



Article Robust Model Selection Criteria Based on Pseudodistances

Aida Toma ^{1,2,*}, Alex Karagrigoriou ³ and Paschalini Trentou ³

- ¹ Department of Applied Mathematics, Bucharest University of Economic Studies, 010164 Bucharest, Romania
- ² "Gh. Mihoc C. Iacob" Institute of Mathematical Statistics and Applied Mathematics, Romanian Academy, 010164 Bucharest, Romania
- ³ Department of Statistics and Actuarial-Financial Mathematics, Lab of Statistics and Data Analysis, University of the Aegean, 83200 Karlovasi, Greece; alex.karagrigoriou@aegean.gr (A.K.); sasm17025@sas.aegean.gr (P.T.)
- * Correspondence: aida.toma@csie.ase.ro

Received: 17 February 2020; Accepted: 3 March 2020; Published: 6 March 2020



Abstract: In this paper, we introduce a new class of robust model selection criteria. These criteria are defined by estimators of the expected overall discrepancy using pseudodistances and the minimum pseudodistance principle. Theoretical properties of these criteria are proved, namely asymptotic unbiasedness, robustness, consistency, as well as the limit laws. The case of the linear regression models is studied and a specific pseudodistance based criterion is proposed. Monte Carlo simulations and applications for real data are presented in order to exemplify the performance of the new methodology. These examples show that the new selection criterion for regression models is a good competitor of some well known criteria and may have superior performance, especially in the case of small and contaminated samples.

Keywords: model selection; minimum pseudodistance estimation; Robustness

1. Introduction

Model selection is fundamental to the practical applications of statistics and there is a substantial literature on this issue. Classical model selection criteria include, among others, the C_p -criterion, the Akaike Information Criterion (AIC), based on the Kullback-Leibler divergence, and the Bayesian Information Criterion (BIC) as well as a General Information Criterion (GIC) which corresponds to a general class of criteria which also estimates the Kullback-Leibler divergence. These criteria have been proposed respectively in [1–4], and represent powerful tools for choosing the best model among different candidate models that can be used to fit a given data set. On the other hand, many classical procedures for model selection are extremely sensitive to outliers and to other departures from the distributional assumptions of the model. Robust versions of classical model selection criteria, which are not strongly affected by outliers, have been proposed for example in [5–7]. Some recent proposals for robust model selection are criteria based on divergences and minimum divergence estimators. We recall here, the Divergence Information Criteria (DIC) based on the density power divergences introduced in [8], the Modified Divergence Information Criteria (MDIC) introduced in [9] and the criteria based on minimum dual divergence estimators introduced in [10].

The interest on statistical methods based on divergence measures has grown significantly in recent years. For a wide variety of models, statistical methods based on divergences have high model efficiency and are also robust, representing attractive alternatives to the classical methods. We refer to the monographs [11,12] for an excellent presentation of such methods, for their importance and applications. The pseudodistances that we use in the present paper were originally introduced in [13],

where they are called "type-0" divergences, and corresponding minimum divergence estimators have been studied. They are also presented and extensively studied in [14] where they are called γ -divergences, as well as in [15] in the context of decomposable pseudodistances. Like divergences, the pseudodistances are not mathematical metrics in the strict sense of the term. They satisfy two properties, namely the nonnegativity and the fact that the pseudodistance between two probability measures equals to zero if and only if the two measures are equal. The divergences are moreover characterized by the information processing property, that is, the complete invariance with respect to statistically sufficient transformations of the observation space. In general, a pseudodistance may not satisfy this property. We have adopted the term pseudodistance for this reason, but in literature we can also encounter the other terms mentioned above.

The pseudodistances that we consider in this paper have also been used to define robustness and efficiency measures, as well as the corresponding optimal robust M-estimators following the Hampel's infinitesimal approach in [16]. The minimum pseudodistance estimators for general parametric models have been studied in [15] and consist of minimizing an empirical version of a pseudodistance between the assumed theoretical model and the true model underlying the data. These estimators have the advantage of not requiring any prior smoothing and conciliate robustness with high efficiency, providing a high degree of stability under model misspecification, often with a minimal loss in model efficiency. Such estimators are also defined and studied in the case of the multivariate normal model, as well as for linear regression models in [17,18], where applications for portfolio optimization models are also presented.

In the present paper we propose new criteria for model selection, based on pseudodistances and on minimum pseudodistance estimators. These new criteria have robustness properties, are asymptotically unbiased, consistent and compare well with some other known model selection criteria, even for small samples.

The paper is organized as follows—Section 2 is devoted to minimum pseudodistance estimators and to their asymptotic properties, which will be needed in the next sections. Section 3 presents new estimators of the expected overall discrepancy using pseudodistances, together with corresponding theoretical properties including robustness, consistency and limit laws. The new asymptotically unbiased model selection criteria are presented in Section 3.3, where the case of the univariate normal model and the case of linear regression models are investigated. Applications based on Monte Carlo simulations and on real data, illustrating the performance of the new methodology in the case of linear regression models, are included in Section 4.

2. Minimum Pseudodistance Estimators

The construction of new model selection criteria is based on using the following family of pseudodistances (see [15]). For two probability measures *P* and *Q* admitting densities *p* and *q* respectively with respect to the Lebesgue measure, the family of pseudodistances of order $\gamma > 0$ is defined by

$$R_{\gamma}(P,Q) = \frac{1}{\gamma+1} \ln\left(\int p^{\gamma} dP\right) + \frac{1}{\gamma(\gamma+1)} \ln\left(\int q^{\gamma} dQ\right) - \frac{1}{\gamma} \ln\left(\int p^{\gamma} dQ\right)$$
(1)

and satisfies the limit relation

$$\lim_{\gamma \to 0} R_{\gamma}(P,Q) = R_0(P,Q), \tag{2}$$

where $R_0(P, Q) := \int \ln \frac{q}{p} dQ$ is the modified Kullback-Leibler divergence.

Let (P_{θ}) be a parametric model indexed by $\theta \in \Theta$, where Θ is a *d*-dimensional parameter space, and p_{θ} be the corresponding densities with respect to the Lebesgue measure λ . Let X_1, \ldots, X_n be a random sample on P_{θ_0} , $\theta_0 \in \Theta$. For $\gamma > 0$ fixed, a minimum pseudodistance estimator of the unknown parameter θ_0 from the law P_{θ_0} is defined by replacing the measure P_{θ_0} in the pseudodistance $R_{\gamma}(P_{\theta}, P_{\theta_0})$ by the empirical measure P_n pertaining to the sample, and then minimizing this empirical quantity with respect to θ on the parameter space. Since the middle term in $R_{\gamma}(P_{\theta}, P_{\theta_0})$ does not depend on θ , these estimators are defined by

$$\widehat{\theta}_{n} = \arg\min_{\theta\in\Theta} \left\{ \frac{1}{\gamma+1} \ln\left(\int p_{\theta}^{\gamma+1} d\lambda\right) - \frac{1}{\gamma} \ln\left(\frac{1}{n} \sum_{i=1}^{n} p_{\theta}^{\gamma}(X_{i})\right), \right\}$$
(3)

or equivalently as

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \{ C_{\gamma}(\theta)^{-1} \cdot \frac{1}{n} \sum_{i=1}^n p_{\theta}^{\gamma}(X_i) \},$$
(4)

where $C_{\gamma}(\theta) = (\int p_{\theta}^{\gamma+1} d\lambda)^{\gamma/(\gamma+1)}$. Denoting $h(x,\theta) := C_{\gamma}(\theta)^{-1} \cdot p_{\theta}^{\gamma}(x)$, these estimators can be written as

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n h(X_i, \theta).$$
(5)

The optimum given above need not be uniquely defined. On the other hand,

$$\arg\max_{\theta\in\Theta}\int h(x,\theta)dP_{\theta_0}(x) = \theta_0$$
(6)

and here θ_0 is the unique optimizer, since $R_{\gamma}(P_{\theta}, P_{\theta_0}) = 0$ implies $\theta = \theta_0$.

Define

$$R_{\gamma}(\theta_0) := \max_{\theta \in \Theta} \int h(x,\theta) dP_{\theta_0}(x) = \int h(x,\theta_0) dP_{\theta_0}(x).$$

An estimator of $R_{\gamma}(\theta_0)$ is defined by

$$\widehat{R}_{\gamma}(\theta_0) := \max_{\theta \in \Theta} \int h(x,\theta) dP_n(x) = \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n h(X_i,\theta) = \frac{1}{n} \sum_{i=1}^n h(X_i,\widehat{\theta}_n).$$
(7)

The following regularity conditions of the model will be assumed throughout the rest of the paper.

(C1) The density $p_{\theta}(x)$ has continuous partial derivatives with respect to θ up to the third order (for all *x* λ -a.e.).

(C2) There exists a neighborhood N_{θ_0} of θ_0 such that the first-, the second- and the third- order partial derivatives with respect to θ of $h(x, \theta)$ are dominated on N_{θ_0} by some P_{θ_0} -integrable functions.

(C3) The integrals $\int \left[\frac{\partial^2}{\partial \theta^2} h(x,\theta)\right]_{\theta=\theta_0} dP_{\theta_0}(x)$ and $\int \left[\frac{\partial}{\partial \theta} h(x,\theta)\right]_{\theta=\theta_0} \left[\frac{\partial}{\partial \theta} h(x,\theta)\right]_{\theta=\theta_0}^t dP_{\theta_0}(x)$ exist.

Theorem 1. Assume that conditions (C1), (C2) and (C3) are fulfilled. Then

- (a) Let $B := \left\{ \theta \in \Theta; \|\theta \theta_0\| \le n^{-1/3} \right\}$. Then, as $n \to \infty$, with probability one, the function $\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta)$ attains a local maximal value at some point $\hat{\theta}_n$ in the interior of B, which implies that the estimator $\hat{\theta}_n$ is $n^{1/3}$ -consistent.
- (b) $\sqrt{n}\left(\hat{\theta}_n \theta_0\right)$ converges in distribution to a centered multivariate normal random variable with covariance matrix

$$V = S^{-1} M S^{-1}, (8)$$

where $S := -\int [\frac{\partial^2}{\partial \theta^2} h(x,\theta)]_{\theta=\theta_0} dP_{\theta_0}(x)$ and $M := \int [\frac{\partial}{\partial \theta} h(x,\theta)]_{\theta=\theta_0} [\frac{\partial}{\partial \theta} h(x,\theta)]_{\theta=\theta_0}^t dP_{\theta_0}(x)$. $\sqrt{n} \left(\widehat{R}_{\alpha}(\theta_0) - R_{\alpha}(\theta_0)\right)$ converges in distribution to a centered normal variable with variance $\sigma^2(\theta_0)$.

(c)
$$\sqrt{n} \left(\widehat{R}_{\gamma}(\theta_0) - R_{\gamma}(\theta_0) \right)$$
 converges in distribution to a centered normal variable with variance $\sigma^2(\theta_0) = \int h(x,\theta_0)^2 dP_{\theta_0}(x) - \left(\int h(x,\theta_0) dP_{\theta_0}(x) \right)^2$.

We refer to [15] for details regarding these estimators and for the proofs of the above asymptotic properties.

3. Model Selection Criteria Based on Pseudodistances

Model selection is a method for selecting the best model among candidate models that can be used to fit a given data set. A model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy, a nonnegative quantity which measures the distance between the true unknown model and a fitted approximating model. If the value of the criterion is small, then the approximated candidate model can be chosen. In the following, by applying the same methodology used for AIC, we construct new criteria for model selection using pseudodistances (1) and minimum pseudodistance estimators.

Let X_1, \ldots, X_n be a random sample from the distribution associated with the true model Q with density q and let p_{θ} be the density of a candidate model P_{θ} from a parametric family (P_{θ}) , where $\theta \in \Theta \subset \mathbb{R}^d$.

3.1. The Expected Overall Discrepancy

For $\gamma > 0$ fixed, we consider the quantity

$$W_{\theta} = \frac{1}{\gamma + 1} \ln\left(\int p_{\theta}^{\gamma + 1} d\lambda\right) - \frac{1}{\gamma} \ln\left(\int p_{\theta}^{\gamma} q d\lambda\right),\tag{9}$$

which is the same as the pseudodistance $R_{\gamma}(P_{\theta}, Q)$ without the middle term that remains constant irrespectively of the model (P_{θ}) used.

The target theoretical quantity that will be approximated by an asymptotically unbiased estimator is given by

$$E[W_{\widehat{\theta}_n}] = E[W_{\theta}|\theta = \widehat{\theta}_n], \tag{10}$$

where $\hat{\theta}_n$ is a minimum pseudodistance estimator defined as in (3). The same pseudodistance is used for both W_{θ} and $\hat{\theta}_n$. The quantity (10) can be seen as an average distance between Q and (P_{θ}) up to a constant and is called *the expected overall discrepancy* between Q and (P_{θ}) .

The next Lemma gives the gradient vector and the Hessian matrix of W_{θ} and is useful for the evaluation of $E[W_{\hat{\theta}_{w}}]$ through Taylor expansion.

Throughout this paper, for a scalar function $\varphi_{\theta}(\cdot)$, the quantity $\frac{\partial}{\partial \theta}\varphi_{\theta}(\cdot)$ denotes the *d*-dimensional gradient vector of $\varphi_{\theta}(\cdot)$ with respect to the vector θ and $\frac{\partial^2}{\partial \theta^2}\varphi_{\theta}(\cdot)$ denotes the corresponding $d \times d$ Hessian matrix. We also use the notations $\dot{\varphi}_{\theta}$ and $\ddot{\varphi}_{\theta}$ for the first and the second order derivatives of φ_{θ} with respect to θ .

We assume the following conditions allowing derivation under the integral sign:

(C4) There exists a neighborhood N_{θ} of θ such that

$$\int \sup_{t\in N_\theta} \left\| \frac{\partial}{\partial t} p_t^{\gamma+1} \right\| \mathrm{d}\lambda < \infty, \quad \int \sup_{t\in N_\theta} \left\| \frac{\partial}{\partial t} [p_t^{\gamma} \dot{p}_t] \right\| \mathrm{d}\lambda < \infty.$$

(C5) There exists a neighborhood N_{θ} of θ such that

$$\int \sup_{t \in N_{\theta}} \left\| \frac{\partial}{\partial t} p_t^{\gamma} \right\| q \mathrm{d}\lambda < \infty, \quad \int \sup_{t \in N_{\theta}} \left\| \frac{\partial}{\partial t} [p_t^{\gamma-1} \dot{p}_t] \right\| q \mathrm{d}\lambda < \infty.$$

Lemma 1. Under (C4) and (C5), the gradient vector and the Hessian matrix of W_{θ} are

$$\frac{\partial}{\partial \theta} W_{\theta} = \frac{\int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda}{\int p_{\theta}^{\gamma+1} d\lambda} - \frac{\int p_{\theta}^{\gamma-1} \dot{p}_{\theta} q d\lambda}{\int p_{\theta}^{\gamma} q d\lambda}$$
(11)

$$\frac{\partial^2}{\partial \theta^2} W_{\theta} = \frac{ \left[\gamma \int p_{\theta}^{\gamma-1} \dot{p}_{\theta} \dot{p}_{\theta}^t d\lambda + \int p_{\theta}^{\gamma} \ddot{p}_{\theta} d\lambda \right] \int p_{\theta}^{\gamma+1} d\lambda - (\gamma+1) \int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda (\int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda)^t}{(\int p_{\theta}^{\gamma+1} d\lambda)^2} \\ - \frac{ \left[(\gamma-1) \int p_{\theta}^{\gamma-2} \dot{p}_{\theta} \dot{p}_{\theta}^t q d\lambda + \int p_{\theta}^{\gamma-1} \ddot{p}_{\theta} q d\lambda \right] \int p_{\theta}^{\gamma} q d\lambda - \gamma \int p_{\theta}^{\gamma-1} \dot{p}_{\theta} q d\lambda (\int p_{\theta}^{\gamma-1} \dot{p}_{\theta} q d\lambda)^t}{(\int p_{\theta}^{\gamma} q d\lambda)^2}.$$

When the true model *Q* belongs to the parametric model (P_{θ}) , hence $Q = P_{\theta_0}$ and $q = p_{\theta_0}$, the gradient vector and the Hessian matrix of W_{θ} simplify to

$$\left[\frac{\partial}{\partial \theta}W_{\theta}\right]_{\theta=\theta_{0}} = 0 \tag{12}$$

$$\left[\frac{\partial^2}{\partial \theta^2} W_{\theta}\right]_{\theta=\theta_0} = M_{\gamma}, (\theta_0)$$
(13)

where

$$M_{\gamma}(\theta_0) := \frac{\left(\int p_{\theta_0}^{\gamma-1} \dot{p}_{\theta_0} \dot{p}_{\theta_0}^t \mathrm{d}\lambda\right) \left(\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda\right) - \left(\int p_{\theta_0}^{\gamma} \dot{p}_{\theta_0} \mathrm{d}\lambda\right) \left(\int p_{\theta_0}^{\gamma} \dot{p}_{\theta_0} \mathrm{d}\lambda\right)^t}{\left(\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda\right)^2}.$$
(14)

In the following Propositions we suppose that the true model Q belongs to the parametric model (P_{θ}) , hence $Q = P_{\theta_0}$, $q = p_{\theta_0}$ and θ_0 is the value of the parameter corresponding to the true model $Q = P_{\theta_0}$. We also say that θ_0 is the true value of the parameter.

Proposition 1. When the true model Q belongs to the parametric model (P_{θ}) , assuming that (C4) and (C5) are fulfilled for $q = p_{\theta_0}$ and $\theta = \theta_0$, the expected overall discrepancy is given by

$$E[W_{\widehat{\theta}_n}] = W_{\theta_0} + \frac{1}{2}E[(\widehat{\theta}_n - \theta_0)^t M_{\gamma}(\theta_0)(\widehat{\theta}_n - \theta_0)] + E[R_n],$$
(15)

where $R_n = o(\|\widehat{\theta}_n - \theta_0\|^2)$, $M_{\gamma}(\theta_0)$ is given by (14).

3.2. Estimation of the Expected Overall Discrepancy

In this section, we introduce an estimator of the expected overall discrepancy, under the hypothesis that the true model Q belongs to the parametric model (P_{θ}) . Hence, $Q = P_{\theta_0}$ and the unknown parameter θ_0 will be estimated by a minimum pseudodistance estimator $\hat{\theta}_n$.

For a given $\theta \in \Theta$, a natural estimator of W_{θ} is defined by

$$Q_{\theta} := \frac{1}{\gamma + 1} \ln \left(\int p_{\theta}^{\gamma + 1} d\lambda \right) - \frac{1}{\gamma} \ln \left(\frac{1}{n} \sum_{i=1}^{n} p_{\theta}^{\gamma}(X_i) \right).$$
(16)

Lemma 2. Assuming (C4), the gradient vector and the Hessian matrix of Q_{θ} are given by

$$\begin{split} \frac{\partial}{\partial \theta} Q_{\theta} &= \frac{\int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda}{\int p_{\theta}^{\gamma+1} d\lambda} - \frac{\sum_{i=1}^{n} p_{\theta}^{\gamma-1}(X_{i}) \dot{p}_{\theta}(X_{i})}{\sum_{i=1}^{n} p_{\theta}^{\gamma}(X_{i})} \\ \frac{\partial^{2}}{\partial \theta^{2}} Q_{\theta} &= \frac{\left[\gamma \int p_{\theta}^{\gamma-1} \dot{p}_{\theta} \dot{p}_{\theta}^{t} d\lambda + \int p_{\theta}^{\gamma} \ddot{p}_{\theta} d\lambda\right] \int p_{\theta}^{\gamma+1} d\lambda - (\gamma+1) \int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda (\int p_{\theta}^{\gamma} \dot{p}_{\theta} d\lambda)^{t}}{(\int p_{\theta}^{\gamma+1} d\lambda)^{2}} - \\ - \frac{\left[(\gamma-1) \sum_{i=1}^{n} p_{\theta}^{\gamma-2}(X_{i}) \dot{p}_{\theta}(X_{i}) \dot{p}_{\theta}(X_{i})^{t} + \sum_{i=1}^{n} p_{\theta}^{\gamma-1}(X_{i}) \ddot{p}_{\theta}(X_{i})\right] \sum_{i=1}^{n} p_{\theta}^{\gamma}(X_{i})}{(\sum_{i=1}^{n} p_{\theta}^{\gamma}(X_{i}))^{2}} \\ + \frac{\gamma(\sum_{i=1}^{n} p_{\theta}^{\gamma-1}(X_{i}) \dot{p}_{\theta}(X_{i}))(\sum_{i=1}^{n} p_{\theta}^{\gamma-1}(X_{i}) \dot{p}_{\theta}(X_{i}))^{t}}{(\sum_{i=1}^{n} p_{\theta}^{\gamma}(X_{i}))^{2}}. \end{split}$$

Proposition 2. When the true model Q belongs to the parametric model (P_{θ}) , by imposing the conditions (C1)-(C5), it holds

$$E[Q_{\theta_0}] = E[Q_{\widehat{\theta}_n}] + \frac{1}{2}E[(\theta_0 - \widehat{\theta}_n)^t M_{\gamma}(\theta_0)(\theta_0 - \widehat{\theta}_n)] + E[R_n],$$
(17)

where $R_n = o(\|\widehat{\theta}_n - \theta_0\|^2)$.

The following result allows to define an asymptotically unbiased estimator of the expected overall discrepancy.

Proposition 3. When the true model Q belongs to the parametric model (P_{θ}) , under (C1)-(C5), it holds

$$E[W_{\widehat{\theta}_{n}}] = E[Q_{\widehat{\theta}_{n}}] + E[(\theta_{0} - \widehat{\theta}_{n})^{t} M_{\gamma}(\theta_{0})(\theta_{0} - \widehat{\theta}_{n})] + \frac{1}{2\gamma n} \left[1 - \frac{\int p_{\theta_{0}}^{2\gamma+1} d\lambda}{\left(\int p_{\theta_{0}}^{\gamma+1} d\lambda\right)^{2}}\right] + E[R_{n}] + \frac{1}{\gamma} E[R'_{n}], \qquad (18)$$

where $R_n = o(\|\widehat{\theta}_n - \theta_0\|^2)$ and $R'_n = o(\|\frac{1}{n}\sum_{i=1}^n p_{\theta_0}^{\gamma}(X_i) - \int p_{\theta_0}^{\gamma+1} d\lambda\|^2).$

3.2.1. Limit Properties of the Estimator $Q_{\hat{\theta}_u}$

Under the hypothesis that the true model Q belongs to the family of models (P_{θ}) , hence $Q = P_{\theta_0}$, we prove the consistency and the asymptotic normality for the estimator $Q_{\hat{\theta}_n}$.

Note that

$$Q_{\widehat{\theta}_n} = \frac{1}{\gamma+1} \ln\left(\int p_{\widehat{\theta}_n}^{\gamma+1} d\lambda\right) - \frac{1}{\gamma} \ln\left(\frac{1}{n} \sum_{i=1}^n p_{\widehat{\theta}_n}^{\gamma}(X_i)\right)$$
(19)

$$= -\ln\left[\frac{\frac{1}{n}\sum_{i=1}^{n}p_{\widehat{\theta}_{n}}(X_{i})}{(\int p_{\widehat{\theta}_{n}}^{\gamma+1}d\lambda)^{\frac{\gamma}{\gamma+1}}}\right]^{\frac{1}{\gamma}} = -\ln[\widehat{R}_{\gamma}(\theta_{0})]^{\frac{1}{\gamma}},$$
(20)

where $\int p_{\hat{\theta}_n}^{\gamma+1} d\lambda = \left[\int p_{\theta}^{\gamma+1} d\lambda \right]_{\theta=\hat{\theta}_n}$ and $\hat{R}_{\gamma}(\theta_0)$ is given by (7).

First we prove that $\widehat{R}_{\gamma}(\theta_0)$ is a consistent estimator of $R_{\gamma}(\theta_0)$. Indeed, using Theorem 1 and the fact that $\int \frac{\partial}{\partial \theta} h(x, \theta_0) dP_{\theta_0}(x) = 0$, a Taylor expansion of $\frac{1}{n} \sum_{i=1}^n h(X_i, \theta)$ in $\widehat{\theta}_n$ around θ_0 gives

$$\widehat{R}_{\gamma}(\theta_0) = \frac{1}{n} \sum_{i=1}^n h(X_i, \theta_0) + o_P(n^{-1/2}).$$
(21)

Using the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}h(X_{i},\theta_{0}) = R_{\gamma}(\theta_{0}) + o_{P}(1).$$
(22)

Combining (21) and (22), we obtain that $\widehat{R}_{\gamma}(\theta_0)$ converges to $R_{\gamma}(\theta_0)$ in probability.

Then, using the continuous mapping theorem, since $g(t) = -\ln t^{\frac{1}{\gamma}}$ is a continuous function, we get

$$Q_{\widehat{\theta}_n} = -\ln[\widehat{R}_{\gamma}(\theta_0)]^{\frac{1}{\gamma}} \to -\ln[R_{\gamma}(\theta_0)]^{\frac{1}{\gamma}} = W_{\theta_0}$$

in probability.

On the other hand, using the asymptotic normality of the estimator $\hat{R}_{\gamma}(\theta_0)$ (according to Theorem 1 (c)) together with the univariate delta method, we obtain the asymptotic normality of $Q_{\hat{\theta}_n}$. The Proposition below summarizes the above asymptotic results.

Proposition 4. Under (C1)-(C3), when $Q = P_{\theta_0}$, it holds

(a) $Q_{\hat{\theta}_n}$ converges to W_{θ_0} in probability.

(b) $\sqrt{n}(Q_{\hat{\theta}_n} - W_{\theta_0})$ converges in distribution to a centered univariate normal random variable with variance $\frac{\sigma^2(\theta_0)}{\gamma^2 R_{\gamma}(\theta_0)^2}$, $\sigma^2(\theta_0)$ being defined in Theorem 1.

3.2.2. Robustness Properties of the Estimator $Q_{\hat{\theta}_n}$

The influence function is a useful tool for describing robustness of an estimator. Recall that, a map *T* defined on a set of probability measures and parameter space valued is a statistical functional corresponding to an estimator $\hat{\theta}_n$ of the parameter θ , whenever $\hat{\theta}_n = T(P_n)$, where P_n is the empirical measure associated to the sample. The influence function of *T* at P_{θ} is defined by

$$\mathrm{IF}(x;T,P_{\theta}):=\left.\frac{\partial T(\widetilde{P}_{\varepsilon x})}{\partial \varepsilon}\right|_{\varepsilon=0},$$

where $\tilde{P}_{\varepsilon x} := (1 - \varepsilon)P_{\theta} + \varepsilon \delta_x$, $\varepsilon > 0$, δ_x being the Dirac measure putting all mass at x. The gross error sensitivity of the estimator is defined by

$$\gamma^*(T, P_\theta) = \sup_x \|\mathrm{IF}(x; T, P_\theta)\|.$$

Whenever the influence function is bounded with respect to x, the corresponding estimator is called B-robust (see [19]).

In what follows, for a given $\gamma > 0$, we derive the influence function of the estimator $Q_{\hat{\theta}_n}$. The statistical functional associated with this estimator, which we denote by U, is defined by

$$U(P) := \frac{1}{\gamma + 1} \ln \left(\int p_{T(P)}^{\gamma + 1} d\lambda \right) - \frac{1}{\gamma} \ln \left(\int p_{T(P)}^{\gamma} dP \right),$$

where *T* is the statistical functional corresponding to the used minimum pseudodistance estimator estimator $\hat{\theta}_n$, namely

$$T(P) := \arg \sup_{\theta} C_{\gamma}(\theta)^{-1} \int p_{\theta}^{\gamma} dP$$

where $C_{\gamma}(\theta) = (\int p_{\theta}^{\gamma+1} d\lambda)^{\gamma/(\gamma+1)}$.

Due to the Fisher consistency of the functional *T*, according to (6), we have $T(P_{\theta_0}) = \theta_0$ which implies that $U(P_{\theta_0}) = W_{\theta_0}$.

Proposition 5. When $Q = P_{\theta_0}$, the influence function of $Q_{\hat{\theta}_n}$ is given by

$$\mathrm{IF}(x; U, P_{\theta_0}) = \frac{1}{\gamma} \left[1 - \frac{p_{\theta_0}^{\gamma}(x)}{\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda} \right].$$
(23)

Note that the influence function of the estimator $Q_{\hat{\theta}_n}$ does not depend on the estimator $\hat{\theta}_n$, but depends on the used pseudodistance. Usually, $p_{\theta_0}^{\gamma}(x)$ is bounded with respect to x and therefore $Q_{\hat{\theta}_n}$ is a robust estimator with respect to W_{θ_0} .



Figure 1. Influence functions in the case of the normal model.

For comparison at the level of the influence function, we consider the AIC criterion which is defined by

$$AIC = -2\ln(\mathcal{L}(\widehat{\theta}_n)) + 2d,$$

where $\mathcal{L}(\hat{\theta}_n)$ is the maximum value of the likelihood function for the model, $\hat{\theta}_n$ the maximum likelihood estimator and *d* the dimension of the parameter. The statistical functional corresponding to the statistic $-2\ln(\mathcal{L}(\hat{\theta}_n))$ is

$$V(P) = -2 \int \ln p_{T(P)} dP$$

where T here is the statistical functional corresponding to the maximum likelihood estimator. The influence function of the functional V is given by

$$IF(x; V, P_{\theta_0}) = 2 \left[\int \ln p_{\theta_0} dP_{\theta_0} - \ln p_{\theta_0}(x) \right].$$
(24)

This influence function is not bounded with respect to *x*, therefore the statistic $-2\ln(\mathcal{L}(\hat{\theta}_n))$ is not robust.

For example, in the case of the univariate normal model, for a positive γ , the influence function (23) writes as

IF
$$(x; U, P_{\theta_0}) = \frac{1}{\gamma} \left(1 - \sqrt{\gamma + 1} \cdot \exp\left(-\frac{\gamma}{2} \left(\frac{x - m}{\sigma}\right)^2\right) \right)$$
 (25)

while the influence function (24) writes as

$$\operatorname{IF}(x; V, P_{\theta_0}) = \left(\frac{x-m}{\sigma}\right)^2 - \frac{2m^2}{\sigma^2} - 1$$
(26)

(here $\theta_0 = (m, \sigma)$). For all the pseudodistances, the influence function (25) is bounded with respect to *x*, therefore the selection criteria based on the statistic $Q_{\hat{\theta}_n}$ will be robust. On the other hand, the influence function (26) is not bounded with respect to *x*, showing the non robustness of AIC in this case. Moreover, the gross error sensitivities corresponding to these influence functions are $\gamma^*(U, P_{\theta_0}) = \frac{1}{\gamma}$ and $\gamma^*(V, P_{\theta_0}) = \infty$. These results show that, in the case of the normal model, when γ increases the gross error sensitivity decreases. Therefore, larger values of γ are associated with more robust procedures. For the particular case m = 0 and $\sigma = 1$, the influence functions (25) and (26) are represented in Figure 1.

3.3. Model Selection Criteria Using Pseudodistances

3.3.1. The Case of Univariate Normal Family

The criteria that we propose in this section correspond to the case where the candidate model is a univariate normal model from the family of normal models (P_{θ}) indexed by $\theta = (\mu, \sigma)$. We also suppose that the true model Q belongs to (P_{θ}).

In the case of the univariate normal model, $M_{\gamma}(\theta_0)$ defined in (14) expresses as

$$M_{\gamma}(\theta_0) = \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} A(\gamma) V^{-1},$$
(27)

where V is the asymptotic covariance matrix given by (8) and the matrix $A(\gamma)$ is given by

$$A(\gamma) = \begin{pmatrix} 1 & 0 \\ 0 & rac{3\gamma^2 + 4\gamma + 2}{2(2\gamma + 1)} \end{pmatrix}.$$

For small positive values of γ , the matrix $A(\gamma)$ can be approximated by the identity matrix *I*. According to Theorem 1, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically multivariate normal and then the statistic $n(\theta_0 - \hat{\theta}_n)^t V^{-1}(\theta_0 - \hat{\theta}_n)$ has approximately a χ_d^2 distribution. For large *n*, it holds

$$E[(\theta_0 - \widehat{\theta}_n)^t M_{\gamma}(\theta_0)(\theta_0 - \widehat{\theta}_n)] \approx \frac{(\gamma + 1)^2}{(2\gamma + 1)^{3/2}} \cdot \frac{d}{n}.$$
(28)

Also, for the normal model, it holds

$$\frac{\int p_{\theta_0}^{2\gamma+1} \mathrm{d}\lambda}{\left(\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda\right)^2} = \frac{\gamma+1}{\sqrt{2\gamma+1}}.$$
(29)

Therefore, (18) becomes

$$E[W_{\hat{\theta}_n}] \cong E[Q_{\hat{\theta}_n}] + \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \cdot \frac{d}{n} + \frac{1}{2\gamma n} \left[1 - \frac{\gamma+1}{\sqrt{2\gamma+1}} \right] + E[R_n] + \frac{1}{\gamma} E[R'_n].$$
(30)

Using the central limit theorem and asymptotic properties of $\hat{\theta}_n$ given in Theorem 1, the following hold

$$n \cdot o(\|\widehat{\theta}_n - \theta_0\|^2) = o_P(1), \qquad (31)$$

$$n \cdot o(\|\frac{1}{n}\sum_{i=1}^{n} p_{\theta_0}^{\gamma}(X_i) - \int p_{\theta_0}^{\gamma+1} d\lambda\|^2) = o_P(1).$$
(32)

Using (30), (31) and (32) we obtain:

Proposition 6. For the univariate normal family, an asymptotically unbiased estimator of the expected overall discrepancy is given by

$$Q_{\hat{\theta}_n} + \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \cdot \frac{d}{n} + \frac{1}{2\gamma n} \left[1 - \frac{\gamma+1}{\sqrt{2\gamma+1}} \right],$$
(33)

where $\hat{\theta}_n$ is a minimum pseudodistance estimator given by (3).

Under the hypothesis that (P_{θ}) is the univariate normal model, as we supposed in this subsection, the function *h* writes as

$$h(x,\theta) = (\sqrt{\gamma+1})^{\gamma/(\gamma+1)} \cdot (\sigma\sqrt{2\pi})^{-\gamma/(\gamma+1)} \cdot \exp\left(-\frac{\gamma}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$
(34)

and it can be easily checked that all the conditions (C1)–(C5) are fulfilled. Therefore we can use all results presented in the preceding subsections, such that Proposition 6 is fully justified.

Moreover, the selection criteria based on (33) are consistent on the basis of Proposition 4. It should also be noted that the bias correction term in (33) decreases slowly as the parameter γ increases staying always very close to zero ($\sim 10^{-2}$). As expected, the larger the sample size the smaller the bias correction. As we saw in Section 3.2.2, since the gross error sensitivity of $Q_{\hat{\theta}_n}$ is $\gamma^*(U, P_{\theta_0}) = \frac{1}{\gamma}$, larger values of γ are associated with more robust procedures. On the other hand, the approximation of $A(\gamma)$ with the identity matrix is realized for values of γ close to zero. Thus, positive values of γ smaller than 0.5 for example could represent choices satisfying the robustness requirement and the approximation of $A(\gamma)$ through the identity matrix, approximation which is necessary to construct the criterion in this case.

3.3.2. The Case of Linear Regression Models

In the following, we adapt the pseudodistance based model selection criterion in the case of linear regression models. Consider the linear regression model

$$Y = \alpha + \beta^t X + e \tag{35}$$

where $e \sim \mathcal{N}(0, \sigma)$ and e is independent of X. Suppose we have a sample given by the i.i.d. random vectors $Z_i = (X_i, Y_i), i = 1, ..., n$, such that $Y_i = \alpha + \beta^t X_i + e_i$.

We consider the joint distribution of the entire data and write a pseudodistance between the theoretical model and the true model corresponding to the data. Let P_{θ} , $\theta := (\alpha, \beta, \sigma)$, be the probability measure associated to the theoretical model given by the random vector Z = (X, Y) and Q the probability measure associated to the true model corresponding to the data. Denote by p_{θ} , respectively by q the corresponding densities. For $\gamma > 0$, the pseudodistance between P_{θ} and Q is defined by

$$R_{\gamma}(P_{\theta},Q) := \frac{1}{\gamma+1} \ln\left(\int p_{\theta}^{\gamma}(x,y)dP_{\theta}(x,y)\right) + \frac{1}{\gamma(\gamma+1)} \ln\left(\int q^{\gamma}(x,y)dQ(x,y)\right) - \frac{1}{\gamma} \ln\left(\int p_{\theta}^{\gamma}(x,y)dQ(x,y)\right).$$
(36)

Similar to [18], since the middle term above does not depend on P_{θ} , a minimum pseudodistance estimator of the parameter $\theta_0 = (\alpha_0, \beta_0, \sigma_0)$ is defined by

$$\widehat{\theta}_{n} = (\widehat{\alpha}_{n}, \widehat{\beta}_{n}, \widehat{\sigma}_{n}) = \arg\min_{\alpha, \beta, \sigma} \left\{ \frac{1}{\gamma + 1} \ln\left(\int p_{\theta}^{\gamma}(x, y) dP_{\theta}(x, y)\right) - \frac{1}{\gamma} \ln\left(\int p_{\theta}^{\gamma}(x, y) dP_{n}(x, y)\right) \right\}, \quad (37)$$

where P_n is the empirical measure associated with the sample. This estimator can be written as

$$\widehat{\theta}_{n} = (\widehat{\alpha}_{n}, \widehat{\beta}_{n}, \widehat{\sigma}_{n}) = \arg\min_{\alpha, \beta, \sigma} \left\{ \frac{1}{\gamma + 1} \ln\left(\int \phi_{\sigma}^{\gamma + 1}(e) de\right) - \frac{1}{\gamma} \ln\left(\frac{1}{n} \sum_{i=1}^{n} \phi_{\sigma}^{\gamma}(Y_{i} - \alpha - \beta^{t} X_{i})\right) \right\}, \quad (38)$$

where ϕ_{σ} is the density of the random variable $e \sim \mathcal{N}(0, \sigma)$. Then, the estimator $Q_{\hat{\theta}_{u}}$ can be written as

$$Q_{\widehat{\theta}_n} = \min_{\alpha,\beta,\sigma} \left\{ \frac{1}{\gamma+1} \ln \left(\frac{1}{(\sigma\sqrt{2\pi})^{\gamma}\sqrt{\gamma+1}} \right) - \frac{1}{\gamma} \ln \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{(\sigma\sqrt{2\pi})^{\gamma}} \cdot \exp \left(-\frac{\gamma}{2\sigma^2} (Y_i - \alpha - \beta^t X_i)^2 \right) \right) \right\}.$$
(39)

In order to construct an asymptotic unbiased estimator of the expected overall discrepancy in the case of the linear regression models, we evaluated the second and the third terms from (18).

For values of γ close to 0 (γ smaller than 0.3), we found the following approximation of the matrix $M_{\gamma}(\theta_0)$

$$M_{\gamma}(\theta_0) \simeq \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} V^{-1} \begin{pmatrix} I & 0\\ 0 & \frac{3\gamma^2 + 4\gamma + 2}{2\gamma + 1} \end{pmatrix}$$
(40)

where *V* is the asymptotic covariance matrix of $\hat{\theta}_n$ and *I* is the identity matrix. We refer to [15] for the asymptotic properties of the minimum pseudodistance estimators in the case of linear regression models. Since $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically multivariate normal distributed, using the χ^2 distribution, we obtain the approximation

$$E[(\widehat{\theta_n} - \theta_0)^t M_{\gamma}(\theta_0)(\widehat{\theta_n} - \theta_0)] \simeq \frac{1}{n} \cdot \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \left[(d-1) + \frac{3\gamma^2 + 4\gamma + 2}{2(\gamma+1)(2\gamma+1)} \right].$$
 (41)

Also, the third term in (18) is given by

$$\frac{1}{2\gamma n} \left[1 - \left(\frac{\gamma + 1}{\sqrt{2\gamma + 1}} \right)^d \right]. \tag{42}$$

Then, according to Proposition 3, an asymptotically unbiased estimator of the expected overall discrepancy is given by

$$Q_{\hat{\theta}_n} + \frac{1}{n} \cdot \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \left[(d-1) + \frac{3\gamma^2 + 4\gamma + 2}{2(\gamma+1)(2\gamma+1)} \right] + \frac{1}{2\gamma n} \left[1 - \left(\frac{\gamma+1}{\sqrt{2\gamma+1}}\right)^d \right], \quad (43)$$

where $Q_{\hat{\theta}_n}$ is given by (39). Note that, using the asymptotic properties of $\hat{\theta}_n$ and the central limit theorem, the last two terms in (18) of Proposition 3 are $o_P(\frac{1}{n})$.

When we compare different linear regression models, as in Section 4 below, we can ignore the terms depending only on *n* and γ in (43). Therefore, we can use as model selection criterion the simplified expression

$$Q_{\widehat{\theta}_n} + \frac{(\gamma+1)^2}{(\sqrt{2\gamma+1})^3} \cdot \frac{d}{n} - \frac{1}{2\gamma n} \left(\frac{\gamma+1}{\sqrt{2\gamma+1}}\right)^d,\tag{44}$$

which we call Pseudodistance based Information Criterion (PIC).

4. Applications

4.1. Simulation Study

In order to illustrate the performance of the PIC criterion (44) in the case of linear regression models, we performed a simulation study using for comparison the model selection criteria AIC, BIC and MDIC. These criteria are defined respectively by

$$AIC = n \log \hat{\sigma}_p^2 + 2 (p+2)$$
$$BIC = n \log \hat{\sigma}_p^2 + (p+2) \log n,$$

where *n* the sample size, *p* the number of covariates of the model and $\hat{\sigma}_p^2$ the classical unbiased estimator of the variance of the model,

$$MDIC = nMQ_{\hat{\theta}} + (2\pi)^{-\alpha/2} (1+\alpha)^{2+p/2} p$$

with $\alpha = 0.25$ and

$$MQ_{\hat{\theta}} = -\left[(1+\alpha^{-1})\frac{1}{n}\sum_{n=1}^{n}f_{\hat{\theta}}^{\alpha}(X_{i})\right],$$

where $\hat{\theta}$ is a consistent estimate of the vector of unknown parameters involved in the model with p covariates and $f_{\hat{\theta}}$ is the associated probability density function. Note that MDIC is based on the well known BHHJ family of divergence measures indexed by a parameter $\alpha > 0$ and on the minimum divergence estimating method for robust parameter estimation (see [20]). The value of $\alpha = 0.25$ was found in [9] to be an ideal one for a great variety of settings. The above three criteria have been chosen to be used in this comparative study with PIC not only due to their popularity, but also due to their special characteristics. Indeed, AIC is the classical representative of asymptotically efficient criteria, BIC is known to be consistent, while MDIC is associated with robust estimations (see e.g., [20–23]).

Let X_1, X_2, X_3, X_4 be four variables following respectively the normal distributions $\mathcal{N}(0,3)$, $\mathcal{N}(1,3)$, $\mathcal{N}(2,3)$ and $\mathcal{N}(3,3)$. We consider the model

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \varepsilon$$

with $a_0 = a_1 = a_2 = 1$ and $\varepsilon \sim \mathcal{N}(0, 1)$. This is the uncontaminated model. In order to evaluate the robustness of the new PIC criterion, we also consider the contaminated model

$$Y = d_1(a_0 + a_1X_1 + a_2X_2 + \varepsilon) + d_2(a_0 + a_1X_1 + a_2X_2 + \varepsilon^*)$$

where $\varepsilon^* \sim \mathcal{N}(5,1)$ and $d_1, d_2 \in [0,1]$ such that $d_1 + d_2 = 1$. Note that for $d_1 = 1$ and $d_2 = 0$ the uncontaminated model is obtained.

The simulated data corresponding to the contaminated model are

$$Y_i = d_1(1 + X_{1,i} + X_{2,i} + \varepsilon_i) + d_2(1 + X_{1,i} + X_{2,i} + \varepsilon_i^*),$$

for i = 1, ..., n, where $X_{1,i}, X_{2,i}, \varepsilon_i, \varepsilon_i^*$ are values of the variables $X_1, X_2, \varepsilon, \varepsilon^*$ independently generated from the normal distributions $\mathcal{N}(1,3), \mathcal{N}(2,3), \mathcal{N}(0,1), \mathcal{N}(5,1)$ correspondingly.

With a set of four possible regressors there are $2^4 - 1 = 15$ possible model specifications that include at least one regressor. These 15 possible models constitute the set of candidate models in our study. More precisely, this set contains the full model (X_1 , X_2 , X_3 , X_4) given by

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + \varepsilon$$

as well as all 14 possible subsets of the full model consisting of one (X_{j_1}) , two (X_{j_1}, X_{j_2}) and three $(X_{j_1}, X_{j_2}, X_{j_3})$ of the four regressors X_1, X_2, X_3 and X_4 , with $j_1 \neq j_2 \neq j_3$, $j_i \in \{1, 2, 3, 4\}$ and i = 1, 2, 3.

In our simulation study, for several values of the parameter γ associated with the pseudodistance, we compared the new criterion PIC with the other model selection criteria. Different levels of contamination and different sample sizes have been considered. In the examples presented in this work, $d_1 \in \{0.8, 0.9, 0.95, 1\}$ and $n \in \{20, 50, 100\}$. Additional examples for n = 30, 75, 200, 500 have been analyzed (results not shown) with similar findings (see below). For each setting, fifty experiments were performed in order to select the best model among the available candidate models. In the framework of each of the fifty experiments, on the basis of the simulated observations, the value of each of the above model selection criteria was calculated for each of the 15 possible models. Then, for each criterion, the 15 candidate models were ranked from 1st to 15th according to the value of the criterion. The model chosen by a given criterion is the one for which the value of the criterion is the lowest among all the 15 candidate models.

Tables 1–12 present the proportions of models selected by the considered criteria. Among the 15 candidate models only 4 were chosen at least once. These four models are the same in all instances and appear in the 2nd column of all Tables.

For small sample sizes (n = 20, n = 30) the criteria PIC and MDIC yield the best results. When the level of contamination is 10% or 20%, the PIC criterion yields very good results and beats the other competitors almost all the time. When the level of contamination is small, for example 5% or when there is no contamination, the two criteria are comparable, in the sense that in many cases the proportions of selected models by the two criteria are very close, so that sometimes PIC wins and sometimes MDIC wins. Tables 1–4 present these results for n = 20, but similar results are obtained for n = 30, too.

For medium sample sizes (n = 50, n = 75), the criteria PIC and BIC yield the best results. The results for n = 50 are given in Tables 5–8. Note that the PIC criterion yields the best results for 0% and 10% contamination. For the other levels of contamination, there are values of γ for which PIC is the best among all the considered criteria. On the other hand, in most cases when BIC wins, the proportions of selections of the true model by BIC and PIC are close.

When the sample size is large (n = 100, n = 200, n = 500), BIC generally yields better results than PIC which stays relatively close behind, but sometimes BIC and PIC have the same performance. Tables 9–12 present the results obtained for n = 100.

Thus, the new PIC criterion works very well for small to medium sample sizes and for levels of contamination up to 20%, but falls behind BIC for large sample sizes. Note that for contaminated data, PIC with $\gamma = 0.15$ prevails in most of the considered cases. On the other hand, for uncontaminated data, it is PIC with $\gamma = 0.2$ that prevails in all the considered instances. It is also worth mentioning that PIC with $\gamma = 0.3$ appears to behave very satisfactorily in most cases irrespectively of the proportion of contamination (0%–20%) and the sample size. Observe also that in all cases, AIC has the highest overestimation rate which is somehow expected (see [24]).

Although the consistency is the main focus of the applications presented in this work, one should point out that if prediction is part of the objective of a regression analysis, then model selection carried out using criteria such as the ones used in this work, have desirable properties. In fact, the case of finite-dimensional normal regression models has been shown to be associated with satisfactory prediction errors for criteria such as AIC and BIC (see [25]). Furthermore, it should be pointed out that in many instances PIC has a behavior quite similar to the above criteria by choosing the same models. Also, according to the presented simulation results, the proportion of choosing the true model by PIC is always better than the proportion of choosing the true model by AIC (even in the case of non contaminated data) and sometimes it is better than the proportion of choosing the true model by BIC. These results imply a satisfactory prediction ability for the proposed PIC criterion.

In conclusion, the new PIC criterion is a good competitor of the well known model selection criteria AIC, BIC and MDIC and may have superior performance especially in the case of small and contaminated samples.

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 90 | 84 | 88 | 84 | 92 | 90 | 86 |
| | X_1, X_2, X_3 | (10) | (16) | (12) | (16) |) (8) | (10) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) | J | J | J | J | J | J |
| AIC | X_1, X_2 | 62 | 56 | 52 | 56 | 66 | 56 | 60 |
| | X_1, X_2, X_3 | (38) | (44) | (48) | (44) | (34) | (44) | (40) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 74 | 76 | 60 | 74 | 72 | 68 | 70 |
| | X_1, X_2, X_3 | (26) | (24) | (40) | (26) | (28) | (32) |) (30) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 86 | 86 | 64 | 78 | 84 | 80 | 74 |
| | X_1, X_2, X_3 | (14) | (14) |) (36) | (22) |) (16) | (20) |) (26) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 1. Proportions of selected models by the considered criteria ($n = 20, d_1 = 0.8$).

Table 2. Proportions of selected models by the considered criteria ($n = 20, d_1 = 0.9$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 80 | 84 | 90 | 82 | 82 | 80 | 80 |
| | X_1, X_2, X_3 | (20) | (16) | (10) | (18) | (18) | (20) | (20) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| AIC | X_1, X_2 | 60 | 52 | 56 | 62 | 64 | 54 | 52 |
| | X_1, X_2, X_3 | (40) | (48) | (44) | (38) |) (36) | (46) | (48) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 76 | 70 | 78 | 72 | 84 | 76 | 76 |
| | X_1, X_2, X_3 | (24) |) (30) |) (22) | (28) | (16) | (24) |) (24) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 86 | 76 | 88 | 74 | 92 | 78 | 86 |
| | X_1, X_2, X_3 | (14) | (24) | (12) | (26) |) (8) | (22) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 82 | 88 | 80 | 94 | 82 | 88 | 86 |
| | X_1, X_2, X_3 | (18) | (12) | (20) |) (6) | (18) | (12) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J |) | J |
| AIC | X_1, X_2 | 78 | 50 | 66 | 70 | 66 | 64 | 66 |
| | X_1, X_2, X_3 | (22) | (50) | (34) | (30) | (34) | (36) |) (34) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J |) | J | J | J |
| BIC | X_1, X_2 | 84 | 64 | 74 | 84 | 84 | 76 | 82 |
| | X_1, X_2, X_3 | (16) | (36) | (26) | (16) | (16) | (24) | (18) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 90 | 74 | 82 | 88 | 88 | 80 | 88 |
| | X_1, X_2, X_3 | (10) | (26) | (18) | (12) | (12) | (20) |) (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 3. Proportions of selected models by the considered criteria (n = 20, $d_1 = 0.95$).

Table 4. Proportions of selected models by the considered criteria ($n = 20, d_1 = 1$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 86 | 86 | 86 | 86 | 88 | 82 | 92 |
| | X_1, X_2, X_3 | (14) | (14) | (14) | (14) | (12) | (18) |) (8) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) |) | J |) | J |) |) |
| AIC | X_1, X_2 | 64 | 74 | 62 | 58 | 64 | 62 | 70 |
| | X_1, X_2, X_3 | (36) | (26) |) (38) | (42) | (36) | (38) |) (30) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J |) | J |) | J |) |) |
| BIC | X_1, X_2 | 78 | 90 | 78 | 80 | 82 | 80 | 74 |
| | X_1, X_2, X_3 | (22) |) (10) |) (22) | (20) | (18) | (20) |) (26) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 84 | 92 | 88 | 88 | 88 | 88 | 80 |
| | X_1, X_2, X_3 | (16) |) (8) | (12) | (12) | (12) | (12) | (20) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|--|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 86 | 96 | 94 | 90 | 88 | 86 | 90 |
| | X_1, X_2, X_3 | (14) | (4) |) (6) | (10) | (12) | (14) | (10) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | $\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4$ |) |) |) |) |) |) |) |
| AIC | X_1, X_2 | 74 | 64 | 82 | 62 | 64 | 78 | 72 |
| | X_1, X_2, X_3 | (26) | (36) | (18) | (38) | (36) | (22) | (28) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 94 | 86 | 96 | 86 | 90 | 88 | 90 |
| | X_1, X_2, X_3 |) (6) | (14) | (4) | (14) | (10) | (12) |) (10) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 94 | 82 | 98 | 82 | 86 | 88 | 90 |
| | X_1, X_2, X_3 |) (6) | (18) |) (2) | (18) | (14) | (12) |) (10) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 5. Proportions of selected models by the considered criteria ($n = 50, d_1 = 0.8$).

Table 6. Proportions of selected models by the considered criteria (n = 50, $d_1 = 0.9$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 92 | 88 | 92 | 90 | 82 | 94 | 86 |
| | X_1, X_2, X_3 |) (8) | (12) |) (8) | (10) | (18) |) (6) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) |) |) |) |) |) |) |
| AIC | X_1, X_2 | 70 | 64 | 62 | 64 | 66 | 74 | 72 |
| | X_1, X_2, X_3 | (30) |) (36) |) (38) | (36) | (34) | (26) | (28) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 92 | 88 | 82 | 92 | 88 | 88 | 86 |
| | X_1, X_2, X_3 |) (8) | (12) | (18) |) (8) | (12) | (12) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 92 | 86 | 76 | 88 | 84 | 88 | 86 |
| | X_1, X_2, X_3 |) (8) | (14) | (24) | (12) | (16) | (12) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | <u> </u> |

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 94 | 92 | 92 | 88 | 84 | 90 | 88 |
| | X_1, X_2, X_3 |) (6) |) (8) | (8) | (12) | (16) | (10) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J |) | J | J | J | J | J |
| AIC | X_1, X_2 | 70 | 62 | 66 | 68 | 70 | 72 | 58 |
| | X_1, X_2, X_3 | (30) | (38) | (34) | (32) | (30) | (28) | (42) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 96 | 82 | 92 | 86 | 92 | 92 | 86 |
| | X_1, X_2, X_3 | (4) | (18) |) (8) | (14) |) (8) |) (8) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_{1}, X_{2} | 90 | 78 | 88 | 86 | 86 | 90 | 82 |
| | X_1, X_2, X_3 | (10) | (22) | (12) | (14) | (14) | (10) | (18) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 7. Proportions of selected models by the considered criteria (n = 50, $d_1 = 0.95$).

Table 8. Proportions of selected models by the considered criteria ($n = 50, d_1 = 1$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 94 | 90 | 80 | 84 | 90 | 94 | 88 |
| | X_1, X_2, X_3 |) (6) | (10) | (20) | (16) | (10) |) (6) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J |) | J | J | J |
| AIC | X_1, X_2 | 64 | 68 | 62 | 68 | 66 | 64 | 62 |
| | X_1, X_2, X_3 | (34) | (32) |) (38) | (32) | (34) | (36) | (38) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 86 | 86 | 86 | 90 | 86 | 94 | 82 |
| | X_1, X_2, X_3 | (14) | (14) | (14) | (10) | (14) |) (6) | (18) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_{1}, X_{2} | 84 | 84 | 82 | 88 | 84 | 90 | 82 |
| | X_1, X_2, X_3 | (16) | (16) | (18) | (12) | (16) | (10) | (18) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 94 | 94 | 94 | 92 | 88 | 88 | 94 |
| | X_1, X_2, X_3 |) (6) |) (6) |) (6) |) (8) | (12) | (12) |) (6) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| AIC | X_1, X_2 | 70 | 82 | 78 | 70 | 68 | 68 | 72 |
| | X_1, X_2, X_3 | (30) | (18) |) (22) | (30) | (32) | (32) | (28) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 90 | 96 | 98 | 90 | 96 | 94 | 88 |
| | X_1, X_2, X_3 | (10) | (4) |) (2) | (10) |) (4) |) (6) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) | J | J |) | J |) | J |
| MDIC | X_1, X_2 | 86 | 96 | 92 | 86 | 92 | 90 | 88 |
| | X_1, X_2, X_3 | (14) | (4) |) (8) | (14) |) (8) | (10) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 9. Proportions of selected models by the considered criteria ($n = 100, d_1 = 0.8$).

Table 10. Proportions of selected models by the considered criteria ($n = 100, d_1 = 0.9$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma=0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|---------------|----------------|
| PIC | X_1, X_2 | 88 | 92 | 96 | 88 | 88 | 88 | 86 |
| | X_1, X_2, X_3 | (12) |) (8) | (4) | (12) | (12) | (12) | (14) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) | J | J |) | J |) |) |
| AIC | X_1, X_2 | 68 | 72 | 78 | 66 | 70 | 78 | 60 |
| | X_1, X_2, X_3 | (32) | (28) |) (22) | (34) | (30) | (22) | (40) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J |) |
| BIC | X_1, X_2 | 98 | 98 | 96 | 88 | 92 | 94 | 92 |
| | X_1, X_2, X_3 | (2) | (2) | (4) | (12) |) (8) |) (6) |) (8) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| MDIC | X_1, X_2 | 90 | 90 | 96 | 84 | 82 | 90 | 82 |
| | X_1, X_2, X_3 | (10) | (10) | (4) | (16) | (18) | (10) | (18) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 90 | 88 | 92 | 90 | 98 | 96 | 92 |
| | X_1, X_2, X_3 | (10) | (12) |) (8) | (10) |) (2) | (4) |) (8) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) |) | J |) |) |) | J |
| AIC | X_1, X_2 | 70 | 78 | 78 | 66 | 82 | 68 | 68 |
| | X_1, X_2, X_3 | (30) | (22) |) (22) | (34) | (18) | (32) | (32) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |
| BIC | X_1, X_2 | 96 | 92 | 92 | 94 | 96 | 94 | 88 |
| | X_1, X_2, X_3 | (4) |) (8) |) (8) |) (6) | (4) |) (6) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J |) | J |) | J |
| MDIC | X_1, X_2 | 90 | 88 | 82 | 90 | 94 | 84 | 88 |
| | X_1, X_2, X_3 | (10) | (12) | (18) | (10) |) (6) | (16) | (12) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

Table 11. Proportions of selected models by the considered criteria ($n = 100, d_1 = 0.95$).

Table 12. Proportions of the selected models by the considered criteria ($n = 100, d_1 = 1$).

| Criteria | Variables | $\gamma = 0.01$ | $\gamma = 0.05$ | $\gamma = 0.1$ | $\gamma = 0.15$ | $\gamma = 0.2$ | $\gamma = 0.25$ | $\gamma = 0.3$ |
|----------|----------------------|-----------------|-----------------|----------------|-----------------|----------------|-----------------|----------------|
| PIC | X_1, X_2 | 94 | 96 | 92 | 92 | 96 | 90 | 94 |
| | X_1, X_2, X_3 |) (6) | (4) |) (8) |) (8) | (4) | (10) |) (6) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J |) |) | J |) |) | J |
| AIC | X_1, X_2 | 78 | 74 | 72 | 74 | 70 | 62 | 74 |
| | X_1, X_2, X_3 | (22) | (26) | (28) | (26) | (30) | (38) | (26) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 |) |) | J |) | J |) | J |
| BIC | X_1, X_2 | 96 | 100 | 92 | 96 | 94 | 90 | 100 |
| | X_1, X_2, X_3 | (4) | |) (8) | (4) |) (6) | (10) | |
| | X_1, X_2, X_4 | } | | } | } | } | } | |
| | X_1, X_2, X_3, X_4 | J | | J | J | J | J | |
| MDIC | X_{1}, X_{2} | 94 | 92 | 86 | 90 | 86 | 80 | 94 |
| | X_1, X_2, X_3 |) (6) |) (8) | (14) | (10) | (14) | (20) |) (6) |
| | X_1, X_2, X_4 | } | } | } | } | } | } | } |
| | X_1, X_2, X_3, X_4 | J | J | J | J | J | J | J |

4.2. Real Data Example

In order to illustrate the proposed method, we used the Hald cement data (see [26]) which represent a popular example for multiple linear regression. This example concern the heat evolved in calories per gram of cement Y as a function of the amount of each of four ingredient in the mix: tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium alumino-ferrite (X_3) and dicalcium silicate (X_4). The data are presented in Table 13.

| X_1 | X_2 | X_3 | X_4 | Y |
|-------|-------|-------|-------|-------|
| 7 | 26 | 6 | 60 | 78.5 |
| 1 | 29 | 15 | 52 | 74.3 |
| 11 | 56 | 8 | 20 | 104.3 |
| 11 | 31 | 8 | 47 | 87.6 |
| 7 | 52 | 6 | 33 | 95.9 |
| 11 | 55 | 9 | 22 | 109.2 |
| 3 | 71 | 17 | 6 | 102.7 |
| 1 | 31 | 22 | 44 | 72.5 |
| 2 | 54 | 18 | 22 | 93.1 |
| 21 | 47 | 4 | 26 | 115.9 |
| 1 | 40 | 23 | 34 | 83.8 |
| 11 | 66 | 9 | 12 | 113.3 |
| 10 | 68 | 8 | 12 | 109.4 |

Table 13. Hald cement data.

Since 4 variables are available, there are 15 possible candidate models (involving at least one regressor) for this data set. Note that the 4 single-variable models should be excluded from the analysis, because cement involves a mixture of at least two components that react chemically (see [27], p. 102). The model selection criteria that have been used are PIC for several values of γ , AIC, BIC and MDIC with $\alpha = 0.25$. Table 14 shows the model selected by each of the considered criteria.

Table 14. Selected models by model selection criteria.

| Criteria | Variables |
|----------------------|-----------------|
| PIC, $\gamma = 0.05$ | X_1, X_2, X_4 |
| PIC, $\gamma = 0.15$ | X_1, X_2, X_4 |
| PIC, $\gamma = 0.2$ | X_1, X_2, X_3 |
| PIC, $\gamma = 0.25$ | X_1, X_2, X_4 |
| PIC, $\gamma = 0.3$ | X_1, X_2, X_4 |
| AIC | X_1, X_2, X_4 |
| BIC | X_1, X_2 |
| MDIC | X_1,X_2,X_3 |

Observe that, in this example, PIC behaves similarly to AIC and MDIC having a slight tendency of overestimation. Note that for this specific dataset the collinearity is quite strong with X_1 and X_3 as well as X_2 and X_4 being seriously correlated. It should be pointed out that the model (X_1, X_2, X_4) is chosen not only by AIC and PIC, but also by C_p Mallows' criterion ([1]) with (X_1, X_2, X_3) coming very close second. Note further that (X_1, X_2, X_4) has also been chosen by cross validation ([28], p. 33) and PRESS ([26], p. 325). Finally, it is worth noticing that these two models share the highest adjusted R^2 values which are almost identical (0.976 for (X_1, X_2, X_4) and 0.974 for (X_1, X_2, X_3)) making the distinction between them extremely hard. Thus, in this example, the new PIC criterion gives results as good as other recognized classical model selection criteria.

5. Conclusions

In this work, by applying the same methodology as for AIC to a family of pseudodistances, we constructed new model selection criteria using minimum pseudodistance estimators. We proved theoretical properties of these criteria including asymptotic unbiasedness, robustness, consistency, as well as the limit laws. The case of the linear regression models was studied in detail and specific selection criteria based on pseudodistance are proposed.

For linear regression models, a comparative study based on Monte Carlo simulations illustrate the performance of the new methodology. Thus, for small sample sizes, the criteria PIC and MDIC yield the best results and in many cases PIC wins, for example when the level of contamination is 10%

or 20%. For medium sample sizes, the criteria PIC and BIC yield the best results. When the sample size is large, BIC generally yields better results than PIC which stays relatively close behind, but sometimes BIC and PIC have the same performance.

Based on the results of the simulation study and on the real data example, we conclude that the new PIC criterion is a good competitor of the well known criteria AIC, BIC and MDIC with an overall performance which is very satisfactory for all possible settings according to the sample size and contamination rate. Also PIC may have superior performance, especially in the case of small and contaminated samples.

An important issue that needs further investigation is the choice of the appropriate value for the parameter γ associated to the procedure. The findings of the presented simulation study show that, for contaminated data, the value $\gamma = 0.15$ leads to very good results, irrespectively of the sample size. Also, $\gamma = 0.3$ produces overall very satisfactory results, irrespectively of the sample size and the contamination rate. We hope to explore further and provide a clear solution to this problem, in a future work. We also intend to extend this methodology to other type of models including nonlinear or time series models.

Author Contributions: A.T. conceived the methodology, obtained the theoretical results. A.T., A.K. and P.T. conceived the application part. A.K. and P.T. implemented the method in R and obtained the numerical results. All authors wrote the paper. All authors have read and approved the final manuscript.

Funding: This research received no external funding.

Acknowledgments: The work of the first author was partially supported by a grant of the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, project number PN-II-RU-TE-2012-3-0007. The work of the third author was completed as part of the activities of the Laboratory of Statistics and Data Analysis of the University of the Aegean.

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

Abbreviations

The following abbreviations are used in this manuscript:

| AIC | Akaike Information Criterion |
|-------------------------|--|
| BIC | Bayesian Information Criterion |
| GIC | General Information Criterion |
| DIC | Divergence Information Criterion |
| MDIC | Modified Divergence Information Criterion |
| PIC | Pseudodistance based Information Criterion |
| BHHJ family of measures | Basu, Harris, Hjort and Jones family of measures |
| | |

Appendix A

Proof of Proposition 1. Using a Taylor expansion of W_{θ} around the true parameter θ_0 and taking $\theta = \hat{\theta}_n$, on the basis of (12) and (13) we obtain

$$W_{\widehat{\theta}_n} = W_{\theta_0} + \frac{1}{2} (\widehat{\theta}_n - \theta_0)^t M_{\gamma}(\theta_0) (\widehat{\theta}_n - \theta_0) + o(\|\widehat{\theta}_n - \theta_0\|^2).$$
(A1)

Then (15) holds. \Box

Proof of Proposition 2. Using a Taylor expansion of Q_{θ} around to $\hat{\theta}_n$ and taking $\theta = \theta_0$, we obtain

$$Q_{\theta_0} = Q_{\widehat{\theta}_n} + \left[\frac{\partial}{\partial \theta}Q_{\theta}\right]_{\theta=\widehat{\theta}_n}^t (\theta_0 - \widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^t \left[\frac{\partial^2}{\partial \theta^2}Q_{\theta}\right]_{\theta=\widehat{\theta}_n} (\theta_0 - \widehat{\theta}_n) + o(\|\widehat{\theta}_n - \theta_0\|^2).$$
(A2)

Note that $\left[\frac{\partial}{\partial \theta}Q_{\theta}\right]_{\theta=\widehat{\theta}_n} = 0$ by the very definition of $\widehat{\theta}_n$.

By applying the weak law of large numbers and the continuous mapping theorem, we get

$$\left[\frac{\partial^2}{\partial\theta^2}Q_\theta\right]_{\theta=\theta_0} - \left[\frac{\partial^2}{\partial\theta^2}W_\theta\right]_{\theta=\theta_0} \xrightarrow{P} 0 \tag{A3}$$

and using (13)

$$\left[\frac{\partial^2}{\partial\theta^2}Q_\theta\right]_{\theta=\theta_0} - M_\gamma(\theta_0) \xrightarrow{P} 0.$$
(A4)

Then, using the consistency of $\hat{\theta}_n$ and (A4), we obtain

$$\left[\frac{\partial^2}{\partial\theta^2}Q_\theta\right]_{\theta=\widehat{\theta}_n} = M_\gamma(\theta_0) + o_P(1).$$
(A5)

Consequently,

$$Q_{\theta_0} = Q_{\widehat{\theta}_n} + \frac{1}{2} (\theta_0 - \widehat{\theta}_n)^t M_{\gamma}(\theta_0) (\theta_0 - \widehat{\theta}_n) + o(\|\widehat{\theta}_n - \theta_0\|^2)$$
(A6)

and we deduce (17). \Box

Proof of Proposition 3. Using Proposition 1 and Proposition 2, we obtain

$$E[W_{\widehat{\theta}_n}] = E[Q_{\widehat{\theta}_n}] + E[(\theta_0 - \widehat{\theta}_n)^t M_{\gamma}(\theta_0)(\theta_0 - \widehat{\theta}_n)] - E[Q_{\theta_0}] + W_{\theta_0} + E[R_n]$$
(A7)

where $R_n = o(\|\widehat{\theta}_n - \theta_0\|^2)$. In order to evaluate $W_{\theta_0} - E[Q_{\theta_0}]$, note that

$$Q_{\theta_0} - W_{\theta_0} = -\frac{1}{\gamma} \left[\ln \left(\frac{1}{n} \sum_{i=1}^n p_{\theta_0}^{\gamma}(X_i) \right) - \ln \left(\int p_{\theta_0}^{\gamma+1} d\lambda \right) \right].$$
(A8)

A Taylor expansion of the function $\ln x$ around to $\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda$ yields

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i})\right) = \ln\left(\int p_{\theta_{0}}^{\gamma+1}d\lambda\right) + \frac{1}{\int p_{\theta_{0}}^{\gamma+1}d\lambda}\left[\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i}) - \int p_{\theta_{0}}^{\gamma+1}d\lambda\right] - \frac{1}{2}\cdot\frac{1}{(\int p_{\theta_{0}}^{\gamma+1}d\lambda)^{2}}\left[\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i}) - \int p_{\theta_{0}}^{\gamma+1}d\lambda\right]^{2} + o(\|\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i}) - \int p_{\theta_{0}}^{\gamma+1}d\lambda\|^{2}).$$
(A9)

Then

$$\begin{split} \mathbf{E}[Q_{\theta_0} - W_{\theta_0}] &= -\frac{1}{\gamma} \mathbf{E}\left[\ln\left(\frac{1}{n}\sum_{i=1}^n p_{\theta_0}^{\gamma}(X_i)\right) - \ln\left(\int p_{\theta_0}^{\gamma+1} d\lambda\right)\right] \\ &= -\frac{1}{\gamma} \left\{\frac{1}{\int p_{\theta_0}^{\gamma+1} d\lambda} \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n p_{\theta_0}^{\gamma}(X_i) - \int p_{\theta_0}^{\gamma+1} d\lambda\right] - \frac{1}{2} \cdot \frac{1}{(\int p_{\theta_0}^{\gamma+1} d\lambda)^2} \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n p_{\theta_0}^{\gamma}(X_i) - \int p_{\theta_0}^{\gamma+1} d\lambda\right)^2\right] + \mathbf{E}[R'_n]\right\} \end{split}$$

where $R'_n = o(\|\frac{1}{n}\sum_{i=1}^n p^{\gamma}_{\theta_0}(X_i) - \int p^{\gamma+1}_{\theta_0} d\lambda\|^2)$. On the other hand,

$$E\left[\left(\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i})-\int p_{\theta_{0}}^{\gamma+1}d\lambda\right)^{2}\right] = \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n}p_{\theta_{0}}^{\gamma}(X_{i})\right] = \frac{1}{n}\operatorname{Var}\left[p_{\theta_{0}}^{\gamma}(X)\right] \\
= \frac{1}{n}\left\{E[p_{\theta_{0}}^{2\gamma}(X)]-E[p_{\theta_{0}}^{\gamma}(X)]^{2}\right\} \\
= \frac{\int p_{\theta_{0}}^{2\gamma+1}d\lambda-(\int p_{\theta_{0}}^{\gamma+1}d\lambda)^{2}}{n}.$$
(A10)

Consequently,

$$\mathbf{E}[Q_{\theta_0}] - W_{\theta_0} = -\frac{1}{2\gamma n} \left[1 - \frac{\int p_{\theta_0}^{2\gamma+1} d\lambda}{\left(\int p_{\theta_0}^{\gamma+1} d\lambda\right)^2} \right] - \frac{1}{\gamma} E\left[R'_n\right].$$
(A11)

Using (A7) and (A11), we obtain (18). \Box

Proof of Proposition 5. For the contaminated model $\tilde{P}_{\varepsilon x} = (1 - \varepsilon)P_{\theta_0} + \varepsilon \delta_x$, it holds

$$U(\widetilde{P}_{\varepsilon x}) = \frac{1}{\gamma + 1} \ln\left(\int p_{T(\widetilde{P}_{\varepsilon x})}^{\gamma + 1} d\lambda\right) - \frac{1}{\gamma} \ln\left(\int p_{T(\widetilde{P}_{\varepsilon x})}^{\gamma} d\widetilde{P}_{\varepsilon x}\right).$$
(A12)

Derivation with respect to ε yields

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} [U(\widetilde{P}_{\varepsilon x})]_{\varepsilon=0} &= \frac{1}{\int p_{\theta_0}^{\gamma+1} d\lambda} \cdot \int p_{\theta_0}^{\gamma} \dot{p}_{\theta_0} d\lambda \cdot \mathrm{IF}(x; T, P_{\theta_0}) - \\ &- \frac{1}{\gamma} \cdot \frac{1}{\int p_{\theta_0}^{\gamma+1} d\lambda} \left\{ - \int p_{\theta_0}^{\gamma+1} d\lambda + \gamma \cdot \int p_{\theta_0}^{\gamma} \dot{p}_{\theta_0} d\lambda \cdot \mathrm{IF}(x; T, P_{\theta_0}) + p_{\theta_0}^{\gamma}(x) \right\} \\ &= \frac{1}{\gamma} \cdot \left[1 - \frac{p_{\theta_0}^{\gamma}(x)}{\int p_{\theta_0}^{\gamma+1} d\lambda} \right]. \end{aligned}$$

Thus we obtain

$$\mathrm{IF}(x; U, P_{\theta_0}) = \frac{1}{\gamma} \left[1 - \frac{p_{\theta_0}^{\gamma}(x)}{\int p_{\theta_0}^{\gamma+1} \mathrm{d}\lambda} \right].$$
(A13)

References

- 1. Mallows, C.L. Some comments on Cp. Technometrics 1973, 15, 661–675.
- Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the* Second International Symposium on Information Theory Petrov; Springer: Berlin/Heidelberger, Germany, 1973; pp. 267–281.
- 3. Schwarz, G. Estimating the dimension of a model. Ann. Stat. 1978, 6, 461–464. [CrossRef]
- 4. Konishi, S.; Kitagawa, G. Generalised information criteria in model selection. *Biometrika* **1996**, *83*, 875–890. [CrossRef]
- 5. Ronchetti, E. Robust model selection in regression. Statist. Probab. Lett. 1985, 3, 21–23. [CrossRef]
- 6. Ronchetti, E.; Staudte, R.G. A robust version of Mallows' Cp. J. Am. Stat. Assoc. 1994, 89, 550–559.
- Agostinelli, C. Robust model selection in regression via weighted likelihood estimating equations. *Stat. Probab. Lett.* 2002, 76, 1930–1934. [CrossRef]
- 8. Mattheou, K.; Lee, S.; Karagrigoriou, A. A model selection criterion based on the BHHJ measure of divergence. *J. Stat. Plann. Inf.* **2009**, *139*, 228–235. [CrossRef]

- 9. Mantalos, P.; Mattheou, K.; Karagrigoriou, A. An improved divergence information criterion for the determination of the order of an AR process. *Commun. Stat.-Simul. Comput.* **2010**, *39*, 865–879. [CrossRef]
- 10. Toma, A. Model selection criteria using divergences. Entropy 2014, 16, 2686–2698. [CrossRef]
- 11. Pardo, L. Statistical Inference Based on Divergence Measures; Chapmann & Hall: London, UK, 2006.
- 12. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapmann & Hall: London, UK, 2011.
- 13. Jones, M.C.; Hjort, N.L.; Harris, I.R.; Basu, A. A comparison of related density-based minimum divergence estimators. *Biometrika* 2001, *88*, 865–873. [CrossRef]
- 14. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
- 15. Broniatowski, M.; Toma, A.; Vajda, I. Decomposable pseudodistances and applications in statistical estimation. *J. Stat. Plan. Infer.* **2012**, *142*, 2574–2585. [CrossRef]
- 16. Toma, A.; Leoni-Aubin, S. Optimal robust M-estimators using Renyi pseudodistances. J. Multivar. Anal. 2013, 115, 359–373. [CrossRef]
- 17. Toma, A.; Leoni-Aubin, S. Robust portfolio optimization using pseudodistances. *PLoS ONE* **2015**, *10*, 1–26. [CrossRef]
- Toma, A.; Fulga, C. Robust estimation for the single index model using pseudodistances. *Entropy* 2018, 20, 374. [CrossRef]
- 19. Hampel, F.R.; Ronchetti, E.; Rousseeuw, P.J.; Stahel, W. *Robust Statistics: The Approach Based on Influence Functions*; Wiley Blackwell: Hoboken, NJ, USA, 1986.
- Basu, A.; Harris, I.R.; Hjort, N.L.; Jones, M.C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* 1998, *85*, 549–559. [CrossRef]
- 21. Karagrigoriou, A. Asymptotic efficiecy of the order selection of a nongaussian AR process. *Stat. Sin.* **1997**, *7*, 407–423.
- 22. Vonta, F.; Karagrigoriou, A. Generalized measures of divergence in survival analysis and reliability. *J. Appl. Prob.* **2010**, *47*, 216–234. [CrossRef]
- 23. Karagrigoriou, A.; Mattheou, K.; Vonta, F. On asymptotic properties of AIC variants with applications. *Open J. Stat.* **2011**, *1*, 105–109. [CrossRef]
- 24. Shibata, R. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **1976**, *63*, 117–126. [CrossRef]
- 25. Speed, T.P.; Yu. B. Model selection and prediction: normal regression. *Ann. Inst. Stat. Math.* **1993**, 45, 35–54. [CrossRef]
- 26. Draper, N.R.; Smith, H. Applied Regression Analysis, 2nd ed.; Wiley Blackwell: Hoboken, NJ, USA, 1981.
- 27. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: Berlin/Heidelberger, Germany, 2002.
- 28. Hjorth, J.S.U. *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap;* Chapman and Hall: London, UK, 1994.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).