UNIVERSITY OF THE AEGEAN
DEPARTMENT OF STATISTICS AND
ACTUARIAL-FINANCIAL MATHEMATICS
STATISTICS AND DATA ANALYSIS

MODEL INDENTIFICATION PROCEDURES: THEORETICAL ASPECTS AND PRACTICAL IMPLICATIONS

MASTER THESIS

PASCHALINI TRENTOU

SAMOS 2019

UNIVERSITY OF THE AEGEAN
DEPARTMENT OF STATISTICS AND
ACTUARIAL-FINANCIAL MATHEMATICS
STATISTICS AND DATA ANALYSIS

Model Indentification Procedures: Theoretical Aspects and Practical Implications

Master Thesis

Paschalini Trentou

Samos 2019

UNIVERSITY OF THE AEGEAN
DEPARTMENT OF STATISTICS AND
ACTUARIAL-FINANCIAL MATHEMATICS
STATISTICS AND DATA ANALYSIS

Model Indentification Procedures: Theoretical Aspects and Practical Implications

Master Thesis
Paschalini Trentou
20 June, Samos

EVALUATION COMMITTEE

Karagrigoriou Alexandros
Professor, University of the Aegean (Supervisor)

Rakitzis Athanasios
Assistant Professor, University of the Aegean

Toma Aida
Professor, Bucharest University of Economic Studies

Contents

Acknowledgements

With boundless love and appreciation, I would like to extend my heartfelt gratitude and appreciation to the people who helped me bring this master thesis into reality.

Firstly, I would like to express the deepest appreciation to my supervisor Professor Alexandros Karagrigoriou of the University of the Aegean, for his patient guidance, his unending support and consistent advices that helped me bring this thesis into success.

I would also like to thank the two readers of this thesis, Assistant Professor Athanasios Rakitzis of the University of the Aegean and especially Professor Aida Toma of Bucharest University of Economic Studies for her continuous interest, her valuable guidance and support throughout the period of working on this thesis.

Moreover, I would like to thank Emmanouil-Nektarios Kalligeris for his support and understanding. Thank you for always being there for me.

In addition, I would like to offer my special thanks to my friends and colleagues Kimon Ntotsis and Dr. Christina Parpoula.

Last but not least, I would like to thank from the bottom of my heart my family. My parents, Vasilis and Anna, my sister Maria and my brother Thanasis, for their love and support all those years.

Abstract

This thesis is conducted at the Department of Statistics and Actuarial-Financial Mathematics of the University of the Aegean, in the context of the MSc program in Statistics and Data Analysis. Its purpose is to present and analyze measures of divergence and information criteria but most importantly to check the performance of a novel information criterion, namely the Pseudodistance Information Criterion (PIC).

The structure of the thesis consists of three Chapters. In Chapter 1, measures of divergence and information criteria are presented. Some of the most well known divergence measures, namely Kullback-Leibler and the BHHJ measures are defined. Also, the most common model selection criteria like Akaike Information Criterion (AIC) and its variations, as well as the Bayesian Information Criterion (BIC), are discussed.

In Chapter 2, recent model selection like the Divergence Information Criterion (DIC) and the Modified Divergence Information Criterion (MDIC) are analyzed. Furthermore, a new model selection criterion, the Pseudodistance Information Criterion (PIC), is presented.

Finally, in Chapter 3, a simulation study and a real case study are performed in order to investigate the appropriateness and the performance of this new criterion.

Περίληψη

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοοικονομικών Μαθηματικών του Πανεπιστημίου Αιγαίου, στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών "Στατιστική και Ανάλυση Δεδομένων". Σκοπός της είναι η ανάλυση μέτρων απόστασης και κριτηρίων πληροφορίας αλλά κυρίως η διερεύνηση των επιδόσεων ενός νέου κριτηρίου πληροφορίας, του Pseudodistance Information Criterion (PIC).

Η δομή της διπλωματικής αποτελείται από τρία Κεφάλαια. Στο Κεφάλαιο 1, παρουσιάζονται μέτρα απόστασης και κριτήρια πληροφορίας. Ορίζονται κάποια από τα πιο γνωστά μέτρα απόστασης όπως το Kullback-Leibler αλλά και το BHHJ μέτρο. Επίσης, συζητιούνται τα πιο συνήθη κριτήρια επιλογής μοντέλου, όπως το Akaike Information Criterion (AIC) και οι παραλλαγές του, καθώς και το Bayesian Information Criterion (BIC).

Στο Κεφάλαιο 2, παρουσιάζονται πρόσφατα κριτήρια επιλογής μοντέλου όπως το Divergence Information Criterion (DIC) και το Modified Divergence Information Criterion (MDIC). Επιπλέον, παρουσιάζεται ένα νέο κριτήριο επιλογής μοντέλου, το Pseudodistance Information Criterion (PIC).

Τέλος, στο Κεφάλαιο 3, πραγματοποιείται μελέτη προσομοίωσης αλλά και μελέτη σε πραγματικά δεδομένα, με σκοπό να διερευνήσει την καταλληλότητα του νέου αυτού κριτηρίου.

Chapter 1

Measures of Divergence and Information Criteria

In this chapter, measures of divergence and information criteria will be presented. Measures of divergence are used as indices of similarity or dissimilarity between populations to measure mutual information about two variables and they can also be used to construct model selection criteria.

A model selection criterion can be considered as an approximately unbiased estimator of the expected overall discrepancy between a candidate model and the true model. If the value of the criterion is small, then the approximated candidate model can be chosen.

In 1951 Kullback-Leibler measure was developed to capture the information that is lost while approximating reality [18]. Some years later, in 1998, BHHJ measure of divergence was introduced [4]. Numerous other measures have been introduced in between. Based on Kullback-Leibler, several criteria have been developed. Other measures, like BHHJ, give rise to other criteria, the most common of which will be presented in Chapter 2. Akaike Information Criterion (AIC) and its variations as well as the Bayesian Information Criterion (BIC) will be discussed in the following sections, together with the associated

measures of divergence.

1.1 Measures of divergence

1.1.1 Kullback-Leibler Measure

Let $\mathbf{x} = (x_1, \ldots, x_n)$ a realization of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ so that the X_i's are independent and identically distributed random variables each with true unknown density function $g(\cdot, \theta_0)$, with $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})'$ the true but unknown value of the p-dimensional parameter of the distribution. Assume a candidate model $f_\theta(\cdot)$ and let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of θ_0 in some hypothesized set $\Theta \in \mathbb{R}^p$, i.e.

$$l(\hat{\theta}; x) = \sum_{i=1}^{n} \log(f_{\hat{\theta}}(x_i)) = \max_{\theta \in \Theta} l(\theta; x)$$

so that $f_{\hat{\theta}}(\cdot)$ could be considered as an estimate of $g(\cdot, \theta_0)$ and $l(\cdot; x)$ the log-likelihood function.

The divergence between the candidate model and the true density can be measured by the Kullback-Leibler (K-L) measure (see [17]).

$$I_X^{KL}(g, f_{\hat{\theta}}) = \int g(y, \theta_0) \log\Big(\frac{g(y, \theta_0)}{f_{\hat{\theta}}(y)}\Big) dy$$

where θ_0 is the true value of the parameter of $g(\cdot)$ and $\hat{\theta}$ the estimate of the parameter.

Notice that $I_X^{KL}(g, f_{\hat{\theta}})$ can also be written in the following form:

$$I_X^{KL}(g, f_{\hat{\theta}}) = E_g[\log(g(X, \theta_0)] - E_g[\log(f_{\hat{\theta}}(X))]. \tag{1.1}$$

where E_g represents the expectation with respect to the true probability distribution g.

Observe that the first term is independent of the candidate model and therefore the divergence can be evaluated using only the second term, which is known as the expected log-likelihood.

In the case of discrete models, K-L can be measured by:

$$I_X^{KL}(g, f_{\hat{\theta}}) = \sum_{i=1}^{\infty} g(y_i, \theta_0) \log\Big(\frac{g(y_i, \theta_0)}{f_{\hat{\theta}}(y_i)}\Big). \tag{1.2}$$

1.1.2 BHHJ Measure

Kullback-Leibler measure for $a \to 0$ is a special case of the well known BHHJ measure intoduced by Basu et *al.* (1998). The BHHJ measure is given by (see [4])

$$I_X^a\,(g, f_{\hat{\theta}}) = \int \left\{ f_{\hat{\theta}}^{1+a}\,(y) - \left(1 + \frac{1}{a}\right) g\,(y)\, f_{\hat{\theta}}^{a}\,(y) + \frac{1}{a} g^{1+a}\,(y) \right\} dy$$

where α is an index usually taken in $(0, 1)$.

1.1.3 Other measures of divergence

Other well known divergence measures between two functions g and f are:

$$\chi^2(g; f) = \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i}, \quad \chi^2 \text{ -statistics [24]}$$

$$I_K(g; f) = \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx, \quad \text{Hellinger distance ([14], [23])}$$

$$D_\varphi(g; f) = \varphi(g) - \varphi(f) - (g - f)^T \bigtriangledown \varphi(f), \quad \text{Bregman divergences [7]}$$

where T denotes the transpose.

1.2 Information Criteria

Choosing a model with too few parameters can involve making non realistic assumptions, bias and poor prediction. Such models are not capable of describing the sample and therefore the entire population [13]. Another common problem in statistical modeling is having too many parameters in the model which increases complexity.

By using different information criteria researchers are looking for a model that adapts to the data while the true model is unknown.

According to the Principle of Parsimony, the model with the lowest number of parameters is selected, which is still capable to describe better the data. Also a simple model is better than a complex one, not only for practical purposes but also for descriptive as well as predictive purposes.

The general form of an information criterion (IC) is the following [13]:

$$IC = -2l(\hat{\theta}) + A_n p$$

where l is the log-likelihood, A_n is a quantity that may or may not depend on the sample size n also known as penalty weight and p is the number of parameters of the model.

The value of A_n changes accordingly to information criterion that has been used. In the table below some typical values of A_n are presented, for different information criteria.

CRITERION	**PENALTY WEIGHT**
AIC [2]	$A_n = 2$
CAIC [6]	$A_n = \log(n) + 1$
BIC [26]	$A_n = \log(n)$
Adjusted BIC [27]	$A_n = \log(\frac{n+2}{24})$

1.2.1 Akaike Information Criterion

The Akaike Information Criterion is one of the most popular criteria. The Kullback-Leibler measure was the one used by Akaike [2] to develop AIC. Akaike proposed the evaluation of the fit of the candidate model using minus twice the *mean* expected log-likelihood given by

$$-2E_g\big[E_g[\log(f_{\hat{\theta}}(X))]\big] = -2\int\ldots\int E_g[\log(f_{\hat{\theta}}(X))]\prod_{i=1}^{n} g(x_i,\theta_0)dx_1\ldots dx_n$$

since the candidate model is close to the true model if the above quantity is small. He also provided an unbiased estimator of the expected log-likelihood so that the resulting criterion is given by:

$$AIC(p) = -2l(\hat{\theta}) + 2p$$

where θ is the vector of model parameters, $L(\hat{\theta})$ is the likelihood of the candidate model and p is the number of parameters of the candidate model.

1.2.1.1 The proof of AIC

According to (1.1) we have

$$I(g,f) = \int g\log\frac{g}{f} = \int g\log g - \int g\log f$$

so that

$$\min I(g,f) = \max\int g\log f \equiv \mathbb{E}_g\log f$$

a natural estimator of which is:

$$\simeq \frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\theta) = \frac{1}{n}\log\prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{n}\text{loglikelihood}$$

Thus, the expected loglikelihood is analogous to likelihood:

$$\mathbb{E}\log f(x;\theta) \simeq \frac{1}{n}\log\prod_{i=1}^{n} f(x_i;\theta) \Rightarrow$$

$$n\mathbb{E}\log f(x;\theta) \simeq \log\prod_{i=1}^{n} f(x_i;\theta) \equiv l(\theta) \Rightarrow$$

$$l^*(\theta) = l(\theta)$$

Let $\hat{\theta}$ be the MLE. Then,

$$l^*(\hat{\theta}) = l(\hat{\theta})$$

$l(\hat{\theta})$ is a good estimate for $l^*(\cdot)$ provided that it is unbiased. For that it should satisfy the property:

$$\mathbb{E}(l^*(\hat{\theta_k})) \simeq \mathbb{E}(l(\hat{\theta_k}))$$

where $\mathbb{E}(l^*(\hat{\theta}))$ is the mean expected log-likelihood. We will show below that $l(\hat{\theta})$ is not an unbiased estimator of $\mathbb{E}(l^*(\hat{\theta_k}))$ but $l(\hat{\theta_k}) - k$ is, i.e.

$$\mathbb{E}(l(\hat{\theta_k}) - k) = \mathbb{E}(l^*(\hat{\theta_k}))$$

To show the above we will consider the simplest case.

Let $x_1, \ldots, x_n$ i.i.d., g be the true model which follows $\mathcal{N}(0,1)$ and f is the approximate or candidate model that follows $\mathcal{N}(\mu,1)$. The log-likelihood of g is given by:

$$l(\mu) = \log\prod_{i=1}^{n} f(x_i,\mu) = \sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}}\exp^{\frac{-(x_i-\mu)^2}{2}}\right)$$

$$= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Recall that $\bar{X}$ is the MLE of μ. Then the above becomes:

$$-\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2$$

$$= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n}{2}(\bar{x} - \mu)^2 \Rightarrow$$

$$l(\mu) = l(\hat{\mu}) - \frac{n}{2}(\bar{x} - \mu)^2.$$

Then, for $\mu = 0$:

$$l(0) = l(\hat{\mu}) - \frac{n}{2}(\bar{x} - 0)^2.$$

Taking expectations on both sides we have:

$$\mathbb{E}(l(0)) = \begin{cases} \mathbb{E}(l(\hat{\mu})) - \frac{n}{2}\cdot\frac{1}{n} = \mathbb{E}(l(\hat{\mu})) - \frac{1}{2} \\ -\frac{n}{2}\log 2\pi - \frac{1}{2}\cdot n \cdot 1 = -\frac{n}{2}\log 2\pi - \frac{n}{2} \end{cases} \tag{1.3}$$

On the other hand,

$$l^*(\mu) = n\mathbb{E}\log f(x;\mu)$$

$$= n\int \frac{1}{\sqrt{2\pi}}\exp^{-\frac{z^2}{2}}\log\left(\frac{1}{\sqrt{2\pi}}\exp^{\frac{-(z-\mu)^2}{2}}\right)dz$$

$$= \ldots = -\frac{n}{2}\log 2\pi - \frac{n}{2}(1+\mu)^2 \tag{1.4}$$

which for $\mu = 0$ and using (1.3) becomes

$$l^*(0) = -\frac{n}{2}\log 2\pi - \frac{n}{2} \equiv \mathbb{E}(l(0)) \equiv \mathbb{E}(l(\hat{\mu})) - \frac{1}{2}. \tag{1.5}$$

Taking now Taylor expansion of $l^*(\mu)$ at point $\mu = 0$ we have:

$$l^*(\mu) \approx l^*(0) + n(\mu - 0)\mathbb{E}\frac{d\log f}{d\mu}\bigg|_{\mu=0} + n(\mu - 0)^2\mathbb{E}\frac{d^2\log f}{d\mu^2}\bigg|_{\mu=0} \tag{1.6}$$

where the derivatives can be easily obtained as follows:

$$\mathbb{E}\frac{d}{d\mu}[\log f]\bigg|_{\mu=0} = \mathbb{E}\frac{d}{d\mu}\bigg[-\frac{1}{2}\log 2\pi - \frac{(z-\mu)^2}{2}\bigg]\bigg|_{\mu=0} = \mathbb{E}(z-\mu)\bigg|_{\mu=0} = 0$$

$$\mathbb{E}\frac{d^2}{d\mu^2}[\log f]\bigg|_{\mu=0} = \mathbb{E}\frac{d}{d\mu}\bigg[+(z-\mu)\bigg]\bigg|_{\mu=0} = -1$$

Thus (1.6) for $\mu = \hat{\mu}$ takes the form:

$$l^*(\hat{\mu}) = l^*(0) - \frac{n}{2}(\mu - 0)^2 = l^*(0) - \frac{n}{2}\mu^2 \Rightarrow$$

$$l^*(\hat{\mu}) = l^*(0) - \frac{n}{2}\hat{\mu}^2 \Rightarrow$$

$$\mathbb{E}(l^*(\hat{\mu})) = l^*(0) - \mathbb{E}\big[\frac{n}{2}\hat{\mu}^2\big]. \tag{1.7}$$

Due to the asymptotic normality of $\hat{\mu}$,

$$\sqrt{n}(\hat{\mu} - 0) \longrightarrow \mathcal{N}(0, 1)$$

we have that (1.7) becomes:

$$\mathbb{E}(l^*(\hat{\mu})) = l^*(0) - \frac{1}{2}.$$

By (1.6) we get:

$$\mathbb{E}(l^*(\hat{\mu})) = \mathbb{E}(l(\hat{\mu})) - \frac{1}{2} - \frac{1}{2} = \mathbb{E}(l(\hat{\mu})) - 1 = \mathbb{E}[(l(\hat{\mu})) - 1]$$

which shows that $l(\hat{\mu}) - 1$ is unbiased for $l^*(\hat{\mu})$.

The above can be easily generalized to

$$\mathbb{E}(l^*(\hat{\theta_k})) = \mathbb{E}[(l(\hat{\theta_k})) - 1]$$

where 1 represents the number of unknown parameters (the mean μ).

Since we need to maximize the expected log-likelihood:

$$\equiv \max_{\text{model}} \mathbb{E} \log f \equiv \max_{\text{model}}(l^*(\theta)) = \max(l^*(\hat{\theta_k}))$$

In order to make the result independent of the specific estimator and the associated sample, we use instead:

$$\max_{\text{model}} \mathbb{E}(l^*(\hat{\theta_k}))$$

which for a model with k parameters is equal to

$$= \max_{\text{model}}[l(\hat{\theta_k}) - k]$$

Thus the unbiasedness is established as long as the $l(\hat{\theta_k})$ is reduced by the number of parameters k. Then, the AIC is given by

$$AIC = -2[l(\hat{\theta_k}) - k] = -2l(\hat{\theta_k}) + 2k$$

with the best model being the one with k parameters that minimizes AIC.

Remark 1. *Based on the above discussion, we have that AIC is an unbiased estimator of the mean expected log-likelihood, namely:*

$$\mathbb{E}(AIC) = -2\mathbb{E}[l(\hat{\theta_k}) - k] = -2[l^*(\hat{\theta_k})] = -2[\textit{Mean Expected Loglikelihood}].$$

1.2.1.2 Akaike Information Criterion and Entropy

As mentioned previously, Akaike used Kullback-Leibler (K-L) to develop Akaike Information Criterion (AIC). K-L informantion quantity is the one called negative entropy and it is based on Boltzmann's entropy in [5]. Indeed,

$$\text{Boltzmann's entropy} = -\log\left(\frac{g(x)}{f(x)}\right) \Leftrightarrow$$

$$-\text{Boltzmann's entropy} = \log\left(\frac{g(x)}{f(x)}\right)$$

In the case where the model is continuous and has a probability density function, K-L information quantity of the true distribution with respect to the model, is given in terms of Boltzmann's entropy by:

$$K - L = \mathbb{E}_g(-\text{Boltzmann's entropy}) = \mathbb{E}_g\left(\log\left(\frac{g(x)}{f(x)}\right)\right)$$

$$= \int g(x)\log\left(\frac{g(x)}{f(x)}\right)dx$$

or for the discrete case:

$$= \sum g_i \log\frac{g_i}{f_i}$$

where $g(x)$ is the true density function and $f(x)$ is the density function that specifies

the model [25]. Consequently K-L's minimization is equal to entropy's maximization. Nevertheless, entropy's maximization can lead to a model with high uncertainty, while K-L's minimization is a better approach as it leads to an approximate model with a little loss of data information.

In the case that the distribution is discrete and n is the sample size, consider the situation where $f_i = \frac{n_i}{n}$ for some integers $n_1, \ldots, n_k$ such that $n_1 + n_2 + \ldots + n_k = n$ and $g = (g_1, \ldots, g_k)$ is our model. The probability that we obtain the frequency distribution $n_1, \ldots, n_k$ from n observations that follow this distribution is:

$$W = \frac{n!}{n_1! \ldots n_k!} g_1^{n_1}, \ldots, g_k^{n_k}$$

By taking logarithm and Stirling's approximation $\log n! \sim n \log n - n$ we have

$$\log W = \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{k} n_i \log g_i$$

$$\sim n \log n - n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i \log g_i$$

$$= -\sum_{i=1}^{k} n_i \log \left(\frac{n_i}{n} \right) + \sum_{i=1}^{k} n_i \log g_i$$

$$= \sum_{i=1}^{k} n_i \log \left(\frac{g_i}{f_i} \right)$$

$$= n \sum_{i=1}^{k} g_i \log \left(\frac{g}{f_i} \right)$$

$$= nI(g; f)$$

Thus it follows that $I(g; f) \sim n^{-1} \log W$.

1.2.1.3 Consistent Akaike Information Criterion

Bozdogan [6] proposed a corrected version of AIC, CAIC in order to overcome its tendency of overestimating the complexity of the underlying model. Akaike Information Criteria (AIC) does not directly depend on sample size and that is the reason for not possesing the consistency property [1]. In formulating CAIC, a correction factor based on the sample size is employed to compensate for the overestimating nature of AIC [3]. Below is presented the form of the Consistent Akaike Information Criterion (CAIC):

$$CAIC(p) = -2 \log L(\hat{\theta}) + ((\log n) + 1)p$$

where θ is the vector of model parameters, n is the sample size, $\log L(\hat{\theta})$ is the log-likelihood $l(\hat{\theta})$ of the candidate model and p is the number of parameters of the candidate model.

AIC differs from CAIC in the second term which now takes into account sample size n.

1.2.1.4 Corrected Akaike's Information Criterion

Hurvich and Tsai [15] found a small sample bias adjustment and this led to a "corrected" criterion (AICc) where the additional penalty is a function of the sample size n. AICc is given by:

$$AICc(p) = AIC + \frac{2(p+1)(p+2)}{n-p-2}$$

where p is the number of estimated parameters in the model and n is the number of observations used in the model. AICc provides a stronger penalty than AIC for smaller sample sizes, and stronger than BIC for very small sample sizes [8]. This value can be used to compare various models for the same data set to determine the best-fitting model. The model having the smallest value is usually the preferred model [3].

1.2.2 Bayesian Information Criterion

Bayesian Information Criterion (BIC) was introduced by Schwarz [26] and it is given by:

$$BIC(p) = -2 \log L(\hat{\theta}) + p \log n$$

where θ is the vector of model parameters, n is the sample size, $\log L(\hat{\theta})$ is the log-likelihood $l(\hat{\theta})$ of the candidate model and p is the number of parameters of the candidate model.

Observe that BIC is based partly on the likelihood function, and it is closely related to Akaike information criterion (AIC) but the penalty is harsher than AIC. Thus, BIC tends to choose simpler models. BIC assumes that one of the models is true and that we are trying to find the model most likely to be true in the Bayesian sense, on the other hand AIC is trying to find the model that predict the best.

The BIC criterion has a number of advantages worth to be mentioned. First of all, it has been shown to be consistent [26] which means that it chooses the correct model with probability 1 as n tends to infinity. Secondly BIC depends on $\log n$ instead of n and therefore it downweights the effect of sample size, which in some cases can prevent the erroneous rejection of the null hypothesis for large sample sizes. A discussion about the properties of AIC and BIC is presented in the following section.

1.2.3 A comparison between AIC and BIC

Classical ways of evaluation of information criteria are consistency (strong or weak) and efficiency. These two properties which are associated with BIC and AIC respectively are briefly presented below.

1.2.3.1 Consistency

There are two types of consistency, strong and weak. Assume that the data were generated from a model assumed to be the true model and belongs to the class of candidate models. By using model selection methods we wish to identify and select this model as the true model. This is related to the concept of consistency. If the selection method is capable of selecting the true model from a set of candidate models with probability tending to 1, then this method is called weakly consistent. On the other hand, we have strong consistency when the true model is selected almost surely [11].

If we don't want to make the assumption that the true model is among the set of the candidate models, we assume there is a candidate model that is closest in the sense of the Kullback-Leibler measure to the true model. Weak consistency is the property that such a closest model is selected by model selection criteria, with probability tending to one.

Both AIC and BIC are constructed using minus twice the mean expected log-likelihood plus a penalty term for the complexity of the model. BIC's penalty is bigger than AIC's. This shows that BIC is not choosing models with too many parameters.

As previously mentioned the general form of information criteria is the following:

$$IC = -2l(\hat{\theta}) + A_n p \equiv -2l(\hat{\theta}) + C_{n,k}$$

where $C_{n,k} > 0$ is the penalty term for candidate model M_k. Specifically,

$$C_{n,k} = \begin{cases} 2(dim(\theta)), \text{for AIC} \\ \log n(dim(\theta)), \text{for BIC} \end{cases} .$$

The two types of consistency are given in Theorem 1 & 2 and Theorem 3 & 4 respectively [11].

Theorem 1. *(Weak Consistency). Assume that there is one model M_{k_0} among the set of the candidate models for which the minimum Kullback-Leibler distance is reached. That is, for this model it holds that:*

1. $$\lim_{n\to\infty} infmin_{k\neq k_0} \frac{1}{n} \sum_{i=1}^{n} (KL(g; f_{k,i}) - KL(g; f_{k_o,i})) > 0$$

   and

2. *The penalty term is $C_{n,k} = O_p(n)$*

Then, the information criterion chooses model M_{k_0} as the best model with probability one.

So, using the above theorem for AIC and BIC it appears that:

$$AIC : \frac{C_{n,k}}{n} = \lim_{n\to\infty} \frac{2}{n} dim(\theta) = 0$$

and

$$BIC : \frac{C_{n,k}}{n} = \lim_{n\to\infty} \frac{\log n}{n} dim(\theta) = 0$$

Therefore, both AIC and BIC are weakly consistent.

Theorem 2. *(Strong Consistency). Assume that there is one model M_{k_0} among the set of the candidate models which reaches the minimum Kullback-Leibler distance. That is, for this model it holds that:*

1.

$$\lim_{n\to\infty} infmin_{k\neq k_0} \frac{1}{n}\sum_{i=1}^{n}(KL(g;f_{k,i}) - KL(g;f_{k_o,i})) > 0$$

and

2. Let penalty $C_{n,k} = O_p(n)$ almost surely.

Then,

$$P(min_{i\neq k_0}(IC(M_{k_o}) - IC(M_i)) > 0, for\ almost\ all\ n) = 1$$

Thus, AIC and BIC are strongly consistent selectors of the model, namely they are best in minimising the Kullback-Leibler distance to the true model.

Theorem 3. *(Weak Consistency). Denote by I the set of candidate models that all reach the minimum Kullback-Leibler distance to the true model, and I_0 is the subset of I containing model with the smallest dimension. Assume that:*

- *For all $k_0 \neq I_0 \subset I$:*

$$\lim_{n\to\infty} sup\frac{1}{n}\sum_{i=1}^{n}(KL(g;f_{k_0,i}) - KL(g;f_{I_o,i})) < \infty$$

- *For all $i \in I_0$ and for all $k \in I\backslash I_0$:*

$$P\left(\frac{C_{n,i} - C_{n,i_0}}{\sqrt{n}} \longrightarrow \infty\right) = 1$$

- *For all $k_0 \neq i_0 \in I$ and for the log-likelihood ratio:*

$$\sum_{i=1}^{n} \log\frac{f_{k_0,i}(y_i;\theta^*_{k_0})}{f_{I_0,i}(y_i;\theta^*_{I_0})} = O_p(1)$$

and that for any $i_0 \in I_0$ *and* $i \in I \backslash I_0$

$$P(C_{n,i} - C_{n,i_0}) = 1$$

Then, with probability $\longrightarrow$ *1 the information criterion will select such small model:*

$$\lim_{n\to\infty} P\left[min_{i\in I\backslash I_0}(IC(M_{i_o}) - IC(M_i)) > 0 \right] = 1$$

Let $C_{n,i}, C_{n,i_0}$ be the penalty terms for models M_i and $M_{i,0}$ respectively.

For AIC

$$C_{n,i} = 2\text{dim}(\theta_i)$$

and

$$C_{n,i_0} = 2\text{dim}(\theta_{i_0})$$

The limit

$$\lim_{n\to\infty}(C_{n,i} - C_{n,i_0}) = \lim_{n\to\infty}(dim(\theta_i) - dim(\theta_{i_0})) = c$$

Therefore,

$$P(\lim_{n\to\infty}(dim(\theta_i) - dim(\theta_{i_0}))) < 1$$

and AIC is not consistent.

For BIC

$$C_{n,i} = \log n\text{dim}(\theta_i)$$

and

$$C_{n,i_0} = \log n\text{dim}(\theta_{i_0})$$

For $dim(\theta_i) > dim(\theta_{i_0})$

$$\lim_{n\to\infty}(C_{n,i} - C_{n,i_0}) = \lim_{n\to\infty}(dim(\theta_i) - dim(\theta_{i_0})) = \infty$$

Therefore,

$$P(\lim_{n\to\infty}(C_{n,i} - C_{n,i_0})) = 1$$

and BIC is consistent.

Theorem 4. *(Strong Consistency). Denote by I the set of candidate models that all reach the minimum Kullback-Leibler distance to the true model, and I_0 is the subset of I containing model with the smallest dimension. Assume that:*

- *For all $k_0 \neq i_0 \subset I$:*

$$\lim_{n\to\infty} sup \frac{1}{\sqrt{n \log\log n}} \sum_{i=1}^{n} (KL(g; f_{k_0,i}) - KL(g; f_{i_o,i})) \leq \infty$$

- *For all $k_0 \neq i_0 \in I$ the log-likelihood ratio:*

$$\sum_{i=1}^{n} \log \frac{f_{k_0,i}(y_i; \theta^*_{k_0})}{f_{I_0,i}(y_i; \theta^*_{I_0})} = 0(\log\log n), \textit{ almost surely.}$$

Then, the require condition on the penalty is that

$$P(C_{n,k} \geq b_n \log\log n, \textit{for almost all } n) = 1$$

where b_n is a random sequence, almost surely bounded below by a strictly positive number.

1.2.3.2 Efficiency

The property where an information criterion behaves 'almost as well' in terms of Mean Squared Error, or expected squared prediction error as the theoretically best model for the type of squared error loss is called efficiency [11].

Let us assume that we wish to choose the best set of candidate variables in the regres-

sion model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i, i = 1, \ldots, n.$$

For prediction, the loss is usually taken to be the squared prediction error. We wish to choose a set of covariates that minimizes the mean squared error.

$$\sum_{i=1}^{n} E[(\hat{Y}_{S,i} - Y_{t,i})^2 \mid Y_{1,\ldots,n}]$$

where

$\hat{Y}_{S,i}$: the predicted value of the true value $Y_{t,i}$ based on data $Y_1, \ldots, Y_n$.

Let $S_0 \subset S$ the set of covariates of the model that were selected by the information criterion and let $S_0^* \subset S$ be the set of covariates that minimizes the mean squared error.

A model selection criterion is called efficient when

$$\frac{\sum_{i=1}^{n} E[(\hat{Y}_{S_0,i} - Y_{t,i})^2]}{\sum_{i=1}^{n} E[(\hat{Y}_{S_0^*,i} - Y_{t,i})^2]} \longrightarrow 1, \text{as n} \longrightarrow \infty.$$

Theorem 5. *Under regularity assumptions (see e.g. Lee and Karagrigoriou, 2001):*

(i) The criteria AIC and AICc are asymptotically efficient.

(ii) BIC is not asymptotically efficient.

Chapter 2

New Developments in Model Selection Criteria

The criteria discussed in the previous chapter are the most popular ones, which are based on the log-likelihood function. In this chapter recent model selection criteria are presented. These criteria are based on measures of divergence or distance. Mattheou et al. [21] proposed the Divergence Information Criterion (DIC), while Mantalos et al. [20] proposed an improvement of DIC, called the Modified Divergence Information Criterion (MDIC).

Finally, based on the so called pseudodistance we will present a novel criterion, the Pseudodistance Information Criterion (PIC), introduced by Toma et al. [29].

2.1 Divergence Information Criterion

In the previous chapter, Basu-Harris-Hjort-Jones (BHHJ) measure of divergence was presented, which is indexed by a positive parameter α. This measure was used for the development of a minimum divergence estimation method for robust parameter estimation. The index α controls the trade-off between robustness and asymptotic efficiency

of the parameter estimators that are the values of θ that minimize the measure over a parametric space Θ.

In order to develop the Divergence Information Criterion [21], the BHHJ measure was used, applying the same methodology used for AIC.

Assume a random sample $X_1, \ldots, X_n$ with a true distribution g and let f_θ be a candidate model, $\theta \in \Theta$, where Θ is a p-dimensional space.

The goodness of fit quantity in this case is given by:

$$W_\theta = \int f_\theta^{1+\alpha}(z) - (1+\alpha^{-1}) g(z) f_\theta^{\alpha}(z) dz, \quad \alpha > 0 \tag{2.1}$$

The above is the same as the BHHJ measure without the last term which is constant irrespectively of the model f_θ used. Note that it can also be written as:

$$W_\theta = E_{f_\theta}(f_\theta^{\alpha}(z)) - (1+\alpha^{-1}) E_g(f_\theta^{\alpha}(z)), \quad \alpha > 0 \tag{2.2}$$

2.1.1 The expected overall discrepancy

The expected overall discrepancy between g and f_θ is given by:

$$EW_{\hat{\theta}} = E(W_\theta | \theta = \hat{\theta}) \tag{2.3}$$

where $\hat{\theta}$ is a consistent and asymptotically normal estimator of θ.

Under standard regularity assumptions, the expected overall discrepancy at $\theta = \hat{\theta}$ is given by:

$$EW_{\hat{\theta}} = W_{\theta_0} + \frac{(\alpha+1)}{2} E[(\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0)] + ER_n \tag{2.4}$$

where $R_n = o(\| \hat{\theta} - \theta_0 \|^2)$, θ_0 is the true value of the parameter and $J(\theta_0)$ is given by

$$J(\theta_0) = \int u_{\theta_0}(z) u'_{\theta_0}(z) f_{\theta_0}^{1+\alpha}(z) dz \tag{2.5}$$

The natural estimator of (2.2) with respect to g is given by:

$$Q_\theta = \int f_\theta^{1+\alpha}(z) dz - \big(1 + \frac{1}{\alpha}\big) \frac{1}{n} \sum_{i=1}^{n} f_\theta^{\alpha}(x_i). \tag{2.6}$$

Theorem 6. *(see Mattheou et al. [21]). The expectation of Q_θ evaluated at the true point θ_0 is given by:*

$$EQ_{\theta_0} = EQ_{\hat{\theta}} + \frac{\alpha + 1}{2} E[(\theta_0 - \hat{\theta})' J(\theta_0)(\theta_0 - \hat{\theta})] + ER_n \tag{2.7}$$

and the expected overall discrepancy evaluated at $\hat{\theta}$ is given by:

$$EW_{\hat{\theta}} = E[Q_{\hat{\theta}} + (\alpha + 1)(\hat{\theta} - \theta_0)' J(\theta_0)(\hat{\theta} - \theta_0) + R_n] \tag{2.8}$$

where R_n and $J(\theta_0)$ the same as mentioned previously.

2.1.2 The development of DIC criterion

An asymptotically unbiased estimator of n-times the expected overall discrepancy evaluated at $\hat{\theta}$ is given by:

$$DIC(p) = nQ_{\hat{\theta}} + (\alpha + 1)(2\pi)^{-\frac{\alpha}{2}} \left(\frac{1 + \alpha}{1 + 2\alpha} \right)^{1 + \frac{p}{2}} p \tag{2.9}$$

where $\hat{\theta}$ is a consistent and asymptotically normal estimator of θ and p the dimension of θ.

If the maximum likelihood estimating method is used, then the correction term is

adjusted accordingly and the adjusted DIC is given by:

$$DIC_{MLE}(p) = nQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}}(1+\alpha)^{-\frac{p}{2}}p \tag{2.10}$$

In order to penalize strictly more complex models DIC and DIC_{MLE} are adjusted accordingly by removing the denominator $[1+2\alpha]$ in (2.9) and they are both given by:

$$DIC_{STR}(p) = nQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}}(1+\alpha)^{2+\frac{p}{2}}p \tag{2.11}$$

2.2 Modified Divergence Information Criterion

2.2.1 The development of MDIC criterion

DIC criterion, mentioned in the previous section, in preliminary simulation studies for regression models [22] showed that for medium sample size DIC had a very good performance for values of α close to zero. The calculation of the first part of $Q_{\hat{\theta}}$ where

$$Q_{\hat{\theta}} = \int f_{\hat{\theta}}^{1+\alpha}(z)dz - \big(1+\frac{1}{\alpha}\big)\frac{1}{n}\sum_{i=1}^{n} f_{\hat{\theta}}^{\alpha}(x_i)$$

namely the integral $\int f_{\hat{\theta}}^{1+\alpha}(z)dz$ is not computationally attractive for practitioners.

Additional by a simulation study shows [22] that the difference in the calculations of the above integral for the different candidate models is negligible compared with the calculation for the entire quantity $Q_{\hat{\theta}}$. That means that the integral term does not affect significantly the choice of the appropriate model and therefore the criterion can be properly modified. For this reason, Mantalos et al. in [20] proposed a modified version of the DIC criterion, namely the Modified Divergence Information Criterion (MDIC), which is

given by:

$$MDIC(p) = nMQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}}(1+\alpha)^{2+\frac{p}{2}}p \tag{2.12}$$

where

$$MQ_{\hat{\theta}} = -\big[\big(1+\frac{1}{\alpha}\big)\frac{1}{n}\sum_{i=1}^{n} f_{\hat{\theta}}^{\alpha}(x_i)\big] \tag{2.13}$$

Note that as usual MDIC can be considered as an approximately unbiased estimator of the expected overall discrepancy up to a constant, a non-negative quantity that measures the distance between the true unknown model and the candidate model. If we choose the model with the smallest estimator of the expected overall discrepancy, we may end up with a selection with an unnecessarily large order.

2.2.2 Optimal choice for the index α

For practical purposes someone has to decide the optimal choice of the positive index α. Therefore, Mantalos et al. proceeded with a simulation study in [20] using 100 observation series for five different time series models with $\alpha \in [0.01,\ 0.5]$. In the figure below we provide the power of the selection according to the specified value of the index α.

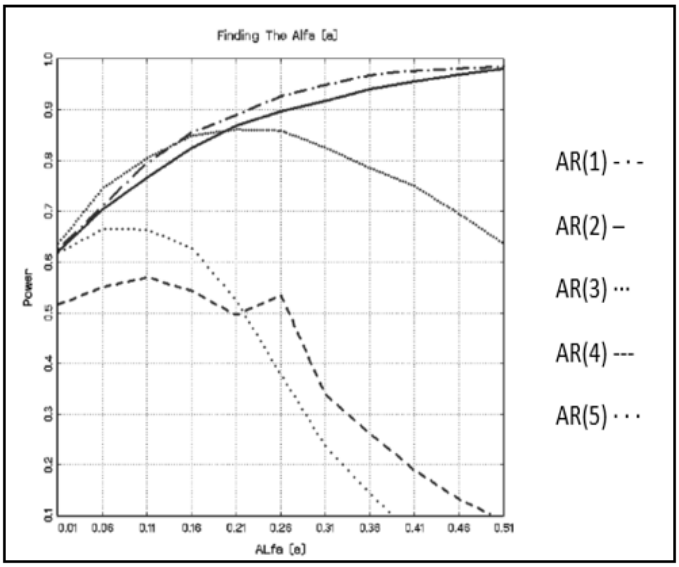

Figure 2.2.1: *Optimal choice of the index* α.

The models considered for the simulation are:

- $AR(1): \ x_t = 1 + 0.65x_{t-1} + \epsilon_t$

- $AR(2): \ x_t = 1 + 1.5x_{t-1} - 0.5x_{t-2} + \epsilon_t$

- $AR(3): \ x_t = 1 + 0.2x_{t-1} + 0.5x_{t-2} - 0.35x_{t-3} + \epsilon_t$

- $AR(4): \ x_t = 1 + 0.2x_{t-1} + 0.5x_{t-2} - 0.35x_{t-3} - 0.22x_{t-4} + \epsilon_t$

- $AR(5): \ x_t = 1 + 0.23x_{t-1} - 0.22x_{t-3} - 0.45x_{t-5} + \epsilon_t$

Figure 2.2.1 shows that the power increases as the value of α increases for small lags (models $AR(1)$ and $AR(2)$). For lags ≥ 3 the power increases up to a value of α and then decreases. In conclusion, Mantalos et al. in [20] shows that an optimal index α value equals 0.25 since it appears to serve a fair balance between small and large lag models.

2.3 Pseudodistance based Information Criterion

2.3.1 Pseudodistances

In order to define new criteria for model selection, we consider the following family of pseudodistances (see [9]). For two probability measures P and Q, admitting densities p and q with respect to the Lebesgue measure, the family of pseudodistances of order $\gamma > 0$ is defined by:

$$R_\gamma(P,Q) = \frac{1}{\gamma+1}\ln\left(\int p^\gamma dP\right) + \frac{1}{\gamma(\gamma+1)}\ln\left(\int q^\gamma dQ\right) - \frac{1}{\gamma}\ln\left(\int p^\gamma dQ\right) \quad (2.14)$$

and satisfies the limit relation

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

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

This family of pseudodistances was originally introduced by Jones et al. [16] where they are called "type 0" divergences. They are also introduced in Broniatowski, Toma and Vajda [9] in the context of decomposable pseudodistances. A pseudodistance satisfies two propeties. It is non-negative and equals to zero if and only if the two measures are equal. Divergences satisfy these two properties, but they are moreover characterized by the so called "information processing property" referring to the invariance of divergences with respect to transformations of the observation space [28]. This property is useful but probably not unavoidable, and pseudodistances may not satisfy this property. That is why the term pseudodistance was adopted, but in literature we can also meet the term "divergence" for this concept.

The concept of pseudodistance cames from the need to place under the same general umbrella all statistical criteria most if not all of which are associated with a distance. Since

the distance used for criteria may or may not be a typical metric, the term "pseudodistance" appeared to be sufficient to include all possible distances that are non-negative with equally if the two arguments coincide.

2.3.2 Minimum Pseudodistance Estimators

Let (P_θ) be a parametric model indexed by $\theta \in \Theta$, where Θ is a p-dimensional parameter space, and p_θ corresponding densities with respect to the Lebesgue measure λ [29]. Let $X_1, \ldots, X_n$ be a random sample on P_{θ_0}, $\theta_0 \in \Theta$. For $\gamma > 0$ fixed, minimum pseudodistance estimators are defined by

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \left\{ \frac{1}{\gamma+1} \ln\left(\int p_\theta^{\gamma+1} \mathrm{d}\lambda \right) - \frac{1}{\gamma} \ln\left(\frac{1}{n} \sum_{i=1}^{n} p_\theta^\gamma(X_i) \right) \right\} \tag{2.15}$$

or equivalently by

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \{ C_\gamma(\theta)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} p_\theta^\gamma(X_i) \} \tag{2.16}$$

where $C_\gamma(\theta) = (\int p_\theta^{\gamma+1} \mathrm{d}\lambda)^{\gamma/(\gamma+1)}$. Denoting $h(x,\theta) := C_\gamma(\theta)^{-1} \cdot p_\theta^\gamma(x)$, these estimators take the form

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta). \tag{2.17}$$

Assume the following regularity conditions of the model (see [29]):

(C1) The density $p_\theta(x)$ has continuous partial derivatives with respect to θ up to 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 $\int [\frac{\partial^2}{\partial^2\theta} h(x,\theta)]_{\theta=\theta_0} \mathrm{d}P_{\theta_0}(x)$ and $\int [\frac{\partial}{\partial\theta} h(x,\theta)]_{\theta=\theta_0} [\frac{\partial}{\partial\theta} h(x,\theta)]^t_{\theta=\theta_0} \mathrm{d}P_{\theta_0}(x)$ exist.

Theorem 7. *(Broniatowski et al., 2012) Assume that the conditions (C1), (C2) and (C3) are fulfilled. Then*

(a) Let $B := \left\{\theta \in \Theta;\ \|\theta - \theta_0\| \leq n^{-1/3}\right\}$. *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* $\widehat{\theta}_n$ *in the interior of* B, *which implies that the estimator* $\widehat{\theta}_n$ *is* $n^{1/3}$*-consistent.*

(b) $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$ *converges in distribution to a centered multivariate normal random variable with covariance matrix*

$$V = S^{-1} M S^{-1} \tag{2.18}$$

where $S := -\int [\frac{\partial^2}{\partial^2 \theta} h(x, \theta)]_{\theta=\theta_0} \mathrm{d}P_{\theta_0}(x)$ *and* $M := \int [\frac{\partial}{\partial \theta} h(x, \theta)]_{\theta=\theta_0} [\frac{\partial}{\partial \theta} h(x, \theta)]^t_{\theta=\theta_0} \mathrm{d}P_{\theta_0}(x)$.

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

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

and

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

2.3.3 The Expected Overall Discrepancy

For $\gamma > 0$ fixed, a simplified form of the pseudodistance is given by:

$$W_\theta = \frac{1}{\gamma + 1} \ln\left(\int p_\theta^{\gamma+1} \mathrm{d}\lambda\right) - \frac{1}{\gamma} \ln\left(\int p_\theta^\gamma q \mathrm{d}\lambda\right) \tag{2.19}$$

which is the same as $R_\gamma(P_\theta, Q)$ given in (2.14) without the middle term that remains constant irrespectively of the model p_θ used.

Under the above setting the Expected Overall Discrepancy, up to a constant is

$$E[W_{\widehat{\theta}_n}] = E[W_\theta|\theta = \widehat{\theta}_n] \tag{2.20}$$

where $\widehat{\theta}_n$ is the minimum pseudodistance estimator, needs to be approximated by an asymptotically unbiased estimator [29]. Note that (2.20) is the same as (2.3).

Proposition 1. *(Toma et al. [29]). When the true model Q belongs to the parametric model (P_θ) for $q = p_{\theta_0}$ and $\theta = \theta_0$, the expected overall discrepancy is given by*

$$E[W_{\widehat{\theta}_n}] = W_{\theta_0} + \frac{1}{2}E[(\widehat{\theta}_n - \theta_0)^t M_\gamma(\theta_0)(\widehat{\theta}_n - \theta_0)] + E[R_n] \tag{2.21}$$

where $R_n = o(\|\widehat{\theta}_n - \theta_0\|^2)$, θ_0 is the true value of the parameter and

$$M_\gamma(\theta_0) := \frac{(\int p_{\theta_0}^{\gamma-1}\dot{p}_{\theta_0}\dot{p}_{\theta_0}^t \mathrm{d}\lambda)(\int p_{\theta_0}^{\gamma+1}\mathrm{d}\lambda) - (\int p_{\theta_0}^{\gamma}\dot{p}_{\theta_0} d\lambda)(\int p_{\theta_0}^{\gamma}\dot{p}_{\theta_0}\mathrm{d}\lambda)^t}{(\int p_{\theta_0}^{\gamma+1}\mathrm{d}\lambda)^2} \tag{2.22}$$

where $\dot{p}_\theta$ is the first order derivatives of p_θ with respect to θ.

As in (2.6), for a given $\theta \in \Theta$, a natural estimator of W_θ is defined by

$$Q_\theta := \frac{1}{\gamma+1}\ln\left(\int p_\theta^{\gamma+1}\mathrm{d}\lambda\right) - \frac{1}{\gamma}\ln\left(\frac{1}{n}\sum_{i=1}^{n} p_\theta^\gamma(X_i)\right). \tag{2.23}$$

and satisfies the following two propositions (see [29]):

Proposition 2. *When Q belongs to (P_θ), under regularity conditions, it holds*

$$\begin{aligned} E[W_{\widehat{\theta}_n}] &= E[Q_{\widehat{\theta}_n}] + E[(\theta_0 - \widehat{\theta}_n)^t M_\gamma(\theta_0)(\theta_0 - \widehat{\theta}_n)] + \\ &\quad + \frac{1}{2\gamma n}\left[1 - \frac{\int p_{\theta_0}^{2\gamma+1}\mathrm{d}\lambda}{\left(\int p_{\theta_0}^{\gamma+1}\mathrm{d}\lambda\right)^2}\right] + E\left[R_n\right] + \frac{1}{\gamma}E\left[R_n'\right], \end{aligned} \tag{2.24}$$

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

Proposition 3. *Under regularity conditions, when $Q = P_{\theta_0}$, it holds*

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

(b) $\sqrt{n}(Q_{\widehat{\theta}_n} - W_{\theta_0})$ converges in distribution to a centered univariate normal with variance $\frac{\sigma^2(\theta_0)}{\gamma^2 R_\gamma(\theta_0)^2}$.

2.3.4 The development of PIC criterion

2.3.4.1 The case of univariate normal family

For the case where the candidate model (P_θ) is the univariate normal model with $\theta = (\mu, \sigma)$ and the true model Q belongs to (P_θ) it is easy to see that

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

where V is the asymptotic covariance matrix given in Theorem 7 and $A(\gamma)$ is a matrix which can be approximated by the identity matrix I_d for small values of $\gamma \in (0, 0.4)$.

The distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is asymptotically multivariate normal so that the statistic $n(\theta_0 - \widehat{\theta}_n)^t V^{-1}(\theta_0 - \widehat{\theta}_n)$ has approximately a χ^2_p distribution (see [29]).

For small values of γ and large n we have:

$$E[(\theta_0 - \widehat{\theta}_n)^t M_\gamma(\theta_0)(\theta_0 - \widehat{\theta}_n)] \approx \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \cdot \frac{d}{n} \tag{2.25}$$

Furthermore, under the normal model we have:

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

Thus, Proposition 2 becomes:

$$E[W_{\widehat{\theta}_n}] \cong E[Q_{\widehat{\theta}_n}] + \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \cdot \frac{d}{n} + \frac{1}{2\gamma n}\left[1 - \frac{\gamma+1}{\sqrt{2\gamma+1}}\right] + E\left[R_n\right] + \frac{1}{\gamma}E\left[R'_n\right]. \quad (2.26)$$

Proposition 4. *An asymptotically unbiased estimator of the expected overall discrepancy in (2.20) is given by*

$$Q_{\widehat{\theta}_n} + \frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}} \cdot \frac{d}{n} + \frac{1}{2\gamma n}\left[1 - \frac{\gamma+1}{\sqrt{2\gamma+1}}\right] \quad (2.27)$$

where $\widehat{\theta}_n$ is a minimum pseudodistance estimator.

The selection criteria based on Proposition 4 are consistent [29].

2.3.4.2 The case of linear regression models

Consider the linear regression model:

$$Y = \alpha + \beta^t X + e \quad (2.28)$$

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 $Y_i = \alpha + \beta^t X_i + e_i$.

Consider the joint distribution of the entire data and write a pseudodistance between the hypothesized model and the true model corresponding to the data. Let P_θ, $\theta := (\alpha, \beta, \sigma)$, be the probability measure associated with the hypothesized model given by the random vector $Z = (X, Y)$ and Q the probability measure associated with the true model corresponding to the data [29]. Denote by p_θ and q the corresponding densities.

For $\gamma > 0$, the pseudodistance between P_θ and Q is defined by:

$$R_\gamma(P_\theta, Q) := \frac{1}{\gamma+1}\ln\left(\int p_\theta^\gamma(x,y)dP_\theta(x,y)\right) + \frac{1}{\gamma(\gamma+1)}\ln\left(\int q^\gamma(x,y)dQ(x,y)\right) - \\ -\frac{1}{\gamma}\ln\left(\int p_\theta^\gamma(x,y)dQ(x,y)\right) \tag{2.29}$$

Then the estimator $Q_{\widehat{\theta}_n}$ can be written as

$$Q_{\widehat{\theta}_n} = \min_{\alpha,\beta,\sigma}\left\{\frac{1}{\gamma+1}\ln\left(\frac{1}{(\sigma\sqrt{2\pi})^\gamma\sqrt{\gamma+1}}\right) - \frac{1}{\gamma}\ln\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\sigma\sqrt{2\pi})^\gamma}\cdot\exp\left(-\frac{\gamma}{2\sigma^2}(Y_i-\alpha-\beta^tX_i)^2\right)\right)\right\} \tag{2.30}$$

An asymptotically unbiased estimator of the expected overall discrepancy in this case is given by:

$$Q_{\widehat{\theta}_n} + \frac{1}{n}\cdot\frac{(\gamma+1)^2}{(2\gamma+1)^{3/2}}\left[(d-1)+\frac{3\gamma^2+4\gamma+2}{2(\gamma+1)(2\gamma+1)}\right] + \frac{1}{2\gamma n}\left[1-\left(\frac{\gamma+1}{\sqrt{2\gamma+1}}\right)^d\right]. \tag{2.31}$$

For different linear regression models, we can ignore the terms depending only on n and γ. Therefore the following model selection criterion can be used:

$$Q_{\widehat{\theta}_n} + \frac{(\gamma+1)^2}{(\sqrt{2\gamma+1})^3}\cdot\frac{d}{n} - \frac{1}{2\gamma n}\left(\frac{\gamma+1}{\sqrt{2\gamma+1}}\right)^d. \tag{2.32}$$

Chapter 3

Case Studies

In this chapter, a simulation study will be presented in order to check how Pseudodistance Information Criterion (PIC) performs. We also applied PIC in a real case study to evaluate the accuracy of PIC in a real problem. Finally, are formulated the conclusions regarding the use of PIC.

3.1 Simulation Study

In order to check the performance of PIC we proceed with a simulation study using the PIC, the Modified Divergence Information Criterion MDIC [20], the Akaike Information Criterion AIC [2] and the Bayesian Information Criterion BIC [26].

The simulation study has the following characteristics:

- A set of 20, 30, 40, 50, 75, 100, 200, 500 observations were used.
- 4 variables X_1, X_2, X_3, X_4 were independently generated from the normal distributions $N(0,3)$, $N(1,3)$, $N(2,3)$, and $N(3,3)$ respectively.
- The first 2 of these variables were planned to be used to generate values of Y_i,

i=1,. . .,(20, 30, 40, 50, 75, 100, 200, 500) using the following model specification:

$$Y_i = a_0 + a_1 X_{1,i} + a_2 X_{2,i} + \varepsilon_i$$

- with $a_0 = a_1 = a_2 = 1$
- and $\varepsilon_i \sim$ N(0,1).

Due though to contamination of the above model from the model

$$Y_i = 1 + X_{1,i} + X_{2,i} + \varepsilon_i^*$$

with $\varepsilon_i^* \sim N(5, 1)$ the simulated values were generated from the model

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

with d_1, $d_2 \in (0, 1)$ such that $d_1 + d_2 = 1$.

Different values of d_1 and d_2 have been used in the present study. Two such examples for $d_1 = 0.8$, $d_2 = 0.2$ and $d_1 = 1$, $d_2 = 0$ are presented below:

$$Y_i = 0.8(1 + X_{1,i} + X_{2,i} + \varepsilon_i) + 0.2(1 + X_{1,i} + X_{2,i} + \varepsilon_i^*)$$

and

$$Y_i = 1 + X_{1,i} + X_{2,i} + \varepsilon_i.$$

In addition to the above, the cases:

- $d_1 = 0.9$, $d_2 = 0.1$
- $d_1 = 0.95$, $d_2 = 0.05$

have also been used in the present simulation study. Different values of γ were also been used in this study.

The reason for introducing contamination into the simulation study was to put into a test the robust features of the PIC criterion.

With a set of 4 possible regressors there are $2^4 - 1 = 15$ possible specifications that include at least one regressor. These 15 possible regression specifications constitute the set of candidate models for the experiment. As a result the candidate set consists of 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 $(X_{j_1}, X_{j_2}, X_{j_3})$, with $j_i \in \{1, 2, 3, 4\}, i = 1, 2, 3$ of the 4 regressors X_1, X_2, X_3, and X_4. 50 such experiments were performed with the intention to select the best model among the available candidate models.

First we consider the standard AIC criterion given by

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

where n the sample size, p the number of covariates of the model and $\hat{\sigma}_p^2$ the estimate of the variance of the model with p variables. The above form of AIC is the one associated with the normal distribution.

We have also chosen to include in the simulations the Bayesian Information Criterion [26] because of its consistency property. The BIC is given by

$$BIC = n \log \hat{\sigma_p}^2 + (p + 2) \log n.$$

The MDIC given by

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

has also been used with $\alpha = 0.25$ with

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

Finally the PIC [29] is given by

$$PIC = Q_{\widehat{\theta}_n} + \frac{(\gamma+1)^2}{(\sqrt{2\gamma+1})^3}\cdot\frac{d}{n} - \frac{1}{2\gamma n}\left(\frac{\gamma+1}{\sqrt{2\gamma+1}}\right)^d. \tag{3.1}$$

For each of the 50 experiments the value of each of the above model selection criteria was calculated for each of the 15 possible regression specifications under consideration. As a result, for each of the 50 experiments and for each model selection criterion the 15 candidate models were ranked from 1st to 15th according to the value of the criterion. Recall that the model chosen by a criterion is the one for which the value of the criterion is the lowest among all 15 candidate models.

3.2 Simulation Results

The tables in this section present for each selection criterion, the proportion of times each candidate model has been selected by the criterion for different values of γ and different values of contamination (1, 0.95, 0.9, 0.8). Observe that all selections contain the correct variables of the model, namely X_1 and X_2.

3.2.1 10% Contamination

Table 3.1: Selected models by model selection criteria ($n = 20$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	80	**84**	**90**	**82**	82	**80**	80	**86**
	X_1, X_2, X_3	(20)	(16)	(10)	(18)	(18)	(20)	(20)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	60	52	56	62	64	54	52	50
	X_1, X_2, X_3	(40)	(48)	(44)	(38)	(36)	(46)	(48)	(50)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	76	70	78	72	84	76	76	74
	X_1, X_2, X_3	(24)	(30)	(22)	(28)	(16)	(24)	(24)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**86**	76	88	74	**92**	78	**86**	80
	X_1, X_2, X_3	(14)	(24)	(12)	(26)	(8)	(22)	(14)	(20)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.2: Selected models by model selection criteria ($n = 30$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**84**	**88**	**86**	**88**	**92**	80	**88**	**86**
	X_1, X_2, X_3	(16)	(12)	(14)	(12)	(8)	(20)	(12)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	66	76	70	62	58	54	60	68
	X_1, X_2, X_3	(34)	(24)	(30)	(38)	(42)	(46)	(40)	(32)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	80	86	82	78	82	78	76	84
	X_1, X_2, X_3	(20)	(14)	(18)	(22)	(18)	(22)	(34)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**84**	86	84	84	86	**84**	82	**86**
	X_1, X_2, X_3	(16)	(14)	(16)	(16)	(14)	(16)	(18)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.3: Selected models by model selection criteria ($n = 40$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**96**	88	80	**92**	**86**	86	**86**	**92**
	X_1, X_2, X_3	(4)	(12)	(20)	(8)	(14)	(14)	(14)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	60	70	62	58	74	64	74
	X_1, X_2, X_3	(30)	(40)	(30)	(38)	(42)	(26)	(36)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	92	**90**	**90**	82	76	88	76	90
	X_1, X_2, X_3	(8)	(10)	(10)	(18)	(24)	(12)	(24)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	86	88	82	76	**92**	78	90
	X_1, X_2, X_3	(10)	(14)	(12)	(18)	(24)	(8)	(22)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.4: Selected models by model selection criteria ($n = 50$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**92**	**88**	**92**	90	82	**94**	**86**	**90**
	X_1, X_2, X_3	(8)	(12)	(8)	(10)	(18)	(6)	(14)	(10)
	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	74
	X_1, X_2, X_3	(30)	(36)	(38)	(36)	(34)	(26)	(28)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**92**	**88**	82	**92**	**88**	88	**86**	88
	X_1, X_2, X_3	(8)	(12)	(18)	(8)	(12)	(12)	(14)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**92**	86	76	88	84	88	**86**	88
	X_1, X_2, X_3	(8)	(14)	(24)	(12)	(16)	(12)	(14)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.5: Selected models by model selection criteria ($n = 75$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	90	**96**	**92**	80	**92**	90	**88**	90
	X_1, X_2, X_3	(10)	(4)	(8)	(20)	(8)	(10)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	72	72	60	70	66	66	68	68
	X_1, X_2, X_3	(28)	(28)	(40)	(30)	(34)	(34)	(32)	(32)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**92**	94	90	**92**	88	**92**	**88**	**92**
	X_1, X_2, X_3	(8)	(6)	(10)	(8)	(12)	(8)	(12)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	92	86	90	84	**92**	84	84
	X_1, X_2, X_3	(10)	(8)	(14)	(10)	(16)	(8)	(16)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.6: Selected models by model selection criteria ($n = 100$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	88	92	**96**	**88**	88	88	86	96
	X_1, X_2, X_3	(12)	(8)	(4)	(12)	(12)	(12)	(14)	(4)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	68	72	78	66	70	78	60	76
	X_1, X_2, X_3	(32)	(28)	(22)	(34)	(30)	(22)	(40)	(24)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**98**	**98**	**96**	**88**	**92**	**94**	**92**	**98**
	X_1, X_2, X_3	(2)	(2)	(4)	(12)	(8)	(6)	(8)	(2)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	90	**96**	84	82	90	82	88
	X_1, X_2, X_3	(10)	(10)	(4)	(16)	(18)	(10)	(18)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.7: Selected models by model selection criteria ($n = 200$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	86	88	86	90	88	82	92	96
	X_1, X_2, X_3	(14)	(12)	(14)	(10)	(12)	(18)	(8)	(4)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	72	80	78	68	78	56	66	66
	X_1, X_2, X_3	(28)	(20)	(22)	(32)	(22)	(44)	(34)	(34)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**98**	**96**	**98**	**92**	**96**	**92**	**96**	**98**
	X_1, X_2, X_3	(2)	(4)	(2)	(8)	(4)	(8)	(4)	(2)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	86	96	80	86	82	88	90
	X_1, X_2, X_3	(10)	(14)	(4)	(20)	(14)	(18)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.8: Selected models by model selection criteria ($n = 500$)

$d_1 = 0.9$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	92	94	94	94	90	94	84	92
	X_1, X_2, X_3	(8)	(6)	(6)	(6)	(10)	(6)	(16)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	78	72	56	70	60	82	76	74
	X_1, X_2, X_3	(22)	(28)	(44)	(30)	(40)	(18)	(24)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**98**	**98**	**96**	**100**	**94**	**100**	**94**	**100**
	X_1, X_2, X_3	(2)	(2)	(4)		(6)		(6)	
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	92	88	82	90	78	94	88	88
	X_1, X_2, X_3	(8)	(12)	(8)	(10)	(22)	(6)	(12)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

3.2.2 5% Contamination

Table 3.9: Selected models by model selection criteria ($n = 20$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	82	**88**	80	**94**	82	**88**	86	86
	X_1, X_2, X_3	(18)	(12)	(20)	(6)	(18)	(12)	(14)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	78	50	66	70	66	64	66	70
	X_1, X_2, X_3	(22)	(50)	(34)	(30)	(34)	(36)	(34)	(30)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	84	64	74	84	84	76	82	84
	X_1, X_2, X_3	(16)	(36)	(26)	(16)	(16)	(24)	(18)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**90**	74	**82**	88	**88**	80	**88**	**90**
	X_1, X_2, X_3	(10)	(26)	(18)	(12)	(12)	(20)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.10: Selected models by model selection criteria ($n = 30$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	80	**88**	**94**	**92**	82	82	**86**	**94**
	X_1, X_2, X_3	(20)	(12)	(6)	(8)	(18)	(18)	(14)	(6)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	66	58	68	72	68	72	58	70
	X_1, X_2, X_3	(34)	(42)	(32)	(28)	(32)	(28)	(42)	(30)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	86	74	82	90	82	**96**	80	84
	X_1, X_2, X_3	(14)	(26)	(18)	(10)	(18)	(4)	(20)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**92**	74	84	90	**86**	**96**	82	86
	X_1, X_2, X_3	(8)	(26)	(16)	(10)	(14)	(4)	(18)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.11: Selected models by model selection criteria ($n = 40$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	82	**90**	**90**	**94**	**92**	**92**	**88**	**90**
	X_1, X_2, X_3	(18)	(10)	(10)	(6)	(8)	(8)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	72	66	64	62	72	78	56	60
	X_1, X_2, X_3	(28)	(34)	(36)	(38)	(28)	(22)	(44)	(40)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**86**	88	88	76	86	88	84	86
	X_1, X_2, X_3	(14)	(12)	(12)	(24)	(14)	(12)	(16)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**86**	88	86	74	86	88	82	84
	X_1, X_2, X_3	(14)	(12)	(14)	(26)	(14)	(12)	(18)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.12: Selected models by model selection criteria ($n = 50$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	94	**92**	**92**	**88**	84	90	**88**	80
	X_1, X_2, X_3	(6)	(8)	(8)	(12)	(16)	(10)	(12)	(20)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	62	66	68	70	72	58	64
	X_1, X_2, X_3	(30)	(38)	(34)	(32)	(30)	(28)	(42)	(36)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**96**	82	**92**	86	**92**	**92**	86	**88**
	X_1, X_2, X_3	(4)	(18)	(8)	(14)	(8)	(8)	(14)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	78	88	86	86	90	82	82
	X_1, X_2, X_3	(10)	(22)	(12)	(14)	(14)	(10)	(18)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.13: Selected models by model selection criteria ($n = 75$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**94**	92	82	**94**	92	**84**	92	86
	X_1, X_2, X_3	(6)	(8)	(18)	(6)	(8)	(16)	(8)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	66	72	54	62	66	62	60	78
	X_1, X_2, X_3	(34)	(28)	(46)	(38)	(34)	(38)	(40)	(22)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	90	**96**	**88**	90	**94**	**84**	**94**	**90**
	X_1, X_2, X_3	(10)	(4)	(12)	(10)	(6)	(16)	(6)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	86	94	74	84	92	82	84	88
	X_1, X_2, X_3	(14)	(6)	(26)	(16)	(8)	(18)	(16)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.14: Selected models by model selection criteria ($n = 100$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	90	88	**92**	90	**98**	**96**	**92**	90
	X_1, X_2, X_3	(10)	(12)	(8)	(10)	(2)	(4)	(8)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	78	78	66	82	68	68	58
	X_1, X_2, X_3	(30)	(22)	(22)	(34)	(18)	(32)	(32)	(42)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**96**	**92**	**92**	**94**	96	94	88	**96**
	X_1, X_2, X_3	(4)	(8)	(8)	(6)	(4)	(6)	(12)	(4)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	88	82	90	94	84	88	88
	X_1, X_2, X_3	(10)	(12)	(18)	(10)	(6)	(16)	(12)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.15: Selected models by model selection criteria ($n = 200$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	94 (6)	92 (8)	92 (8)	84 (16)	92 (8)	**96** (4)	92 (8)	86 (14)
AIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	74 (26)	74 (26)	62 (38)	68 (32)	72 (28)	64 (36)	86 (14)	84 (16)
BIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	**100**	**96** (4)	**96** (4)	**96** (4)	**96** (4)	94 (6)	**100**	**100**
MDIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	94 (6)	88 (12)	80 (20)	88 (12)	86 (14)	90 (10)	96 (4)	94 (6)

Table 3.16: Selected models by model selection criteria ($n = 500$)

$d_1 = 0.95$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	92 (8)	92 (8)	88 (12)	94 (6)	92 (8)	88 (12)	86 (14)	92 (8)
AIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	74 (26)	68 (32)	72 (28)	66 (34)	62 (38)	78 (22)	66 (34)	64 (36)
BIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	**96** (4)	**98** (2)	**100**	**100**	**96** (4)	**98** (2)	**98** (2)	**100**
MDIC	X_1, X_2 X_1, X_2, X_3 X_1, X_2, X_4 X_1, X_2, X_3, X_4	94 (6)	86 (14)	90 (10)	90 (10)	78 (22)	88 (12)	86 (14)	86 (14)

3.2.3 0% Contamination (no-contaminated data)

Table 3.17: Selected models by model selection criteria ($n = 20$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**86**	86	86	86	**88**	82	**92**	**80**
	X_1, X_2, X_3	(14)	(14)	(14)	(14)	(12)	(18)	(8)	(20)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	64	74	62	58	64	62	70	60
	X_1, X_2, X_3	(36)	(26)	(38)	(42)	(36)	(38)	(30)	(40)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	78	90	78	80	82	80	74	72
	X_1, X_2, X_3	(22)	(10)	(22)	(20)	(18)	(20)	(26)	(28)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	84	**92**	**88**	**88**	**88**	**88**	80	72
	X_1, X_2, X_3	(16)	(8)	(12)	(12)	(12)	(12)	(20)	(28)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.18: Selected models by model selection criteria ($n = 30$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	82	**86**	**92**	**82**	82	**84**	**84**	**90**
	X_1, X_2, X_3	(18)	(14)	(8)	(18)	(18)	(16)	(16)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	64	70	56	58	66	64	66	64
	X_1, X_2, X_3	(36)	(30)	(44)	(42)	(34)	(36)	(34)	(36)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**88**	**86**	78	78	82	74	82	82
	X_1, X_2, X_3	(12)	(14)	(22)	(22)	(18)	(26)	(18)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**88**	84	80	78	**84**	74	80	82
	X_1, X_2, X_3	(12)	(16)	(20)	(22)	(16)	(26)	(20)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.19: Selected models by model selection criteria ($n = 40$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	82	88	88	**80**	**90**	**84**	82	**92**
	X_1, X_2, X_3	(18)	(12)	(12)	(20)	(10)	(16)	(18)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	78	62	56	62	64	72	64
	X_1, X_2, X_3	(30)	(22)	(38)	(44)	(38)	(36)	(28)	(36)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**88**	88	**92**	74	86	**84**	**92**	84
	X_1, X_2, X_3	(12)	(12)	(8)	(26)	(14)	(16)	(8)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	82	**90**	88	72	80	**84**	90	82
	X_1, X_2, X_3	(18)	(10)	(12)	(28)	(20)	(16)	(10)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.20: Selected models by model selection criteria ($n = 50$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**94**	**90**	80	84	**90**	**94**	**88**	**90**
	X_1, X_2, X_3	(6)	(10)	(20)	(16)	(10)	(6)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	64	68	62	68	66	64	62	44
	X_1, X_2, X_3	(34)	(32)	(38)	(32)	(34)	(36)	(38)	(56)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	86	86	**86**	**90**	86	**94**	82	82
	X_1, X_2, X_3	(14)	(14)	(14)	(10)	(14)	(6)	(18)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	84	84	82	88	84	90	82	74
	X_1, X_2, X_3	(16)	(16)	(18)	(12)	(16)	(10)	(18)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.21: Selected models by model selection criteria ($n = 75$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**90**	88	**92**	78	84	86	86	82
	X_1, X_2, X_3	(10)	(12)	(8)	(22)	(16)	(14)	(14)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	66	56	70	68	74	70	66	78
	X_1, X_2, X_3	(34)	(44)	(30)	(32)	(26)	(30)	(34)	(22)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	88	**94**	90	**90**	**90**	**92**	**92**	**92**
	X_1, X_2, X_3	(12)	(6)	(10)	(10)	(10)	(8)	(8)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	86	86	82	82	86	80	84	86
	X_1, X_2, X_3	(14)	(14)	(18)	(18)	(14)	(20)	(16)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.22: Selected models by model selection criteria ($n = 100$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	94	96	**92**	92	**96**	**90**	94	82
	X_1, X_2, X_3	(6)	(4)	(8)	(8)	(4)	(10)	(6)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	78	74	72	74	70	62	74	72
	X_1, X_2, X_3	(22)	(26)	(28)	(26)	(30)	(38)	(26)	(28)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**96**	**100**	**92**	**96**	94	**90**	**100**	**94**
	X_1, X_2, X_3	(4)		(8)	(4)	(6)	(10)		(6)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	94	92	86	90	86	80	94	90
	X_1, X_2, X_3	(6)	(8)	(14)	(10)	(14)	(20)	(6)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.23: Selected models by model selection criteria ($n = 200$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	90	88	86	88	90	92	**90**	82
	X_1, X_2, X_3	(10)	(12)	(14)	(12)	(10)	(8)	(10)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	54	68	68	70	78	64	60	64
	X_1, X_2, X_3	(46)	(32)	(32)	(30)	(22)	(36)	(40)	(36)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**100**	**96**	**96**	**98**	**94**	**100**	**90**	**92**
	X_1, X_2, X_3		(4)	(4)	(2)	(6)		(10)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	82	90	82	86	88	92	78	82
	X_1, X_2, X_3	(18)	(10)	(18)	(14)	(12)	(8)	(22)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.24: Selected models by model selection criteria ($n = 500$)

$d_1 = 1$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**98**	82	82	94	96	88	84	98
	X_1, X_2, X_3	(2)	(18)	(18)	(6)	(4)	(12)	(16)	(2)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	82	74	76	68	74	62	70	78
	X_1, X_2, X_3	(18)	(26)	(24)	(32)	(26)	(38)	(30)	(22)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	94	**96**	**98**	**96**	**98**	**96**	**100**	**100**
	X_1, X_2, X_3	(6)	(4)	(2)	(4)	(2)	(4)		
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	90	82	90	84	86	76	88	94
	X_1, X_2, X_3	(10)	(18)	(10)	(16)	(14)	(24)	(12)	(6)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

3.2.4 20% Contamination

Table 3.25: Selected models by model selection criteria ($n = 20$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**90**	84	**88**	**84**	**92**	**90**	**86**	74
	X_1, X_2, X_3	(10)	(16)	(12)	(16)	(8)	(10)	(14)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	62	56	52	56	66	56	60	60
	X_1, X_2, X_3	(38)	(44)	(48)	(44)	(34)	(44)	(40)	(40)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	74	76	60	74	72	68	70	74
	X_1, X_2, X_3	(26)	(24)	(40)	(26)	(28)	(32)	(30)	(26)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	86	**86**	64	78	84	80	74	**82**
	X_1, X_2, X_3	(14)	(14)	(36)	(22)	(16)	(20)	(26)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.26: Selected models by model selection criteria ($n = 30$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**88**	88	**92**	80	78	72	82	82
	X_1, X_2, X_3	(12)	(12)	(8)	(20)	(22)	(28)	(18)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	68	68	66	66	64	56	60	62
	X_1, X_2, X_3	(32)	(32)	(34)	(34)	(36)	(44)	(40)	(38)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**88**	84	82	82	**82**	**88**	82	84
	X_1, X_2, X_3	(12)	(16)	(18)	(18)	(18)	(12)	(18)	(16)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**88**	**94**	86	**88**	**82**	**88**	**88**	**86**
	X_1, X_2, X_3	(12)	(6)	(14)	(12)	(18)	(12)	(12)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.27: Selected models by model selection criteria ($n = 40$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**92**	**94**	**92**	**86**	**94**	**92**	84	86
	X_1, X_2, X_3	(8)	(6)	(8)	(14)	(6)	(8)	(16)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	68	64	66	54	70	66	62	58
	X_1, X_2, X_3	(32)	(36)	(34)	(46)	(30)	(34)	(38)	(42)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	84	86	86	80	90	**92**	**88**	86
	X_1, X_2, X_3	(16)	(14)	(14)	(20)	(10)	(8)	(12)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	86	86	88	80	90	90	**88**	**88**
	X_1, X_2, X_3	(14)	(14)	(12)	(20)	(10)	(10)	(12)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.28: Selected models by model selection criteria ($n = 50$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	86	**96**	94	**90**	88	86	**90**	88
	X_1, X_2, X_3	(14)	(4)	(6)	(10)	(12)	(14)	(10)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	74	64	82	62	64	78	72	68
	X_1, X_2, X_3	(26)	(36)	(18)	(38)	(36)	(22)	(28)	(32)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**94**	86	**96**	86	**90**	**88**	**90**	**90**
	X_1, X_2, X_3	(6)	(14)	(4)	(14)	(10)	(12)	(10)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**94**	82	98	82	86	**88**	**90**	88
	X_1, X_2, X_3	(6)	(18)	(2)	(18)	(14)	(12)	(10)	(12)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.29: Selected models by model selection criteria ($n = 75$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	92	82	92	**94**	90	80	**92**	**92**
	X_1, X_2, X_3	(8)	(18)	(8)	(6)	(10)	(20)	(8)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	74	72	74	58	78	68	68	62
	X_1, X_2, X_3	(26)	(28)	(26)	(42)	(22)	(32)	(32)	(38)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**98**	**88**	**98**	90	**94**	**92**	**92**	**92**
	X_1, X_2, X_3	(2)	(12)	(2)	(10)	(6)	(8)	(8)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	96	**88**	94	88	92	90	90	90
	X_1, X_2, X_3	(4)	(12)	(6)	(12)	(8)	(10)	(10)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.30: Selected models by model selection criteria ($n = 100$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	**94**	94	94	**92**	88	88	**94**	**92**
	X_1, X_2, X_3	(6)	(6)	(6)	(8)	(12)	(12)	(6)	(8)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	82	78	70	68	68	72	72
	X_1, X_2, X_3	(30)	(18)	(22)	(30)	(32)	(32)	(28)	(28)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	90	**96**	**98**	90	**96**	**94**	88	90
	X_1, X_2, X_3	(10)	(4)	(2)	(10)	(4)	(6)	(12)	(10)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	86	96	92	86	92	90	88	82
	X_1, X_2, X_3	(14)	(4)	(8)	(14)	(8)	(10)	(12)	(18)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.31: Selected models by model selection criteria ($n = 200$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	92	94	88	88	94	84	92	88
	X_1, X_2, X_3	(14)	(8)	(14)	(6)	(8)	(2)	(14)	(42)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	78	66	66	74	62	68	64	78
	X_1, X_2, X_3	(22)	(34)	(34)	(26)	(38)	(32)	(36)	(22)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**96**	**98**	**96**	**92**	**96**	**94**	**96**	**98**
	X_1, X_2, X_3	(4)	(2)	(4)	(8)	(4)	(6)	(4)	(2)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	**96**	90	88	**92**	88	84	86	96
	X_1, X_2, X_3	(4)	(10)	(12)	(8)	(12)	(16)	(14)	(4)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

Table 3.32: Selected models by model selection criteria ($n = 500$)

$d_1 = 0.8$									
		%							
Criteria	**Variables**	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.15$	$\gamma = 0.2$	$\gamma = 0.25$	$\gamma = 0.3$	$\gamma = 0.4$
PIC	X_1, X_2	88	92	90	92	90	94	88	94
	X_1, X_2, X_3	(12)	(8)	(10)	(8)	(10)	(6)	(12)	(6)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
AIC	X_1, X_2	70	78	72	64	76	60	58	64
	X_1, X_2, X_3	(30)	(22)	(28)	(36)	(24)	(40)	(42)	(36)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
BIC	X_1, X_2	**98**	**96**	**92**	**100**	**98**	**100**	**98**	**98**
	X_1, X_2, X_3	(2)	(4)	(8)		(2)		(2)	(2)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								
MDIC	X_1, X_2	84	92	86	90	90	88	86	86
	X_1, X_2, X_3	(16)	(8)	(14)	(10)	(10)	(12)	(14)	(14)
	X_1, X_2, X_4								
	X_1, X_2, X_3, X_4								

The simulation study showed that for $d_1 = 0.9$ and $d_1 = 0.95$ PIC has the highest success rate for small sample sizes. For medium sample sizes PIC is comparable to BIC, while for large sample sizes PIC ranked second and BIC ranked first.

For the case where we don't have contamination and $d_1 = 1$ for small sample sizes PIC has the highest success rate while for medium and large sample sizes PIC ranked second and BIC was first.

For the case where $d_1 = 0.8$ PIC has the highest success rate for small sample sizes, while the second one was MDIC. For medium sample sizes PIC is comparable to BIC. Finally, PIC ranked second for large sample sizes and BIC was first.

Also, the results of the simulation study show that for small and medium sample sizes, values of γ between 0.1 and 0.25 are associated with a very good performance of the new PIC criterion both in the uncontaminated case, as well as for a fraction of contamination smaller than 20%. This leads to the idea of choosing of an optimal value of the tuning parameter γ in order to assure the robustness and the efficiency of the procedure. This issue will be considered in future research studies.

For a visual evaluation of the proposed PIC model selection criterion against all other competitors we provide a representative set of four (4) graphs for sample size $n = 40$, one for each of the contamination proportions considered in this work (Figures 3.1-3.4), where on the horizontal axis are the different values of γ. The graphs show the proportion of the correct model selection for each γ. All figures show the satisfactory performance of the proposed PIC criterion (blue) in comparison with AIC (red), BIC(grey) and MDIC(yellow).

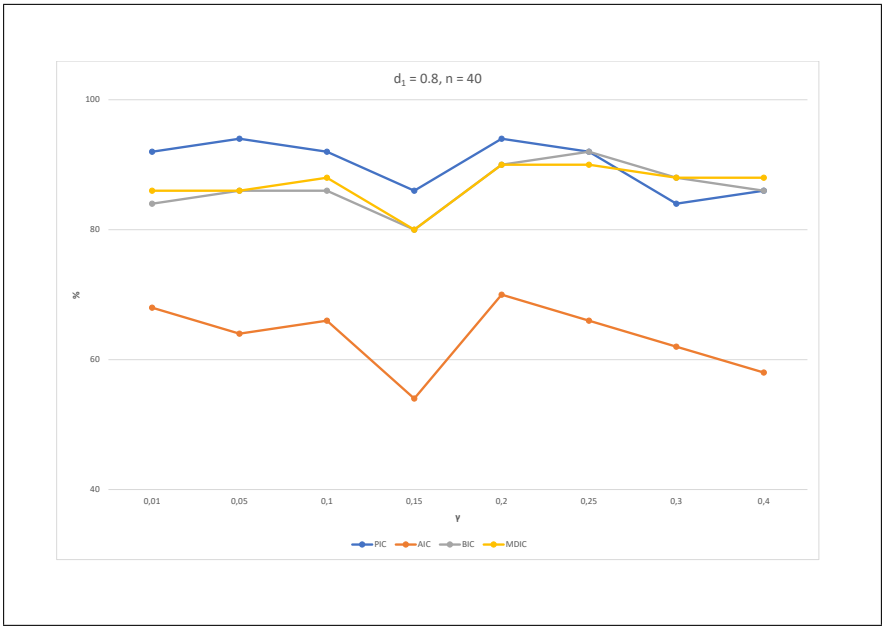

Figure 3.1: Proportion of the correct model selected ($n = 40$, 20% contamination).

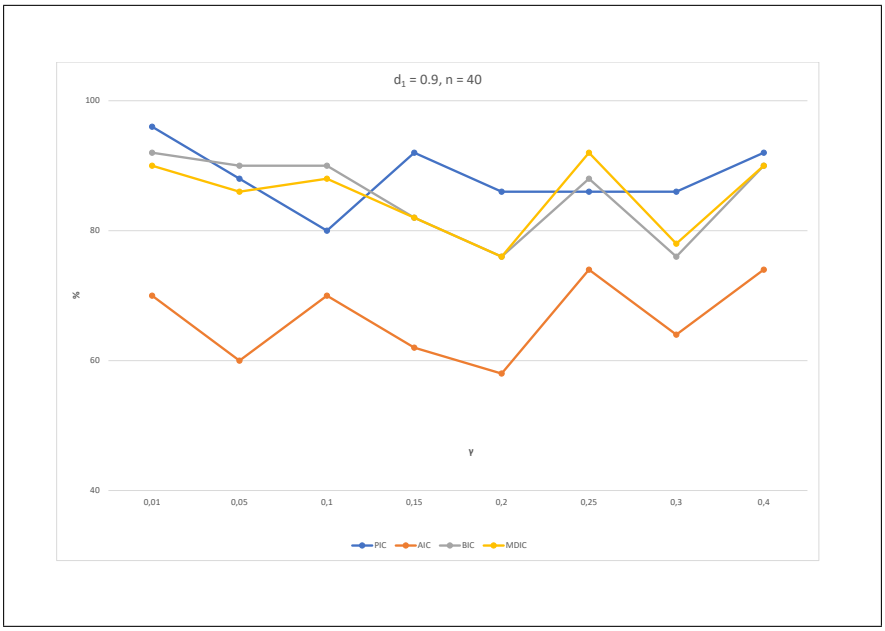

Figure 3.2: Proportion of the correct model selected ($n = 40$, 10% contamination).

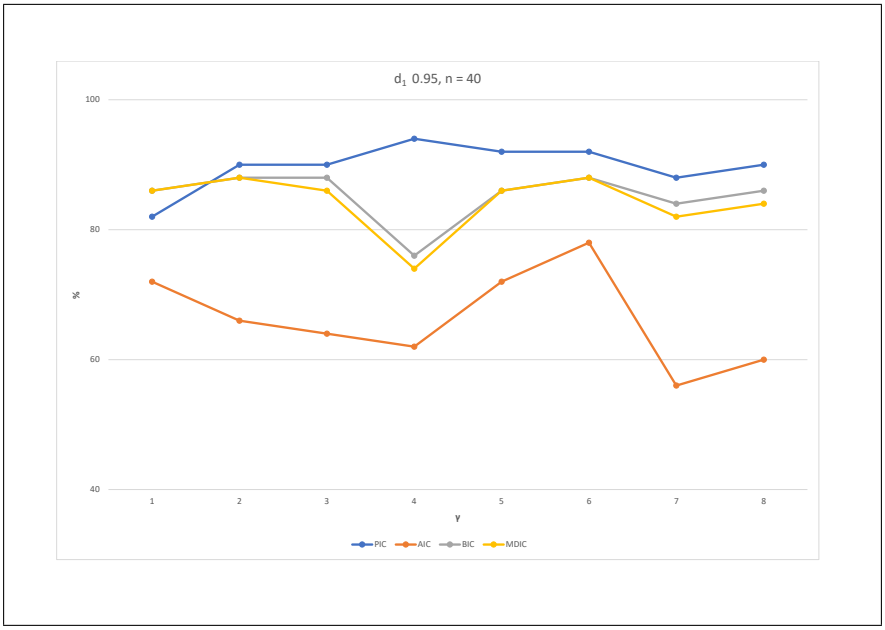

Figure 3.3: Proportion of the correct model selected ($n = 40$, 5% contamination).

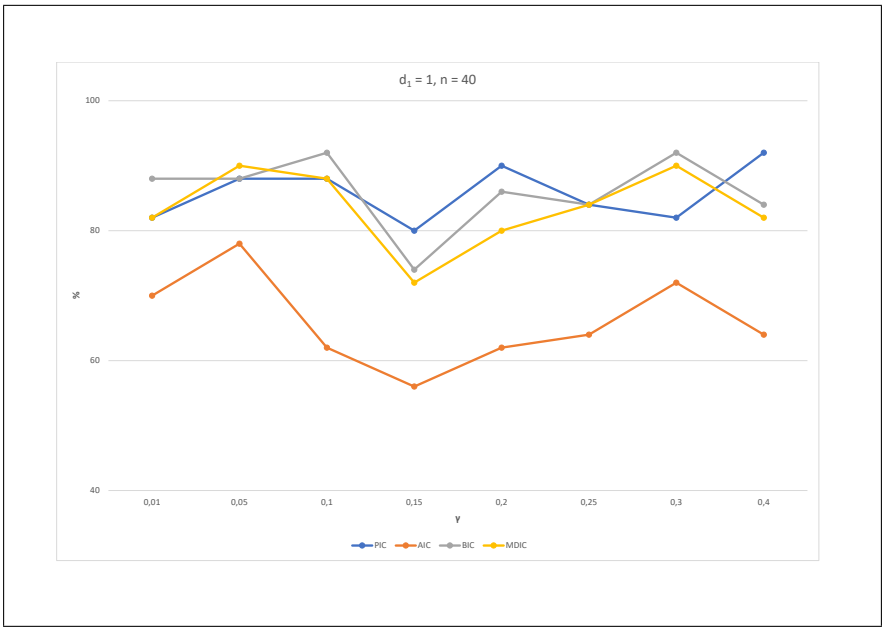

Figure 3.4: Proportion of the correct model selected ($n = 40$, 0% contamination).

3.3 Real Case Study

3.3.1 Cement Hardening Data

This example is a small set of data on variables thought to be related to the heat evolved during the hardening of Portland cement [30]. This data set has been used by various authors (see [10], [12]) and represents a popular example for multiple linear regression analysis.

Table 3.33: Cement hardening data

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	**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

The data include 4 predictor variables with a sample size of 13. The predictor variables (as a percentage of the weight) are:

- X_1 = calcium aluminate ($3CaO \cdot Al_2O_3$)
- X_2 = tricalcium silicate ($3CaO \cdot SiO_2$)
- X_3 = tetracalcium alumino ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$)
- X_4 = dicalcium silicate ($2CaO \cdot SiO_2$,)

while the response variable is

- Y = total calories given off during hardening per gram of cement after 180 days.

3.3.2 Set of Candidate Models and Model Selection Criteria

Because there are 4 variables available, we considered all $(2^4 - 1 = 15)$ 15 possible candidate models, involving at least one regressor for this dataset.

The model selection criteria that have been used are PIC [29], AIC [2], BIC [26] and MDIC [20]. In addition, because of the small sample size we also added AICc [15] in the study.

3.3.3 Results

The following table shows which model was selected by PIC, AIC, AICc, BIC and MDIC.

Observe that PIC chooses different models as the value of γ changes. We note, however, that the 4 single-variable models should be excluded from the analysis because cement involves a mixture of *at least two* compounds that react chemically (see [10]).

- PIC for $\gamma = 0.3$ chooses $X_1X_2X_4$ model as AIC.
- PIC for $\gamma = 0.4$ chooses $X_1X_2X_3$ model like MDIC.
- AICc and BIC both choose X_1X_2 model.

Table 3.34: Selected models by model selection criteria

Criteria	γ	**Variables**	**Criteria**	**Variables**
PIC	0.01	X_1	**AIC**	$X_1X_2X_4$
			AICc	X_1X_2
	0.05-0.25	X_2		
			BIC	X_1X_2
	0.3	$X_1X_2X_4$	**MDIC**	$X_1X_2X_3$
	0.4	$X_1X_2X_3$		

We observe that PIC behaves similarly to AIC and MDIC having a slight tendency of overestimation. Note though that for the specific dataset the collinearity is quite strong with X_1 and X_3 as well as X_2 and X_4 being seriously correlated.

3.4 Conclusions

In this work we performed a comparative study of model selection criteria in order to investigate the practical implications of the new criterion PIC.

The simulation study showed that for $d_1 = 0.9$ and $d_1 = 0.95$ PIC has the highest success rate for small sample sizes. For medium sample sizes PIC is comparable to BIC, while for large sample sizes PIC ranked second and BIC ranked first.

For the case where we don't have contamination and $d_1 = 1$ for small sample sizes PIC has the highest success rate while for medium and large sample sizes PIC ranked second and BIC was first.

For the case where $d_1 = 0.8$ PIC has the highest success rate for small sample sizes, while the second one was MDIC. For medium sample sizes PIC is comparable to BIC. Finally, PIC ranked second for large sample sizes and BIC was first.

The real case study showed that PIC gives different results for the different values of γ, but it has the same behaviour as AIC for $\gamma = 0.3$ and the same behaviour as MDIC

for $\gamma = 0.4$.

Based on the results of the simulation study we conclude that the performance of PIC is satisfactory for all possible settings according to the sample size, underlying error distribution and contamination rate. An important issue that needs further investigation is the choice of the appropriate value for the tunning parameter γ.

We hope to address and provide a satisfactory solution to this problem in a future work. We also intent to explore the capabilities of PIC in various other settings.

Bibliography

[1] Acquah, H. D. (2013). On the Comparison of Akaike Information Criterion and Consistent Akaike Information Criterion in Selection of an Asymmetric Price Relationship: Bootstrap Simulation Results. *Agris On-line Papers in Economics and Informatics*, **5(1)**, 3–9.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proc. of the* 2^{nd} *Intern. Symposium on Information Theory*, (Petrov B. N. and Csaki F., eds.), 267-281, Akademiai Kaido, Budapest.

[3] Anderson, D. R., Burnham, K. P., White, G. C. (1998). Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics*, **25:2**, 263–282.

[4] Basu, A., Harris, I. R., Hjort, N. L., Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.

[5] Boltzmann, L. (1877). Uber die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Warmetheorie und der Wahrscheinlichkeitsrechnung respektive den Satzen uber das Warmgleichgewicht. Weiner Berichte. **76**, 373–435.

[6] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52:3**, 345–370.

[7] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

[8] Brewer, M. J., Butler, A., Cooksley, S. L. (2016). The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, **7**, 679-692.

[9] Broniatowski, M., Toma, A., Vajda, I. (2012). Decomposable pseudodistances and applications in statistical estimation. *Journal of Statistical Planning and Inference*, **142:9**, 2574–2585.

[10] Burnham, K. P., Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. New York, Springer.

[11] Claeskens, G., Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge University Press.

[12] Draper, N. R., Smith, H. (1981). Applied Regression Analysis. 2nd Edition, Wiley, New York.

[13] Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R. (2017). Sensitivity and Specificity of Information Criteria. PeerJ Preprints 5:e1103v3.

[14] Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, **136**, 210−271.

[15] Hurvich, C. M., Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

[16] Jones, M. C., Hjort, N. L., Harris, I. R., Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika*, **88**, 865–873.

[17] Konishi, S., Kitagawa, G. (2008). Information criteria and statistical modeling. Springer.

[18] Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

[19] Lee, S., Karagrigoriou, A. (2001). An asymptotically optimal selection of the order of a linear process. Sankhya A, **63**, 93-106.

[20] Mantalos, P., Mattheou, K., Karagrigoriou, A. (2010). An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics – Simulation and Computation*, 39, 865–879.

[21] Mattheou, K., Lee, S., Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, 139, 128–135.

[22] Mattheou, K. (2007). On New Developments in Statistical Inference for Measures of Divergence, Ph.D. thesis, Department of Mathematics and Statistics, University of Cyprus.

[23] Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. Ann. Inst. Statist. Math., **19**, 181−192.

[24] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 157−175.

[25] Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). Akaike Information Criterion Statistics. (Tokyo, KTK Scientific).

[26] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statist.*, **6**, 461–464.

[27] Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

[28] Toma, A., Fulga, C. (2018). Robust Estimation for the Single Index Model Using Pseudodistances. *Entropy*, 374-394.

[29] Toma, A., Karagrigoriou, A., Trentou, P. (2019). Robust criteria for model selection using pseudodistances. (In preparation).

[30] Woods, H., Steinour, H. H., and Starke, H. R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry*, 24, 1207–1214.