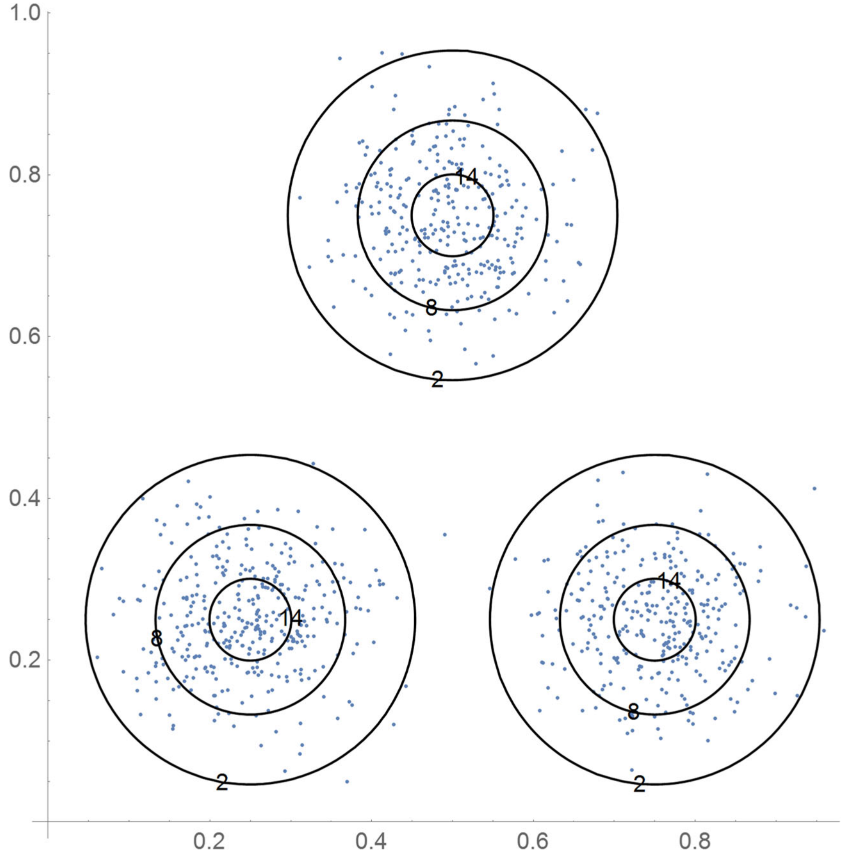

Figure 3: A realization of a nonhomogeneous Poisson point process with $\xi = 1025$ and contours of the three different bivariate normal kernels. There are $n = 1000$ events observed in the window $(0,1) \times (0,1)$

with mean vectors μ_i, $i = 1,2,3$, covariance matrices Σ_i, $i = 1,2,3$. Using the classic method of simulating from an IPPP by Lewis and Shedler (1979), the simulation study based on the $LDiv.IC$ has the following characteristics:

- Generate the number of points $N(\mathcal{W}) \backsim$ Poisson $(\Lambda(\mathcal{W}))$, where $\Lambda(\mathcal{W}) = \int_{\mathcal{W}} \xi \lambda_\theta(x)dx$. Say $N(\mathcal{W}) = n$.
- Find $\lambda^* = \max\limits_{x \in \mathcal{W}} \lambda_\theta(x)$.
- Generate a point $x_0 \in \mathcal{W}$ on $\mathcal{W}$ uniformly and $U \backsim \mathcal{U}(0,1)$. Check if $U < \frac{\lambda_\theta(x_0)}{\lambda^*}$ and keep the point. Repeat until we retain n-points.

These points form a point pattern from the entertained Poisson point process.

- Calculate the value of the AIC and DIC for $a = 0.1$.
- Compute the $LDiv.IC$ for several kernels of the bivariate normal. More precisely, we choose μ_0 to be one of the vectors $(0.25, 0.25)$, $(0.75, 0.75)$ and $(0.5, 0.75)$, and for the variance-covariance matrix Σ_0 we choose the matrix $\Sigma_0 = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$.

We present the results in Table 3, where the values of the local criterion are calculated for each kernel in the window $\mathcal{W} = (0, 1) \times (0, 1)$. We notice that in the area with center $(0.25, 0.25)$, the first normal component is selected, in the area with center $(0.75, 0.25)$, the second normal component is selected, whereas the third normal component is selected in the area with center $(0.5, 0.75)$. As a result, the true model is selected in each of the focus areas.

In addition, we present the values of the AIC and DIC. It must be noted that both global criteria (AIC and DIC) select the component with the highest weight in the true model. Therefore, global selection criteria provide erroneous results in this case since they lose track of the local characteristics of a model, and such criteria do not involve a mechanism that can capture local behavior. In contrast, the $LDiv.IC$ is able by construction to identify and include local characteristics in making an informed decision about competing models. Finally, note that the selection of the "best" model in each focus area enables us to estimate the parameters of the model in a

Table 3: Displaying the values of the $LDiv.IC$ for three candidate models and several kernel choices

	Model 1	Model 2	Model 3	Selection
Driving kernel: bivariate normal $(\mu_0, \Sigma_0 = [(0.01, 0), (0, 0.01)])$				
	$LDiv.IC$	$LDiv.IC$	$LDiv.IC$	Model selected
$(0.25, 0.25)$	-42416.8	-5340.27	-2933.20	1
$(0.75, 0.25)$	-4832.02	-36591.1	-2632.57	2
$(0.5, 0.75)$	-2805.56	-2749.72	-34964.1	3
AIC	30594.2	33136.6	37800.5	1
DIC	-7508.30	-6865.79	-6500.77	1

Model 1 refers to a bivariate normal that corresponds to the first normal component of the mixture model, model 2 refers to the second component, and model 3 refers to the third component. The driving kernel is a bivariate normal with varying means μ_0 and fixed Σ_0. In the last column, we display the model selected. See the text for the interpretation of the results "robust" way as the focus area changes, since we are able to recover the parameters of the true component that realizes the data every time. Although this example is somewhat artificial, it serves the purpose in exemplifying the aforementioned issues when performing model selection locally or globally.

5 Application

We discuss applications of the methodology developed to real life data. Before we present the data analysis, we briefly discuss the use and interpretation of simple differences of the local information criterion LDiv.IC, following Burnham and Anderson (2002, p. 70). In particular, define $\Delta LDiv.IC_i = LDiv.IC_i - LDiv.IC_{min}$, where $LDiv.IC_i$ the value of the local information criterion for model i, and $LDiv.IC_{min} = \underset{j}{min} LDiv.IC_j$. Then, model i is estimated to be the best if it has $\Delta LDiv.IC_i = 0$. The larger $\Delta LDiv.IC_i$ is, the less likely it is that the i^{th} model is the best locally.

Rough rules of thumb can be built in a similar fashion as in the (global) AIC case. In particular, we conducted a simulation study based on the examples of the simulation section (results omitted), in order to help us identify meaningful ranges of values for the $\Delta LDiv.IC_i$, since in these examples, we know the true models, and therefore, we can validate the model selection criteria. More precisely, our simulations showed that values of the $\Delta LDiv.IC_i$ in the $[0, 1)$ range show strong support for model i, values in the interval $[1, 5)$ show marginal support for model i, whereas, values above 5 indicate no support for model i. We compute and present these values in what follows and use them as evidence to suggest the best model in each case. It should be noted that the recommended intervals are a first attempt at identifying these important ranges for the value of the $\Delta LDiv.IC_i$, and they are by no means a panacea. Further study is required in order to fully appreciate the behavior of the $\Delta LDiv.IC_i$ and its range of values in different settings.

5.1. Univariate Case: Galaxy Data The galaxy data was first studied by Postman et al. (1986). The data describe the velocity in km^3/sec of 82 galaxies from the six conical areas of the corona Borealist constellation. In 1992, Roeder (cf. Roeder, 1992) was the first to apply a mixture model to this data and since then, this has constituted a benchmark example for papers working with mixture distributions. For a comparative presentation on the different approaches using the Galaxy data, we refer to Aitkin (2001).

Most researchers believe that these observations arise from a mixture model with normal components. Figure 4 presents a histogram of the data. In this histogram, we can see several gaps within the main bulk of the observations (from 15 to about 26), as well as some outliers which are concentrated near velocities 10 and 32. Therefore, it is reasonable to assume that this is a mixture of at least three components, and this observation is the only one that several researchers seem to agree on. In particular, most papers in the literature consider the estimate for the number of the components to be between 3 and 8.

We consider five candidate models for this data: normal distribution, mixture of two normal components, mixture of three normal components, mixture of four normal components, and mixture of five normal components. The data histogram as well as the densities of the five candidate models is illustrated in Fig. 4, and in Table 4, we present the parameters of the respective models. Note that the AIC suggests the mixture of four normal components, with value $AIC = 418.916$. The results are similar in Table 5, where we display the $\Delta LDiv.IC$, with the values allowing us to easily and quickly identify which is the best model proposed in each case.

Furthermore, we apply the local selection criterion ($LDiv.IC$) for several normal kernels (see Table 4). Initially, we choose to divide the window of observation based on three normal kernels with means 10, 20, and 30 and constant standard deviation $\sigma = 1.67$, respectively. In this case and using the local criterion, a different mixture of normal components is selected for each interval as the best model, where the intervals are determined by the driving normal kernel (see part a of Table 4).

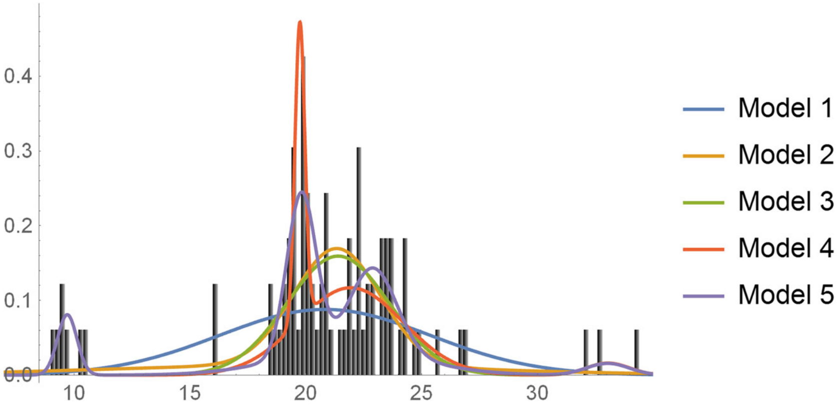

Figure 4: Histogram of the galaxy data and the densities of the five candidate models, where model i consists of a mixture with i univariate normal components, $i = 1, \ldots, 5$

Table 4: Displaying the values of the *LDiv.IC* for the galaxy data for selecting five candidate models for several kernels using MLE

Galaxy data	Model 1	Model 2	Model 3	Model 4	Model 5	Model selected
(a) Kernel: normal ($\mu_0 = 10, 20, 30,\ \sigma_0^2 = 1.67^2$)						
$(10, 1.67^2)$	− 9.8125	− 9.9070	− 11.3280	− 11.3286	− 11.3127	4
$(20, 1.67^2)$	− 78.9007	− 79.3333	− 79.7111	− 80.2637	− 80.3412	5
$(30, 1.67^2)$	− 1.0762	− 1.5513	− 1.6083	− 1.5919	− 1.5102	3
(b) Kernel: normal ($\mu_0 = 10, 16, 22, 28,\ \sigma_0^2 = 1^2$)						
$(10, 1^2)$	− 15.4213	− 15.3812	− 17.4966	− 17.4969	− 17.4888	4
$(16, 1^2)$	− 4.0021	− 5.3062	− 5.2868	− 5.2169	− 5.2779	2
$(22, 1^2)$	− 77.116	− 76.910	− 77.049	− 77.3723	− 77.2399	4
$(28, 1^2)$	− 1.9359	− 2.6320	− 2.7668	− 2.7533	− 2.6161	3
(c) Kernel: normal ($\mu_0 = 10, 12.5, 15, 17.5, 20, 22.5, 25, 27.5, 30,\ \sigma_0^2 = 0.4175^2$)						
$(10, 0.4175^2)$	− 23.2203	− 22.9663	− 25.6755	− 25.6759	− 25.6697	4
$(12.5, 0.4175^2)$	0.9708	0.2343	0.0054	0.0035	0.0755	4
$(15, 0.4175^2)$	1.9243	− 0.0748	− 0.1721	− 0.2238	− 0.0871	4
$(17.5, 0.4175^2)$	2.8029	0.1507	0.8093	− 0.0772	− 0.2833	5
$(20, 0.4175^2)$	− 142.619	− 144.392	− 145.232	− 146.537	− 146.858	5
$(22.5, 0.4175^2)$	− 85.8838	− 85.9618	− 86.5534	− 86.402	− 85.971	3
$(25, 0.4175^2)$	− 20.4891	− 20.9452	− 20.8027	− 20.701	− 20.657	2
$(27.5, 0.4175^2)$	− 2.9717	− 3.7274	− 3.7838	− 3.7863	− 3.6396	4
$(30, 0.4175^2)$	0.6793	0.2212	0.0676	0.0718	0.1642	3
AIC	484.676	472.553	424.360	418.916	422.132	4

The driving kernel is a normal with varying means μ_0 and fixed σ_0^2. The models considered are mixtures of univariate normals with 1–5 components. In the last column, we have the model that is selected from the following candidate models: model 1: $\hat{\theta}_1 = (\hat{\mu_1}, \hat{\sigma_1}^2) = (20.83,\ 4.54^2)$, model 2: $\hat{\theta}_2 = (\hat{\mu_2},\ \hat{\mu_3},\ \hat{\sigma_2}^2,\ \hat{\sigma_3}^2,\ \hat{w_2},\ \hat{w_3}) = (21.35,\ 19.36,\ 1.88^2,\ 8.15^2,\ 0.74,\ 0.26)$, model 3: $\hat{\theta}_3 = (\hat{\mu_4},\ \hat{\mu_5},\ \hat{\mu_6},\ \hat{\sigma_4}^2,\ \hat{\sigma_5}^2,\ \hat{\sigma_6}^2, \hat{w_4},\ \hat{w_5},\ \hat{w_6}) = (33.04,\ 21.40,\ 9.71,\ 0.92^2,\ 2.20^2,\ 0.42^2,\ 0.037,\ 0.878,\ 0.085)$, model 4: $\hat{\theta}_4 = (\hat{\mu_7}, \hat{\mu_8},\ \hat{\mu_9},\ \widehat{\mu_{10}},\ \hat{\sigma_7}^2,\ \hat{\sigma_8}^2,\ \hat{\sigma_9}^2,\ \widehat{\sigma_{10}}^2, \hat{w_7},\ \hat{w_8},\ \hat{w_9},\ \widehat{w_{10}}) = (33.05,\ 21.94,\ 19.75,\ 9.71,\ 0.92^2,\ 2.27^2,\ 0.45^2,\ 0.42^2,\ 0.037,\ 0.665,\ 0.213,\ 0.085)$, and model 5: $\hat{\theta}_5 = (\widehat{\mu_{11}},\ \widehat{\mu_{12}},\ \widehat{\mu_{13}},\ \widehat{\mu_{14}},\ \widehat{\mu_{15}},\ \widehat{\sigma_{11}}^2,\ \widehat{\sigma_{12}}^2,\ \widehat{\sigma_{13}}^2,\ \widehat{\sigma_{14}}^2,\ \widehat{\sigma_{15}}^2,\ \widehat{w_{11}},\ \widehat{w_{12}},\ \widehat{w_{13}},\ \widehat{w_{14}},\ \widehat{w_{15}}) = (33.04,\ 22.92,\ 21.85,\ 19.82,\ 9.71,\ 0.92^2,\ 1.02^2,\ 3.05^2,\ 0.63^2,\ 0.42^2,\ 0.036,\ 0.289,\ 0.245,\ 0.344,\ 0.085)$. See the text for the interpretation of the results

Next, we concentrate on finer areas of the domain by selecting a smaller standard deviation in the normal driving kernel ($\sigma = 1$ and $\sigma = 0.4175$), and consider similar means. In parts b and c of Table 4, we obtain the results where different models are selected depending on the focus area (see column 5 parts b and c of Table 4). In particular, for all parts a, b, and c of the table, the mixture with 4 normal components emerges as the model mostly selected (7/16 times) regardless of the area we focus to compute the local selection criterion. This conclusion is also supported by the AIC.

Table 5: Displaying the values of the $\Delta LDiv.IC_{ij}$ for five candidate models $j = 1, 2, \dots, 5$, and sixteen kernels $i = 1, 2, \dots, 16$

Galaxy data	Model 1	Model 2	Model 3	Model 4	Model 5	Model selected
(a)						
$\Delta LDiv.IC_{1j}$	1.5161	1.4216	0.0006	0	0.0159	4
$\Delta LDiv.IC_{2j}$	1.4405	1.0079	0.6301	0.0775	0	5
$\Delta LDiv.IC_{3j}$	0.5321	0.0570	0	0.0164	0.0981	3
(b)						
$\Delta LDiv.IC_{4j}$	2.0756	2.1157	0.0003	0	0.0081	4
$\Delta LDiv.IC_{5j}$	1.3041	0	0.0194	0.0893	0.0283	2
$\Delta LDiv.IC_{6j}$	0.2563	0.4623	0.3233	0	0.1324	4
$\Delta LDiv.IC_{7j}$	0.8309	0.1348	0	0.0135	0.1507	3
(c)						
$\Delta LDiv.IC_{8j}$	2.4556	2.7096	0.0004	0	0.0062	4
$\Delta LDiv.IC_{9j}$	0.9673	0.2308	0.0019	0	0.0720	4
$\Delta LDiv.IC_{10j}$	2.1481	0.1490	0.0517	0	0.1367	4
$\Delta LDiv.IC_{11j}$	3.0862	0.4340	1.0926	0.2061	0	5
$\Delta LDiv.IC_{12j}$	4.2390	2.4660	1.6260	0.3210	0	5
$\Delta LDiv.IC_{13j}$	0.6696	0.5916	0	0.1514	0.5824	3
$\Delta LDiv.IC_{14j}$	0.4561	0	0.1425	0.2442	0.2882	2
$\Delta LDiv.IC_{15j}$	0.8146	0.0589	0.0025	0	0.1467	4
$\Delta LDiv.IC_{16j}$	0.6117	0.1536	0	0.0042	0.0966	3
ΔAIC_j	65.760	53.637	5.4440	0	3.2160	4

See the text for the interpretation of the results

5.2. Multivariate Case: Iris Data The iris data (Fisher, 1936) is perhaps the best known data set to be found in the multivariate analysis literature. The iris data set was introduced by R. A. Fisher as an example for discriminant analysis. The data set includes 3 categories of 50 cases each, where each class refers to a type of iris plant. Measurements in centimeters were taken from each sample on the length and width of the sepal and petals. In particular, the data contains the following variables: (1) sepal length, (2) sepal width, (3) petal length, (4) petal width, and (5) class: iris setosa, iris versicolor, and iris virginica. In our example we will only focus on the variables petal width and petal length.

In order to evaluate the performance of the proposed local divergence information criterion, we performed a study using the $LDiv.IC$ in the iris data. Figure 5 illustrates $n = 150$ observations for variables petal width (x-axis) and petal length (y-axis). We assume that the observations arise from a mixture model of three normal components given by

$$f_\theta(x) = \frac{1}{3}N((0.246, 1.462), \Sigma_1) + \frac{1}{3}N((1.326, 4.260), \Sigma_2) + \frac{1}{3}N((2.026, 5.552), \Sigma_3),$$

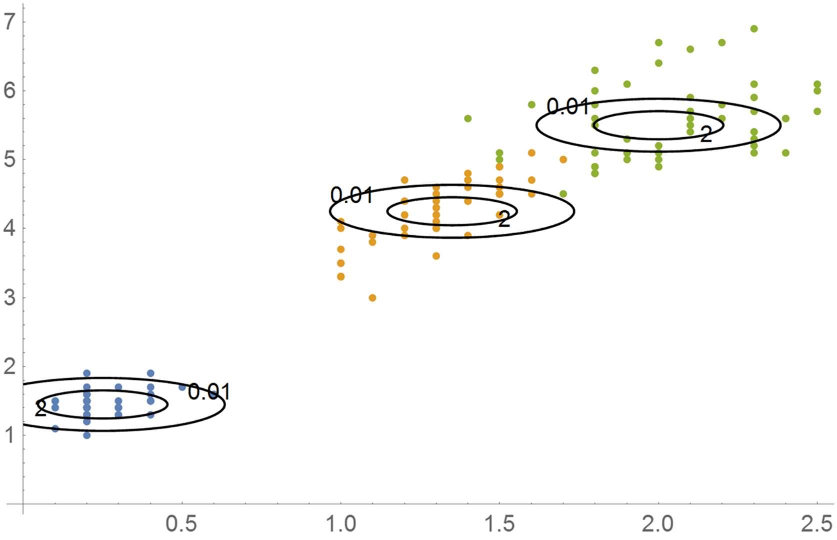

Figure 5: Plot of the variables petal width (x-axis) vs petal length (y-axis) and contours of the three different bivariate normal kernels. The data set contains 50 observations of each of the three species of iris, setosa (blue), virginica (yellow), and versicolor (green)

with $\Sigma_1 = \begin{pmatrix} 0.0109 & 0.0059 \\ 0.0059 & 0.0296 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.0383 & 0.0716 \\ 0.0716 & 0.2164 \end{pmatrix}$ and $\Sigma_3 = \begin{pmatrix} 0.0739 & 0.0478 \\ 0.0478 & 0.2985 \end{pmatrix}$. The estimates of the mixture model parameter means and covariance matrices were obtained using the data augmentation approach (Dempster et al., 1977), whereas each component is given the same probability (since each group has 50 cases).

We consider the problem of choosing between three candidate models, locally. As candidate models, we consider the bivariate normal components of $f_\theta(x)$, i.e.,

$$f(x; \mu_i, \Sigma_i) = (2\pi)^{-1} |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i)\right),$$

with mean vectors μ_i and variance-covariance matrices Σ_i, $i = 1, 2, 3$. The study based on the $LDiv.IC$ has the following characteristics:

- Calculate the value of the AIC and DIC for $a = 0.1$.

- Compute the $LDiv.IC$ for several kernels of the bivariate normal. More precisely, we choose μ_0 to be one of the vectors $(0.25, 1.45)$, $(1.35, 4.25)$,

Table 6: Displaying the values of the *LDiv.IC* for selecting three candidate models for several kernels

Iris data	Model 1	Model 2	Model 3	Selection
Driving kernel: bivariate normal $(\mu_0, \Sigma_0 = [(0.05, 0), (0, 0.05)])$				
μ_0	*LDiv.IC*	*LDiv.IC*	*LDiv.IC*	Model selected
$(0.25, 1.45)$	-1114.76	0.00001	0	1
$(1.35, 4.25)$	-0.00152	-805.079	-6.74515	2
$(2, 5.5)$	0	-3.80021	-272.447	3
AIC	48426.18	2692.16	4358.32	2
DIC	-709.715	-917.613	-792.693	2

Model 1 refers to a bivariate normal that corresponds to the first normal component of the mixture model, model 2 refers to the second component, and model 3 refers to the third component. The driving kernel is a bivariate normal with varying means μ_0 and fixed Σ_0. In the last column, we note the model selected by the *LDiv.IC*. See the text for the interpretation of the results

and $(2, 5.5)$, and for the variance-covariance matrix Σ_0, we choose the matrix $\Sigma_0 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$ (see Table 6).

We present the results in Table 6, where the values of the local criterion are calculated for each kernel. We notice that in the area with center $(0.25, 1.45)$, the first normal component is selected, in the area with center $(1.35, 4.25)$, the second normal component is selected, whereas the third normal component is selected in the area with center $(2, 5.5)$. In addition, we present the values of the AIC and DIC and notice that both global criteria select the second component. Therefore, the global selection criteria tend to select the model corresponding to the center of the data, since these procedures do not take into account local characteristics of the data. Similar results are obtained using Table 7 via the $\Delta LDiv.IC$ values.

5.3. Local Model Selection for a Point Process: Redwoodfull Data Next we evaluate the performance of the proposed local divergence information criterion using an example from the point process literature, namely the redwoodfull data. This data represents the locations of 195 seedlings and saplings of California redwood trees in a square sampling region, and it was

Table 7: Displaying the values of the $\Delta LDiv.IC_{ij}$ for three candidate models $j = 1, 2, 3$, and three kernels $i = 1, 2, 3$

Iris data	Model 1	Model 2	Model 3	Selection
$\Delta LDiv.IC_{1j}$	0	1114.76001	1114.76	1
$\Delta LDiv.IC_{2j}$	805.077	0	798.334	2
$\Delta LDiv.IC_{3j}$	272.447	268.647	0	3
ΔAIC_j	45734.02	0	1666.16	2

See the text for the interpretation of the results

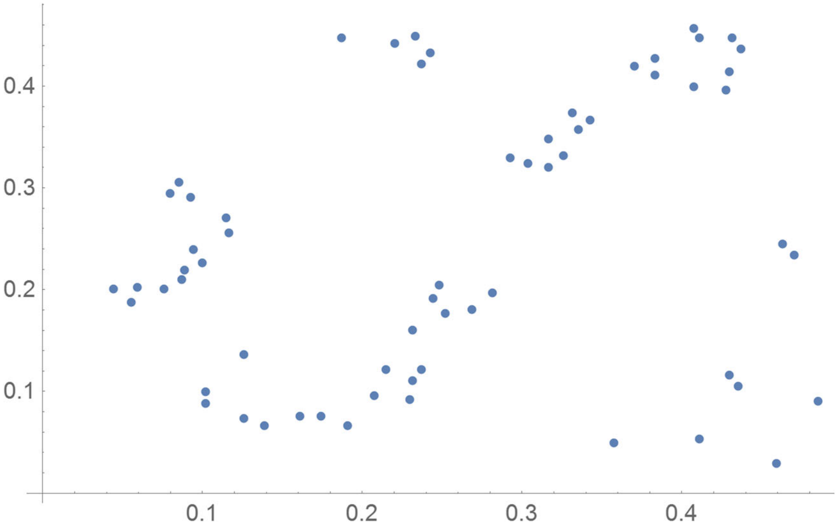

Figure 6: Displaying locations of redwood trees in California. There are $n = 63$ events observed

originally described and analyzed by Strauss (1975). In our study, we use 63 observations in the reduced window $\mathcal{W} = (0, 0.5) \times (0, 0.5)$, presented in Fig. 6.

The study based on the *LDiv.IC* has the following characteristics:

- As candidate models for the intensity of the Poisson process, we consider six models of mixtures of bivariate normal models, where model i consists of a mixture with i bivariate normal components $i = 1, \ldots, 6$.

Table 8: Displaying the values of the *LDiv.IC* for six candidate models and several kernels

Redwood	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Selection
Driving kernel: bivariate normal $(\mu_0, \Sigma_0 = [(0.01, 0), (0, 0.01)])$							
μ_0	*LDiv.IC*	*LDiv.IC*	*LDiv.IC*	*LDiv.IC*	*LDiv.IC*	*LDiv.IC*	Model selected
$(0.08, 0.25)$	– 2941.13	– 2958.20	– 3008.57	– 3050.93	– 3039.29	– 3010.76	4
$(0.2, 0.45)$	– 1645.45	– 1647.45	– 1612.76	– 1659.03	– 1640.06	– 1612.99	4
$(0.2, 0.15)$	– 3689.42	– 3690.03	– 3686.59	– 3738.56	– 3703.70	– 3670.25	4
$(0.35, 0.35)$	– 3170.21	– 3200.66	– 3111.32	– 3260.05	– 3237.78	– 3209.04	4
$(0.40, 0.40)$	– 2958.16	– 2994.83	– 2915.98	– 3078.63	– 3061.47	– 3032.35	4
$(0.45, 0.10)$	– 1177.27	– 1197.46	– 1142.21	– 1183.27	– 1188.05	– 1122.75	2
AIC	– 147.732	– 158.662	– 137.785	– 173.835	– 162.56	– 149.903	4

The driving kernel is a bivariate normal with varying means μ_0 and fixed Σ_0. In the last column, we note the model selected by the *LDiv.IC*. See the text for the interpretation of the results

Table 9: Displaying the values of the $\Delta LDiv.IC_{ij}$ for six candidate models, $j = 1, 2, \ldots, 6$, and six kernels $i = 1, 2, \ldots, 6$

Redwood	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Selection
$\Delta LDiv.IC_{1j}$	109.80	92.73	42.36	0	11.64	40.17	4
$\Delta LDiv.IC_{2j}$	13.58	11.58	46.27	0	18.97	46.04	4
$\Delta LDiv.IC_{3j}$	49.14	48.53	51.97	0	34.86	68.31	4
$\Delta LDiv.IC_{4j}$	89.84	59.39	148.73	0	22.27	51.01	4
$\Delta LDiv.IC_{5j}$	120.47	83.80	162.65	0	17.16	46.28	4
$\Delta LDiv.IC_{6j}$	20.19	0	55.25	14.19	9.41	74.71	2
ΔAIC_j	26.10	15.17	36.05	0	11.28	23.93	4

See the text for the interpretation of the results

The estimated parameters of the six candidate models are given in the Appendix (Section A.1).

- Calculate the value of the AIC for $a = 0.1$.

- Compute the $LDiv.IC$ for several bivariate normal kernels. More precisely, we choose $\mu_0 = (0.08, 0.25)$, $(0.20, 0.45)$, $(0.20, 0.15)$, $(0.35, 0.35)$, $(0.4, 0.4)$, or $(0.45, 0.10)$ and fixed variance-covariance matrix $\Sigma_0 = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$ (see Table 8).

We present the results in Table 8, where the values of the local criterion are calculated for each kernel. We notice that in the focus areas based on the kernels with means $\mu_0 = (0.08, 0.25)$, $(0.20, 0.45)$, $(0.20, 0.15)$, $(0.35, 0.35)$, $(0.4, 0.4)$), and fixed Σ_0, model 4 is selected, whereas in the focus area with mean $\mu_0 = (0.45, 0.10)$, model 2 is selected. In addition, the AIC suggests model 4 with value $AIC = -173.835$. As a result, the model with 4 mixtures of normal components emerges as the most appropriate both globally and locally. We reach same conclusion by looking $\Delta LDiv.IC$ values presented in Table 9.

6 Conclusions

In this paper, we introduced the concept of local model selection. In order to create the local criterion, we introduced a broad class of local divergence measures between two probability measures, based on the BHHJ power divergence. Explicit expressions of the proposed local divergences were derived when the underline distributions are members of the exponential family of distributions or they are described by multivariate normal models.

The local model selection criterion ($LDiv.IC$) we developed was exemplified via simulations and three classic examples from the literature.

In particular, the *LDiv.IC* can be used to propose the best number of components of a mixture model (locally), as well as provide a robust estimating procedure of model parameters, e.g., for a point process model that is affected by the choice of window over which we observe the point pattern, we are able to perform model selection regardless of the focus area; we simply select an appropriate driving kernel and then assess the performance of the model in a particular area of the observation window. This criterion enables us to select the "best" model among several candidate models, on a specific area of their common domain. In our simulations and applications, we have illustrated the robust behavior of the proposed local criterion, identifying the true model locally regardless of the kernel choice.

Finally, there are several theoretical and practical considerations that we have not fully explored in this paper. Firstly, the estimation methods used in the calculation of the *LDiv.IC* are based on a specific local φ-divergence measure (BHHJ), and therefore, in subsequent works, we will consider other measures of similarity. Second, we have explored only a few cases where the *LDiv.IC* was applied in practice. There are several areas where local model selection is crucial, including extreme value theory (tail behavior), robust estimation, spatial statistics, and time series analysis. Moreover, we have not explored the Bayesian paradigm in the creation of local model selection criteria. These are areas of great interest and will be explored elsewhere.

7 Proofs

This section presents the proofs of the main results of Section 3.

7.1. Proof of Lemma 4 This subsection provides a detailed proof of Lemma 4.

Proof. Based on equation (9) of Mattheou et al. (2009), a Taylor expansion of the quantity $Q^\omega_\alpha(\theta)$ defined in (3.12), about the estimator $\hat{\theta}$, yields the approximation:

$$\begin{aligned} Q^\omega_a(\theta) \quad = \quad & Q^\omega_a(\hat{\theta}) + (\theta-\hat{\theta})^t[\nabla_\theta Q^\omega_a(\theta)]_{\theta=\hat{\theta}} + \frac{1}{2}(\theta-\hat{\theta})^t[\nabla^2_\theta Q^\omega_a(\theta)]_{\theta=\hat{\theta}}(\theta-\hat{\theta}) \\ & + o\left(\|\hat{\theta}-\theta\|^2\right),\ a>0. \end{aligned} \tag{7.1}$$

Moreover, based on equation (12) of Mattheou et al. (2009), and Lemmas 1 and 3, we have

$$[\nabla_\theta Q^\omega_a(\theta)]_{\theta=\theta_0} \xrightarrow[n\to\infty]{P} [\nabla_\theta W^\omega_\alpha(\theta)]_{\theta=\theta_0} \text{ and } [\nabla^2_\theta Q^\omega_a(\theta)]_{\theta=\theta_0} \xrightarrow[n\to\infty]{P} [\nabla^2_\theta W^\omega_\alpha(\theta)]_{\theta=\theta_0}, \tag{7.2}$$

as $n \to \infty$. By the fact that $\hat{\theta} \to \theta_0$, and formulas (3.7), (3.8) and (7.2), the following holds

$$[\nabla_\theta Q_a^\omega(\theta)]_{\theta=\hat{\theta}} \xrightarrow[n\to\infty]{P} 0 \text{ and } [\nabla_\theta^2 Q_a^\omega(\theta)]_{\theta=\hat{\theta}} \xrightarrow[n\to\infty]{P} (a+1)J^\omega(\theta_0). \tag{7.3}$$

Taking into account (7.3), we reformulate (7.1) as

$$Q_a^\omega(\theta) = Q_a^\omega(\hat{\theta}) + \frac{a+1}{2}(\theta-\hat{\theta})^t J^\omega(\theta_0)(\theta-\hat{\theta}) + o\left(\|\hat{\theta}-\theta\|^2\right). \tag{7.4}$$

Letting $\theta = \theta_0$ in the above equation, we have

$$Q_a^\omega(\theta_0) = Q_a^\omega(\hat{\theta}) + \frac{a+1}{2}(\theta_0-\hat{\theta})^t J^\omega(\theta_0)(\theta_0-\hat{\theta}) + o\left(\|\hat{\theta}-\theta_0\|^2\right), \tag{7.5}$$

and consequently

$$E_g(Q_a^\omega(\theta_0)) = E_g(Q_a^\omega(\hat{\theta})) + \frac{a+1}{2}E_g((\theta_0-\hat{\theta})^t J^\omega(\theta_0)(\theta_0-\hat{\theta})) + E_g(R_n), \tag{7.6}$$

where $R_n = (\|\hat{\theta}-\theta_0\|^2)$. Moreover, using Lemma 2, we obtain

$$E_g(W_a^\omega(\hat{\theta})) = W_a^\omega(\theta_0) + \frac{a+1}{2}E_g((\hat{\theta}-\theta_0)^t J^\omega(\theta_0)(\hat{\theta}-\theta_0)) + E_g(R_n). \tag{7.7}$$

On the other hand, $E_g(Q_a^\omega(\theta_0)) = W_a^\omega(\theta_0)$. Indeed,

$$\begin{aligned}
E_g(Q_a^\omega(\theta_0)) &= E_g\left(E_{f_{\theta_0}}(h_\omega(X)f_{\theta_0}^{1+a}(X))\right) - (1+a^{-1})E_g\left(\frac{1}{n}\sum_{i=1}^n h_\omega(X_i)f_{\theta_0}^a(X_i)\right) \\
&= E_{f_{\theta_0}}(h_\omega(X)f_{\theta_0}^{1+a}(X)) - (1+a^{-1})\frac{1}{n}\sum_{i=1}^n E_g\left(h_\omega(X_i)f_{\theta_0}^a(X_i)\right) \\
&= E_{f_{\theta_0}}(h_\omega(X)f_{\theta_0}^{1+a}(X)) - (1+a^{-1})\frac{1}{n}\sum_{i=1}^n E_g\left(h_\omega(X)f_{\theta_0}^a(X)\right) \\
&= E_{f_{\theta_0}}(h_\omega(X)f_{\theta_0}^{1+a}(X)) - (1+a^{-1})E_g\left(h_\omega(X)f_{\theta_0}^a(X)\right) \\
&= W_a^\omega(\theta_0).
\end{aligned}$$

Now, combining equations (7.6) and (7.7), we have the desired result.

7.2. Proof of Proposition 5 This subsection provides a detailed proof of Proposition 5.

Proof. i) Lemma 4 and the fact that $no\left(\|\hat{\theta}-\theta_0\|^2\right) = no_p(n^{-1}) = o_p(1)$ (cf. Pardo, 2006, p. 411–412) lead to the desired result.

ii) By Theorem 4.2 of Basu et al. (1998), we have that $\hat{\theta}$ is a consistent estimator of the parameter θ, with

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow[n\to\infty]{L} N(0, AVar(\theta_0)),$$

and asymptotic variance

$$AVar(\theta_0) = J^{-1}(\theta_0)K(\theta_0)J^{-1}(\theta_0),$$

where $J(\theta_0)$ and $K(\theta_0)$ are given by

$$J(\theta_0) = \int u_{\theta_0}(x)u^t_{\theta_0}(x)f^{1+a}_{\theta_0}(x)d\mu(x),$$

and

$$\begin{aligned} K(\theta_0) &= \int u_{\theta_0}(x)u^t_{\theta_0}(x)f^{1+2a}_{\theta_0}(x)d\mu(x) \\ &\quad - \int u_{\theta_0}(x)f^{1+a}_{\theta_0}(x)d\mu(x)\int u^t_{\theta_0}(x)f^{1+a}_{\theta_0}(x)d\mu(x). \end{aligned}$$

In view of Corollary 2.1 of Dik and Gunst (1985), we obtain

$$n(\hat{\theta}-\theta_0)^T J^{\omega}(\theta_0)(\hat{\theta}-\theta_0) \xrightarrow[n\to\infty]{L} \sum_{i=1}^{r}\beta_i Z_i^2,$$

where $Z_1, \dots Z_r$, are iid standard normal random variables,

$$r = rank(AVar(\theta_0)J^{\omega}(\theta_0)AVar(\theta_0)),$$

and $\beta_1, \beta_2, \dots, \beta_r$ denote the non-zero eigenvalues of the matrix

$$J^{\omega}(\theta_0)AVar(\theta_0).$$

Taking into account the above discussion, we finally have

$$E_g[n(\hat{\theta}-\theta_0)^t J^{\omega}(\theta_0)(\hat{\theta}-\theta_0)] = \sum_{i=1}^{r}\beta_i. \tag{7.8}$$

7.3. Proof of (3.20) This subsection provides a proof of (3.20) in Remark 3.

Proof. We have (cf. Basu et al., 1998),

$$
\begin{aligned}
&D_0^\omega(g, f_\theta) = \lim_{a\to 0} D_a^\omega(g, f_\theta)\\
&= \lim_{a\to 0}\left\{\int\limits_{\mathcal{X}} h_\omega(x)\left(f_\theta^{1+a}(x) - (1+\frac{1}{a})g(x)f_\theta^a(x) + \frac{1}{a}g^{1+a}(x)\right)dx\right\}\\
&= \lim_{a\to 0}\int\limits_{\mathcal{X}} h_\omega(x)f_\theta^{1+a}(x)dx - \lim_{a\to 0}\int\limits_{\mathcal{X}} h_\omega(x)g(x)f_\theta^a(x)dx\\
&+\lim_{a\to 0}\int\limits_{\mathcal{X}} h_\omega(x)\frac{g(x)(g^a(x) - f_\theta^a(x))}{\alpha}dx\\
&= \int\limits_{\mathcal{X}} h_\omega(x)f_\theta(x)dx - \int\limits_{\mathcal{X}} h_\omega(x)g(x)dx + \int\limits_{\mathcal{X}} h_\omega(x)g(x)\lim_{a\to 0}\frac{(g^a(x) - f_\theta^a(x))}{\alpha}dx\\
&= E_{f_\theta}(h_\omega(X)) - E_g(h_\omega(X)) + \int\limits_{\mathcal{X}} h_\omega(x)g(x)\lim_{a\to 0}\{g^a(x)\log g(x)\\
&-f_\theta^a(x)\log f_\theta(x)\}dx\\
&= E_{f_\theta}(h_\omega(X)) - E_g(h_\omega(X)) + \int\limits_{\mathcal{X}} h_\omega(x)g(x)\log\frac{g(x)}{f_\theta(x)}dx.
\end{aligned}
$$

Acknowledgments. The authors would like to thank three referees for their helpful comments and suggestions which improved the exposition of an earlier version of the manuscript.

References

AITKIN, M. (2001). Likelihood and Bayesian analysis of mixtures. *Stat. Model.* **1**, 287–304.

AKAIKE, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Proceeding of the Second International Symposium on Information Theory*, (Petrov, B.N. and Csaki, F., eds.). Akademiai Kaido, Budapest.

AVLOGIARIS, G., MICHEAS, A. and ZOGRAFOS, K. (2016a). On local divergences between two probability measures. *Metrika* **79**, 303–333.

AVLOGIARIS, G., MICHEAS, A. and ZOGRAFOS, K. (2016b). On testing local hypotheses via local divergence. *Stat. Methodol.* **31**, 20–42.

BASU, A., HARRIS, I.R., HJORT, N.L. and JONES, M.C. (1998). Robust and efficient estimation by minimizing a density power divergence. *Biometrika* **85**, 549–559.

BASU, A., SHIOYA, H. and PARK, C. (2011). *Statistical Inference: the Minimum Distance Approach.* Chapman & Hall/CRC, London.

BENGTSSON, T. and CAVANAUGH, J.E. (2006). An improved Akaike information criterion for state-space model selection. *Comput. Statist. Data Anal.* **50**, 2635–2654.

BURNHAM, P.K. and ANDERSON, R.D. (2002). *Model Selection and Multimodel Inference a Practical Information-Theoretic Approach*, 2nd edn. Springer, Berlin.

CAVANAUGH, J.E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Aust. N. Z. J. Stat.* **46**, 257–274.

CLAESKENS, G. and HJORT, N.L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, 464, 900–916.

CLAESKENS, G. and HJORT, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

CRESSIE, N. (1993). *Statistics for Spatial Data*, 2nd edn. Wiley, New York.

CRESSIE, N. and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**, 440–464.

CSISZÁR, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **8**, 85–108.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2**, 299–318.

DEMPSTER, P.A., LAIRD, M.N. and RUBIN, B.D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38.

DIK, J.J. and GUNST, M.C.M. (1985). The distribution of general quadratic forms in normal variables. *Stat. Neerl.* **39**, 14–26.

DIGGLE, P.J. (2013). *Statistical Analysis of Spatial and Spatial-Temporal Point Patterns*, 3rd edn. CRC Press, Boca Raton.

FISHER, R.A. (1936). The use of multiple measurements Axonomic problems. *Ann. Eugen.* **7**, 179–188.

JIMÉNEZ-GAMERO, M.D., PINO-MEJIAS, R., ALBA-FERNÁNDEZ, V. and MORENO REBOLLO, J.L. (2011). Minimum φ-divergence estimation in misspecified multinomial models. *Comput. Statist. Data Anal.* **55**, 3365–3378.

KAGAN, A.M. (1963). On the theory of Fisher's information quantity. *Dokl. Akad. Nauk SSSR* **151**, 277–278.

KONISHI, S. and KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

KULLBACK, S. and LEIBLER, R.A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22**, 79–86.

LEWIS, P.A.W. and SHEDLER, G.S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist.* **26**, 3, 403–413.

MATTHEOU, K., LEE, S. and KARAGRIGORIOU, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *J. Statist. Plann. Inference* **139**, 228–235.

MCLACHLAN, G. and PEAL, D. (2000). *Finite Mixture Models.* Wiley, New York.

MICHEAS, A. (2014). Hierarchical Bayesian modeling of marked non-homogeneous Poisson processes with finite mixtures and inclusion of covariate information. *J. Appl. Stat.* **41**, 12, 2596–2615.

NIELSEN, F. and NOCK, R. (2011). On Rényi and Tsallis entropies and divergences for exponential families. arXiv:1105.3259v1 [cs.IT] 17 May 2011.

PARDO, L. (2006). *Statistical Inference Based on Divergence Measures.* Chapman & Hall/CRC, London.

POSTMAN, M., GELLER, M. and HUCHRA, J. (1986). The cluster-cluster correlation function. *The Astronomical Journal* **91**, 1267–1273.

ROEDER, K. (1992). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of Statistical Association* **85**, 617–624.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

SEGHOUANE, A.K. and BEKARA, M. (2004). A small sample model selection based on the Kullback symmetric divergence. *IEEE Trans. Signal Process* **52**, 3314–3323.

SHANG, J. (2008). Selection criteria based on Monte Carlo simulation and cross validation in mixed models. *Far East J. Theor. Stat.* **25**, 51–72.

SHANG, J. and CAVANAUGH, J.E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Comput. Statist. Data Anal.* **52**, 2004–2021.

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. B* **64**, 583–639.

STRAUSS, D.J. (1975). A model for clustering. *Biometrika* **62**, 467–475.

TAKEUCHI, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences)* **153**, 12–18. (in Japanese).

TOMA, A. (2014). Model selection criteria using divergences. *Entropy* **16**, 5, 2686–2698.

TOMA, A. and BRONIATOWSKI, M. (2011). Dual divergence estimators and tests: robustness results. *J. Multivariate Anal.* **102**, 1, 20–36.

VUONG, Q.H. and WANG, W. (1993). Minimum chi-square estimation and tests for model selection. *J. Econometrics* **56**, 141–168.

Appendix A

A.1 Estimated Parameters of the Candidate Models in the Redwood Trees Application

Model 1: bivariate normal with estimated parameters $\hat{\mu}_1 = (0.26, 0.25)$ and $\hat{\Sigma}_1 = \begin{pmatrix} 0.0170 & 0.0049 \\ 0.0049 & 0.0179 \end{pmatrix}$.

Model 2: mixture of two bivariate normal components with estimated parameters $\hat{w}_2 = 0.598$, $\hat{\mu}_2 = (0.21, 0.17)$, $\hat{\Sigma}_2 = \begin{pmatrix} 0.0156 & -0.0032 \\ -0.0032 & 0.0068 \end{pmatrix}$, $\hat{w}_3 = 0.402$, $\hat{\mu}_3 = (0.32, 0.36)$ and $\hat{\Sigma}_3 = \begin{pmatrix} 0.0087 & 0.0007 \\ 0.0007 & 0.0056 \end{pmatrix}$.

Model 3: mixture of three bivariate normal components with estimated parameters $\hat{w}_4 = 0.2764$, $\hat{\mu}_4 = (0.12, 0.27)$, $\hat{\Sigma}_5 = \begin{pmatrix} 0.0036 & 0.0040 \\ 0.0040 & 0.0084 \end{pmatrix}$, $\hat{w}_6 = 0.2506$, $\hat{\mu}_6 = (0.34, 0.36)$, $\hat{\Sigma}_6 = \begin{pmatrix} 0.0083 & -0.0016 \\ -0.0016 & 0.0123 \end{pmatrix}$, $\hat{w}_7 = 0.4730$, $\hat{\mu}_7 = (0.28, 0.22)$ and $\hat{\Sigma}_7 = \begin{pmatrix} 0.0105 & 0.0079 \\ 0.0079 & 0.0142 \end{pmatrix}$.

Model 4: mixture of four bivariate normal components with estimated parameters $\hat{w}_8 = 0.1565$, $\hat{\mu}_8 = (0.38, 0.18)$, $\hat{\Sigma}_8 = \begin{pmatrix} 0.0055 & 0.0000 \\ 0.0000 & 0.0105 \end{pmatrix}$, $\hat{w}_9 = 0.3004$, $\hat{\mu}_9 = (0.34, 0.35)$, $\hat{\Sigma}_9 = \begin{pmatrix} 0.0050 & 0.0047 \\ 0.0047 & 0.0071 \end{pmatrix}$, $\hat{w}_{10} =$

0.2715, $\hat{\mu}_{10} = (0.12, 0.29)$, $\hat{\Sigma}_{10} = \begin{pmatrix} 0.0040 & 0.0049 \\ 0.0049 & 0.0083 \end{pmatrix}$, $\hat{w}_{11} = 0.2715$, $\hat{\mu}_{11} = (0.22, 0.15)$ and $\hat{\Sigma}_{11} = \begin{pmatrix} 0.0053 & 0.0029 \\ 0.0029 & 0.0049 \end{pmatrix}$.

Model 5: mixture of five bivariate normal components with estimated parameters $\hat{w}_{12} = 0.2588$, $\hat{\mu}_{12} = (0.12, 0.28)$, $\hat{\Sigma}_{12} = \begin{pmatrix} 0.0037 & 0.0046 \\ 0.0046 & 0.0077 \end{pmatrix}$, $\hat{w}_{13} = 0.1435$, $\hat{\mu}_{13} = (0.40, 0.17)$, $\hat{\Sigma}_{13} = \begin{pmatrix} 0.0045 & 0.0010 \\ 0.0010 & 0.0094 \end{pmatrix}$, $\hat{w}_{14} = 0.2776$, $\hat{\mu}_{14} = (0.33, 0.34)$, $\hat{\Sigma}_{14} = \begin{pmatrix} 0.0050 & 0.0052 \\ 0.0052 & 0.0073 \end{pmatrix}$, $\hat{w}_{15} = 0.0916$, $\hat{\mu}_{15} = (0.24, 0.25)$, $\hat{\Sigma}_{15} = \begin{pmatrix} 0.0057 & 0.0057 \\ 0.0057 & 0.0095 \end{pmatrix}$, $\hat{w}_{16} = 0.2284$, $\hat{\mu}_{16} = (0.23, 0.15)$ and $\hat{\Sigma}_{16} = \begin{pmatrix} 0.0049 & 0.0034 \\ 0.0034 & 0.0055 \end{pmatrix}$.

Model 6: mixture of six bivariate normal components with estimated parameters $\hat{w}_{17} = 0.2478$, $\hat{\mu}_{17} = (0.12, 0.28)$, $\hat{\Sigma}_{17} = \begin{pmatrix} 0.0036 & 0.0044 \\ 0.0044 & 0.0074 \end{pmatrix}$, $\hat{w}_{18} = 0.0883$, $\hat{\mu}_{18} = (0.25, 0.29)$, $\hat{\Sigma}_{18} = \begin{pmatrix} 0.0061 & 0.0067 \\ 0.0067 & 0.0108 \end{pmatrix}$, $\hat{w}_{19} = 0.2527$, $\hat{\mu}_{19} = (0.33, 0.34)$, $\hat{\Sigma}_{19} = \begin{pmatrix} 0.0054 & 0.0059 \\ 0.0059 & 0.0083 \end{pmatrix}$, $\hat{w}_{20} = 0.1372$, $\hat{\mu}_{20} = (0.41, 0.16)$, $\hat{\Sigma}_{20} = \begin{pmatrix} 0.0039 & 0.0016 \\ 0.0016 & 0.0092 \end{pmatrix}$, $\hat{w}_{21} = 0.1943$, $\hat{\mu}_{21} = (0.23, 0.16)$, $\hat{\Sigma}_{21} = \begin{pmatrix} 0.0046 & 0.0042 \\ 0.0042 & 0.0067 \end{pmatrix}$, $\hat{w}_{22} = 0.0797$, $\hat{\mu}_{22} = (0.24, 0.20)$ and $\hat{\Sigma}_{22} = \begin{pmatrix} 0.0055 & 0.0059 \\ 0.0059 & 0.0090 \end{pmatrix}$.

G. Avlogiaris
K. Zografos
Department of Mathematics,
University of Ioannina,
Ioannina, Greece
E-mail: avlo2000@yahoo.gr
kzograf@uoi.gr

A. C. Micheas
Department of Statistics,
University of Missouri,
Columbia, MO, USA
E-mail: micheasa@missouri.edu

Paper received: 12 July 2017.