JIRSS (2021)
Vol. 20, No. 02, pp 43-63
DOI:10.52547/jirss.20.2.43

Testing Several Rival Models Using the Extension of Vuong's Test and Quasi Clustering

Abdolreza Sayyareh [1]

[1] Department of Computer Sciences and Statistics, K. N. Toosi University of Technology, Tehran-Iran.

Received: 11/07/2019, Revision received: 10/12/2021, Published online: 09/02/2022

Abstract. The two main goals in model selection are firstly introducing an approach to test homogeneity of several rival models and secondly selecting a set of reasonable models or estimating the best rival model to the true one. In this paper we extend Vuong's method for several models to cluster them. Based on the working paper of Katayama (2008), we propose an approach to test whether rival models have expected relations. The multivariate extension of Vuong's test gives the opportunity to examine some hypotheses about the rival models and their relations with respect to the unknown true model. On the other hand, the standard method of model selection provides an implementation of Occam's razor, in which parsimony or simplicity is balanced against goodness of fit. Therefore, we are interested in clustering the rival models based on their divergence from the true model to select a suitable set of rival models. In this paper we have introduced two approaches to select suitable sets of rival models based on the multivariate extension of Vuong's test and quasi clustering approach.

Keywords. Akaike Information Criterion, Clustering, Kullback-Leibler Divergence, Mis-specified Models, Non-nested Models.

Corresponding Author: Abdolreza Sayyareh (asayyareh@kntu.ac.ir)

MSC: 62F03, 62H30.

1 Introduction

Vuong (1989) has presented a test based on the likelihood ratio to compare two proposed models. Katayama (2008) has provided an extension of Vuong's model selection test for the multivariate case. Lorestani and Sayyareh (2017) have extended the Vuong's (1989) model selection test to three models in accordance to the union-intersection principle. They have shown that the distribution of the test statistic is asymptotically equal to the distribution of the maximum of dependent random variables with bivariate folded standard normal distribution. As a part of the Vuong's test, he demonstrated that two competing models are equally close to the true data generating model. Consider the simplest case where we have two non-nested rival moedels, denoted as $\mathcal{M}_1, \mathcal{M}_2$.

We may test the equivalence of the two rival models or test wether one model is better than the other one. The selected set of reasonable models will be either $\{\mathcal{M}_1\}$,$\{\mathcal{M}_2\}$ or $\{\mathcal{M}_1, \mathcal{M}_2\}$. The standardized differences of the expected Kullbal-Leibler, $\mathcal{EKL}$'s, of the two rival models will be used as the criterion to select between the rival models. See Vuong (1989) and Commenges et al. (2008). We have developed analogous procedures based on the Katayama's extention to more general null hypothesis testing problem. It is not difficult to extend Linhart's (1988) test to the case where we have $\mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_w$ as non-nested rival models, see, Shimodaira (1997). For each $\mathcal{M}_i, i = 1, 2, ..., w$, we will decide whether the rival model $\mathcal{M}_i$ is equivalent to the model $\mathcal{M}_j, i \neq j = 1, 2, ..., w$ or not. If we find two equivalent rival models, we will set these models at the same subset of models and we will continue our search to achieve k sets of equivalent models in $\mathcal{KL}$ sense.

In the literature there are many criteria to evaluate the best model. The well known criterion is Kullback-Leibler (1951), say $\mathcal{KL}$, risk or divergence. This criterion has an estimator as Akaike (1973) information criterion, AIC. Commenges et al. (2008) have recently considered the differences of the $\mathcal{KL}$ risks between two rival models and Sayyareh et al. (2011) and Sayyareh (2012) compare some tests and criteria to model selection.

This paper examines a common scenario in which there is more than one candidate model. The literature on non-nested hypothesis testing in statistics was pioneered by Cox (1961, 1962) and Atkinson (1970), this which was applied by Pesaran (1974) and Pesaran and Deaton (1978). Shimodaira (1998, 2001) has considered the sampling error

of *AIC* in multiple comparisons and has constructed a set of good models rather than choosing a single model. The asymptotic distribution of *AIC* in linear regression models and the bias correction of this statistics are discussed by Yanagihara and Ohomoto (2005). Recently, Commenges et al. (2008) has considered the normalized difference of *AIC* as an estimate of a difference of Kullback-Leibler risks between two models. Cox (1961, 1962) has modified the classical hypothesis testing to test the non-nested hypotheses, Vuong (1989) tested the equivalence of two models, and the information criterion (*AIC*), Akaike (1973), is introduced to select the best model under parsimony. An essential problem in model selection arises from the phenomenon that Zucchini (2000) refers to as selection bias. If one begins with a very large collection of rival models, then he can be sure that the reasonable model will have accidentally high maximum likelihood term. The selection bias can be expected to be less if we begin with a small set of rival models. But we faced with a problem: how to select the few models that go into this set? After selecting the admissible set of rival models, the problem is simplified to finding the best model in this set. One problem with *AIC* is that its value has no intrinsic meaning and values of *AIC* depend on the number of observations. If the specific structure of the models is of interest, it may be interesting to measure how far from the truth each model is. This may not be possible, but we can quantify the difference of risks between two models. Estimating the difference of risks will be informative only if we have an idea of what a large or a small difference is. It is shown that a normalized difference of *AIC*'s is an estimate of a difference of Kullback-Leibler ($\mathcal{KL}$) criterion,(Kullback-Leibler (1968)), see Commenges et al. (2008). Some recent investigations have made to extend the model selection tests and criteria.

In this paper we have considered a subset model selection approach which is an open problem in model selection. The subset selection leads to the consideration of multivariate extension of Vuong's test which gives the opportunity to examine some hypotheses about the rival models and their relations with respect to the unknown true model. Therefore, we are interested in clustering the rival models based on its divergence from the true model to select a suitable set of rival models. We have considered an approach to test whether competing models have expected relations. We have extended Vuong's (1989) results to various cases in non-nested situations. In many situations, we have a few rival models. For each model we could compute the mean corrected losses which is equal to $1/(2n)$AIC of each rival model. Katayama (2008), has considered a deep mathematical and asymptotic study on the extension of Vuong's (1989) model selection test. As a part of this work Katayama considered the equality of all rival models. In this work we have considered a statistical test which is a little different from Katayama's work, so that we could discuss many models

and different tests. Also, we have proposed an approach which helps us answer this unsolved problem in statistics: How can we select an admissible set of models which are more reasonable to consider as a rival set of models? This set of rival models leads us to decrease the risk in model selection. This approach lets us consider a large set of models as the candidate set and return out some of them because of their large divergence from the true model.

A few articles have examined the selection of an admissible set of rival models. Barmalzan and Sayyareh (2011) have supposed a random sample of a population with true but unknown density $h(.)$. In general, the true density is unknown and we have to consider a parametric model, say, $f(.;\theta)$ as an approximation of $h(.)$. Clearly, $f(.;\theta)$ should be close to the true density. The suggestion of a model as an approximation or estimation of the true density might result in a great risk in model selection. For this reason, they consider k nonnested rival models and investigate the model which is closer to the true model. In fact, they have considered this main question in model selection that how it is possible to gain a collection of approximate models for the estimation of the true model. Pho et al. (2019) have compared Akaike information criterion, Bayesian information criterion and Vuong's test in model selection. Sayyareh (2017) has considered the sample and the non-nested rival models as blocks and treatments, respectively, and introduce the extended Friedman test version to compare with the results of the test based on the linear sign rank test. Clarke and Signorino (2010) have considered the problem of choosing rival statistical models that are non-nested in terms of their functional forms. They assess the ability of two tests, one parametric and one distribution free, to discriminate between such models.

The rest of this paper is organized as follow: Section 2 presents the basic framework, models and assumptions. Section 3 addresses the Kullback-Leibler risks. Section 4, which describes the normalized difference of *AIC's* as an estimate of the difference of Kullback-Leibler risk. In Section 5, we illustrate the extension of Vuong's test. In Section 6, we propose a cluster approach to construct a suitable set of models which are close to the true one. We also present a simulation study in the framework of densities which makes it possible to illustrate our approach.

2 Basic Framework

Suppose the focus of the analysis is to be used instead of the $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, ...\mathcal{M}_w\}$ as w parametric rival models. They are candidates to use instead the data generating model (the true model). If the specific structure of the models is of interest, as it tell us

something about the observed phenomena, it would be important to measure how far from the truth each model is. The models are considered as the conditional models, which are all based on the same conditioning variables, x_t, and differ only insofar as they are based upon different p.d.fs. According to Vuong (1989) and Katayama (2008) we consider, following assumptions:

Assumption A1 (*a*) The p-dimensional random vectors $W_t' = (Y_t', X_t'), t = 1, 2, ...$ are independent and identically distributed (*i.i.d.*) with common true distribution $H_W^0(W, \sigma)$.

(*b*) For H_X-almost all x, $H_{Y|X}(y|x)$ has a Radon-Nikodym density $h^0(y|x)$ relative to v_Y, which is strictly positive for v_Y-almost all y.

Considering w parametric families of conditional distributions defined on $\sigma_Y \times X$ for Y_t given X_t: $\mathcal{M}_i = \left\{g_i^{\beta_i}(y|x), \beta_i \in \mathcal{B}_i \subset \Re^{p_i}\right\} = (g_i)_{\beta_i \in \mathcal{B}_i}, i = 1, 2, ...w$, where p_is are positive integers such that $p_i \leq p_j, i < j$ and β_i is the parametric space for model $\mathcal{M}_i$.

Assumption A2 1. For a every $\beta_i \in \mathcal{B}_i$ and for H_X-almost all x, the conditional distribution $G_{Y|X}^i$ has a Radon-Nikodym density $g_i^{\beta_i}$ relative to v_Y, which is strictly positive for V_y-almost all y. (*b*) $\mathcal{B}_i$ is a compact subset of $\Re^{p_i}$, and the conditional density $g_i^{\beta_i}(y|x)$ is continuous in β_i for H_W-almost all (y, x).

2. For H_W-almost all (y, x), $|\log g_i^{\beta_i}(y|x)|$ is dominated by an H_W-integrable function independent of $\beta_i \in \mathcal{B}_i$. 2.(*b*) The function $x_g = \int \log g_i^{\beta_i}(y|x) H_W^0(dx)$ has a unique maximum on $\beta_i^* \in \mathcal{B}_i$.

3. For H_W-almost all (y, x), $|\log g_i^{\beta_i}(y|x)|$ is twice continuously differentiable on $\mathcal{B}_i$. (*b*) For H_W-almost all (y, x), $|\nabla_{\beta_i} \log g_i^{\beta_i}(y|x)$. $\nabla_{\beta_j'} \log g_j^{\beta_j}(y|x)|, |\nabla^2_{\beta_i \beta_i'} \log g_i^{\beta_i}(y|x)|$ and $|\log g_i^{\beta_i}(y|x)|$ are dominated by H_W-integrable functions independent of $\beta_i \in \mathcal{B}_i$ and $\beta_j \in \mathcal{B}_j$.

4.(*a*) β_{i*} is an interior point of $\mathcal{B}_i$.

(*b*) β_{i*} is a regular point of $A_i(\beta_i)$, where, $A_i(\beta_i) = \mathcal{E}_h\left[\nabla^2_{\beta_i \beta_i'} \log g_i^{\beta_i}(Y|x)\right]$ and $\mathcal{E}_h$ denotes the expectation with respect to H_W.

Definition 2.1. The parametric family $(i)\,(g_i)_{\beta_i \in \beta_i}$ is nested in $(g_j)_{\beta_j \in \beta_j}$ if $(g_i) \subset (g_j)$; $(ii)\,(g_i)$ is well specified if there is value $\beta_0 \in \mathcal{B}_i$ such that $g_i^{\beta_i}(y|x) = h(y|x)$; otherwise it is mis-specified.

3 Kullback-Leibler Risk and Likelihood Function

In decision theory, estimators are chosen as minimizing some risk function. The most important risk functions are based on the Kullback-Leibler, (1951), $\mathcal{KL}$, divergence. Let a probability $\mathbf{P}'$ be absolutely continuous with respect to a probability $\mathbf{P}$, $\mathcal{F}_1$ be a sub-σ-field of $\mathcal{F}$, the loss of using $\mathbf{P}'$ instead of $\mathbf{P}$ is the $\mathcal{L}_{\mathcal{F}}^{\mathbf{P}/\mathbf{P}'} = \log \frac{d\mathbf{P}}{d\mathbf{P}'|\mathcal{F}}$. Its expectation is $\mathcal{E}_{\mathbf{P}}\{\mathcal{L}_{\mathcal{F}}^{\mathbf{P}/\mathbf{P}'}\} = \mathcal{KL}(\mathbf{P}, \mathbf{P}'; \mathcal{F})$. This is the Kullback-Leibler risk. If $\mathcal{F}$ is the largest sigma-field on the space, then we omit it in the notation. If Y is random variable with p.d.f. f_Y and g_Y under $\mathbf{P}$ and $\mathbf{P}'$, respectively, then we $\frac{d\mathbf{P}}{d\mathbf{P}'|\mathcal{F}} = \frac{h^0(y|x)}{g_i^{\beta_i}(y|x)}$ and the divergence of the distribution $\mathbf{P}'$ relative to $\mathbf{P}$ can be written as $\mathcal{KL}(P, P') = \int \log \frac{h(y|x)}{g_i^{\beta_i}(y|x)}\, d(y)$. We know that $\mathcal{KL}(\mathbf{P}, \mathbf{P}'; \mathcal{F}) = \mathcal{KL}(P, P')$ if $\mathcal{F}$ is the σ-field generated by y on $(\Omega, \mathcal{F})$. Base on continuity arguments, we take $0 \log \frac{0}{r} = 0$ for all $r \in R$ and $t \log \frac{t}{0} = \infty$ for all non-zero t. Hence, $\mathcal{KL}$ divergence takes its value in $[0, \infty]$ and $\mathcal{KL}(h(y|x), g_i^{\beta_i}(y|x)) = 0$ implies that $h(y|x) = g_i^{\beta_i}(y|x)$. The $\mathcal{KL}$ divergence is not a metric, but it is additive over marginals of product measures. We assume that there is a value $\beta_{i*} \in \mathcal{B}_i$ which minimizes $\mathcal{KL}(h(.), g_i^{\beta_i}(y|x))$. If the model is well specified $\beta_{i0} = \beta_{i*}$. The MLE $\hat{\beta}_{in}$ is a consistent estimator for β_*. If the model is well specified $\beta_* = \beta_0$; if the model is misspecified, $\mathcal{KL}(h, g^{\beta}) > 0$. The Quasi Maximum Likelihood Estimator (QMLE), $\hat{\beta}_n$, is a consistent estimator of β_*, see White (1982^a, 1982^b).

4 Differences of AIC Criteria

Consider a sample of independently distributed random variables $\underline{Y}_n = (Y_1, ..., Y_n)$ having probability distribution function, pdf, $h = h(\cdot)$. Let us consider k rival models: $(g_i) = (g_i^{\beta_j}(.))_{\beta_j \in B_j}, B_j \subset \Re^{p_j}, i = 1, ..., k$ and $j = 1, ..., J_i$.

Definition 4.1. (i) (g_i) is nested in (g_j) if $(g_i) \subset (g_j)$; (ii) (g_j) is well specified if there is a value $\beta_{\star_j} \in B_j$ such that $g^{\beta_{\star_j}} = h$; otherwise it is misspecified.

Kullback-Leibler divergence, is the log-likelihood loss of the rival model g^{β_j} relatively to h for observation Y, is the expectation of $\log \frac{h(Y)}{g^{\beta_j}(Y)}$; officially defined as

$$\mathcal{KL}(h, g^{\beta_j}) = \mathcal{E}_h\Big\{\log \frac{h(Y)}{g^{\beta_j}(Y)}\Big\}.$$

This criterion is nonnegative and is zero when $h = g^{\beta_j}$, that is, $\beta_j = \beta_{\star_j}$. The Kullback-Leibler divergence is not a distance between the two probability measures, because it is not symmetric, but generally it is not a drawback: there is no symmetry between h, the true pdf and g^{β_j} a possible pdf. This indicates that we may think about the Kullback-Leibler divergence as a an expected loss rather than a distance. We also assume that there is a $\beta_0 \in B$ which minimize $\mathcal{KL}(h, g^{\beta_j})$. If the model is well specified, $\beta_0 = \beta_*$; if not, $\mathcal{KL}(h, g^{\beta_{\star j}}) > 0$. White (1982) has shown that the MLE $\hat{\beta}_n$ is a consistent estimator of β_* and β_0. We shall say (g_i) is closer to h than (g_j) if $\mathcal{KL}(h, g^{\beta_i}) < \mathcal{KL}(h, g^{\beta_j})$. Considering $\mathcal{KL}(h, g^{\beta_j}) = \mathcal{E}_h\Big\{\log h(Y)\Big\} - \mathcal{E}_h\Big\{\log g^{\beta_j}(Y)\Big\}$, we can estimate the difference of risks $\triangle(g^{\beta_{0_j}}, g^{\beta_{0_i}}) = \mathcal{KL}(h, g^{\beta_{0_j}}) - \mathcal{KL}(h, g^{\beta_{0_i}})$ by $-\frac{1}{n}(L_{\underline{Y}_n}^{g^{\hat{\beta}_{jn}}} - L_{\underline{Y}_n}^{g^{\hat{\beta}_{in}}})$, where L denotes the log-likelihood function. As an estimation, we use $g^{\hat{\beta}_{jn}}$ instead of $g^{\beta_{0_j}}$. Thus we consider $\mathcal{E}_h\Big\{\log \frac{h(Y)}{g^{\hat{\beta}_{jn}}(Y)}\Big\}$, the expected Kullback-Leibler loss, and that we denote by $\mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}})$, which is introduced by Akaike (1973). Linhart and Zucchini (1986) have show that

$$\mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}}) = \mathcal{KL}(h, g^{\beta_j}) + \frac{1}{2}n^{-1}Tr(I_{g_j}^{-1}J_{g_j}) + o(n^{-1}),$$

where

$$I_{g_j} = -\mathcal{E}_h\Big\{\frac{\partial^2 \log g^{\beta_j}(Y)}{\partial\beta^2} \mid_{\beta_0}\Big\},$$

and

$$J_{g_j} = \mathcal{E}_h\Big\{\Big[\frac{\partial \log g^{\beta_j}(Y)}{\partial\beta} \mid_{\beta_0}\Big]\Big[\frac{\partial \log g^{\beta_j}(Y)}{\partial\beta} \mid_{\beta_0}\Big]^T\Big\}.$$

Two essential terms in $\mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}})$ are interpreted as the misspecification risk and the statistical risk, respectively. We also have

$$\begin{aligned}&\mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}})\\ &= -\mathcal{E}_h(n^{-1}L_{\underline{Y}_n}^{g^{\hat{\beta}_{jn}}}) + \mathcal{E}_h\Big\{\log h(Y)\Big\} + n^{-1}Tr(I_{g_j}^{-1}J_{g_j}) + o_p(n^{-1}).\end{aligned} \tag{4.1}$$

Akaike criterion as $AIC(g^{\hat{\beta}_{jn}}) = -2L_{\underline{Y}_n}^{g^{\hat{\beta}_{jn}}} + 2p$, follows from (4.1) as $\mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}})$. Our interest is $\triangle(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}}) = \mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{jn}}) - \mathcal{E}_h\mathcal{KL}(h, g^{\hat{\beta}_{in}})$. Using (4.1), we obtain

$$\begin{aligned}&\mathcal{E}_h\Big\{-\frac{1}{n}\Big[(L_{\underline{Y}_n}^{g^{\hat{\beta}_{jn}}} - L_{\underline{Y}_n}^{g^{\hat{\beta}_{in}}}) - [Tr(I_{g_j}^{-1}J_{g_j}) - Tr(I_{g_i}^{-1}J_{g_i})]\Big]\Big\}\\ &= \triangle(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}}) + o_p(n^{-1}).\end{aligned} \tag{4.2}$$

Akaike in his suggestion has noted that if $Tr(I_{g_j}^{-1}J_{g_j}) \approx p_j$, then $Tr(I_{g_i}^{-1}J_{g_j}) \approx p_i$. Using these approximations, we obtain a simple estimator of $\triangle(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$ as

$$\begin{aligned}\hat{\triangle}(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}}) &= \frac{1}{2}n^{-1}\Big\{AIC(g^{\hat{\beta}_{nj}}) - AIC(g^{\hat{\beta}_{ni}})\Big\} \\ &= -n^{-1}\Big\{L_{\underline{Y}_n}^{g^{\hat{\beta}_{jn}}} - L_{\underline{Y}_n}^{g^{\hat{\beta}_{in}}} - (p_j - p_i)\Big\}. \end{aligned} \tag{4.3}$$

Note that the precise value of *AIC* has no clear interpretation, but the expectation of $\hat{\triangle}(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$ tracks the quantity of main interest $\triangle(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$. In fact, the bias of $\hat{\triangle}(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$ is of order $o_p(n^{-1})$. All that has been said can be extended to regression models by directly defining the $\mathcal{KL}$ divergence in terms of conditional densities.

5 Extension of Vuong's Test

From now on, we use conditional density, $g_i^{\beta_i}(Y|\cdot)$, based on the Vuong's paper. Consider the case where we have k possibly non-nested models to be compared. Sometimes, we only need to say that some models are equivalent as a candidate for the true model. This equivalence in $\mathcal{KL}$ sense means that

$$\mathcal{KL}\,(h, g_i^{\beta_i}) = \mathcal{KL}\,(h, g_j^{\beta_j}),$$

or

$$\mathcal{E}_h\,(\log g_i^{\beta_i}(Y|x)) = \mathcal{E}_h\,(\log g_j^{\beta_j}(Y|x)).$$

Let $g_j^{\beta_j}(\cdot|\cdot) \neq g_i^{\beta_i}(\cdot|\cdot)$ and define the maximum log-likelihood function for model (g) as

$$LL_n(i) = \sum_{t=1}^{n} \log g_i^{\hat{\beta}_{in}}(Y_t|\cdot).$$

In general, we say that (g_i) is closer to h than (g_j) if $\mathcal{KL}\,(h, g_i^{\beta_{i*}}) < \mathcal{KL}\,(h, g_j^{\beta_{j*}})$. We cannot estimate $\mathcal{KL}\,(h, g_i^{\beta_i})$ because the entropy of h, which is equal to $\mathcal{E}_h(\log h(\cdot))$, cannot be correctly estimated. Whereas, the second term of the $\mathcal{KL}$ risk has a known estimation as Akaike information criterion as *AIC*, which is given by

$$AIC(j) = -2LL_n(j) + 2p_j,$$

where $p_j = \dim \beta_j$. Then it is more relevant to consider the risk $\mathcal{E}\{\log \frac{h(Y|\cdot)}{g_j^{\hat{\beta}_{jn}}(Y_t|\cdot)}\}$ that we may call the expected Kullback-Leibler risk and that we denote it by $\mathcal{EKL}\,(h, g_i^{\hat{\beta}_{in}})$. *AIC* was revisited by Linhart and Zucchini (1986) who showed that

$$\mathcal{EKL}\,(h, g_i^{\hat{\beta}_{in}}) = \mathcal{EKL}\,(h, g^{\beta_{i*}}) + \frac{1}{2n} Tr(B_{g_i} A_{g_i}^{-1}) + o(n^{-1}), \tag{5.1}$$

where

$$B_{g_i} = \mathcal{E}_h\{[\frac{\partial \log g_i^{\beta}(Y|\cdot)}{\partial \beta}].[\frac{\partial \log g_i^{\beta}(Y|\cdot)}{\partial \beta'}]|_{\beta_{i*}}\},$$

and

$$A_{g_i} = -\{\mathcal{E}_h[\frac{\partial^2 \log g_i^{\beta}(Y|\cdot)}{\partial \beta \partial \beta'}|_{\beta_{i*}}]\}.$$

We also have

$$\mathcal{EKL}\,(h, g_i^{\hat{\beta}_{in}}) = F(h) - \mathcal{E}_h\{n^{-1} LL_n(i)\} + \frac{1}{n} Tr(B_{g_i} A_{g_i}^{-1}) + o_p(n^{-1}). \tag{5.2}$$

Akaike information criterion follows from (5.2) by multiplying by $2n$, deleting constant term $F(h)$, replacing second term in the right by $n^{-1} LL_n(i)$ and replacing $Tr(B_{g_i} A_{g_i}^{-1})$ by p_i. The term $\frac{1}{n} Tr(B_{g_i} A_{g_i}^{-1})$ is the sum of the mis-specification risk and the statistical risk. Note that if (g_i) is well-specified, the mis-specification risk is zero and $B_g = A_g$, and thus, $\mathcal{EKL}\,(h, g_i^{\hat{\beta}_{in}}) = p_i/2n + o(n^{-1})$. Define

$$LL_{wn} = (LL_n(1), LL_n(2), ..., LL_n(w)),$$

$$LL_*(i) = \sum_{t=1}^{n} \log g_i^{\beta_{i*}}(Y_t|\cdot),$$

$$\hat{\beta}_n = (\hat{\beta}'_{1n}, \hat{\beta}'_{2n}, ..., \hat{\beta}'_{wn}),$$

and

$$\beta_* = (\beta'_{1*}, \hat{\beta}'_{2*}, ..., \hat{\beta}'_{w*}).$$

In misspecified case, $\hat{\beta}_{in}$ is referred to as quasi maximum likelihood estimator, *QMLE*, and its probability limit under the true model, which we denote by β_{i0}, is known as pseudo true value of parameter. These pseudo true values are defined by

$$\beta_{i*} = \arg\max_{\beta_i \in \mathcal{B}_i} \mathcal{E}_h\{\frac{1}{n} \sum_{t=1}^{n} \log g_i^{\beta_i}(Y_t|\cdot)\}.$$

To ensure global identifiability of the pseudo true value, it will be assumed that β_{i*} provides the unique maxima of $\mathcal{E}_{g_i}\{\frac{1}{n}\sum_{t=1}^{n}\log g_i^{\beta_i}(Y_t|\cdot)\}$. We prepare the notation as: $\mu_w = E_h\{\log g_i^{\beta_{i*}}(Y_t|\cdot)\}$, $E_{*w} = (\mu_1, \mu_2, ..., \mu_w)$. For more detail about the following theorem, see Katayama (2008).

Theorem 5.1. *Under Assumptions A1 and A2 the vector*

$$Z_n = \sqrt{n}\left((\hat{\beta}_n - \beta_*)', (\frac{1}{n}LL_{wn} - E_{*w})'\right)',$$

converges to a standard normal distribution:

$$Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta L} \\ \Sigma'_{\beta L} & \Sigma_{LL} \end{pmatrix},$$

$$\Sigma_{\beta\beta} = \left(A_i^{-1}(\beta_{i*})C_{ij}(\beta_{i*}, \beta_{j*})A_j^{-1}(\beta_{j*})\right),$$

$$B_{ij}(\beta_i, \beta_j) = E\left[\nabla_{\beta_i}\log g_i^{\beta_i}(Y_t|\cdot)\nabla_{\beta'_j}\log g_j^{\beta_j}(Y_t|\cdot)\right],$$

$$A_i(\beta_i) = E\left[\nabla^2_{\beta_i\beta'_i}\log g_i^{\beta_i}(Y_t|\cdot)\right],$$

$$C_{ij}(\beta_i, \beta_j) = A_i^{-1}(\beta_i)B_{ij}(\beta_i, \beta_j)A_j^{-1}(\beta_j),$$

$$\Sigma_{\beta L} = \left(Cov(-A_i^{-1}(\beta_{i*})(\frac{\partial \log g_i^{\beta_{i*}}(Y_t|\cdot)}{\partial\beta_{i*}}), \log g_{j+1}^{\beta_{j+1*}}(Y_t|\cdot))\right),$$

$$\Sigma_{LL} = \left(Cov(\log g_i^{\beta_{i*}}(Y_t|\cdot), \log g_j^{\beta_{j*}}(Y_t|\cdot))\right).$$

Proof. From the fact that,

$$n^{1/2}(\hat{\beta}_{in} - \beta_{i*}) \xrightarrow{L} \mathcal{N}(0, I_{g_j}^{-1}J_{g_j}),$$

we have $n^{1/2}(\hat{\beta}_{in} - \beta_{i*})$ and $-A_i^{-1}(\beta_{i*})n^{-1/2}\nabla_{\beta_{i*}}(LL_*(j))$ are asymptotically equivalent:

$$\frac{n^{1/2}(\hat{\beta}_{in} - \beta_{i*})}{-A_i^{-1}(\beta_{i*})n^{-1/2}\nabla_{\beta_{i*}}(LL_*(j))} \to 1.$$

Set,

$$AL_*(i) = (-A_1^{-1}(\beta_{1*})\nabla_{\beta_{1*}} \log g_1^{\beta_{1*}}(Y_t|\cdot), -A_2^{-1}(\beta_{2*})\nabla_{\beta_{2*}} \log g_2^{\beta_{2*}}(Y_t|\cdot), ...,$$
$$- A_w^{-1}(\beta_{w*})\nabla_{\beta_{w*}} \log g_w^{\beta_{w*}}(Y_t|\cdot))',$$

and

$$LE_*(i) = \left(\log g_1^{\beta_{1*}}(Y_t|\cdot) - \mu_1, \log g_2^{\beta_{2*}} - \mu_2, ..., \log g_w^{\beta_{w*}}(Y_t|\cdot) - \mu_w\right)'.$$

From Vuong's approach we have:

$$Z_n = n^{-1/2} \sum_{i=1}^{n} \{AL_*(i), LE_*(i)\}'$$
$$= \sqrt{n}\left((\hat{\beta}_n - \beta_*)', (\frac{1}{n}LL_{wn} - E_{*w})'\right)' \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

□

5.1 Estimating $A_i(\beta_i), B_{ij}(\beta_i, \beta_j)$ and Σ

Related to the Theorem 5.1. we need to estimate $A_i(\beta_i), B_{ij}(\beta_i, \beta_j)$ and Σ. The strong consistent estimators of $A_i(\beta_i), B_{ij}(\beta_i, \beta_j)$ are given by

$$A_i(\hat{\beta}_{in}) = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)}{\partial\beta\partial\beta'},$$

$$B_{ij}(\hat{\beta}_{in}, \hat{\beta}_{jn}) = \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)}{\partial\beta} \frac{\partial \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)}{\partial\beta'}.$$

Also, Σ is obtained from the sample analogs of the submatrics of Σ. Then, $\hat{\Sigma}_{\beta\beta} = \hat{C}_{ij}(\hat{\beta}_i, \hat{\beta}_j)$ which is the empirical version of the $C_{ij}(\beta_i, \beta_j)$ evaluated at maximum likelihood estimate of β. Similarly, (i, j)-th submatrics of $\Sigma_{\beta L}$ and Σ_{LL} are respectively

$$\hat{\sigma}_{\beta L}^{ij} = -\frac{1}{n} A_i^{-1}(\hat{\beta}_{in}) \sum_{t=1}^{n} \frac{\partial \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)}{\partial\beta} \log g_j^{\hat{\beta}_{jn}}(Y_t|\cdot)$$
$$+ \left\{ A_i^{-1}(\hat{\beta}_{in}) \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)}{\partial\beta} \right\} \left\{ \frac{1}{n} \sum_{t=1}^{n} \log g_j^{\hat{\beta}_j n}(Y_t|\cdot) \right\},$$

and

$$\hat{\sigma}^{ij}_{LL} = \frac{1}{n}\sum_{t=1}^{n} \log g_i^{\hat{\beta}_i n}(Y_t|\cdot) \log g_j^{\hat{\beta}_j n}(Y_t|\cdot) - \left\{\frac{1}{n}\sum_{t=1}^{n} \log g_i^{\hat{\beta}_i n}(Y_t|\cdot)\right\}\left\{\frac{1}{n}\sum_{t=1}^{n} \log g_j^{\hat{\beta}_j n}(Y_t|\cdot)\right\}.$$

Consider a_i as a $(\sum_{i=1}^{w} p_i + w)$-vector, where $(\sum_{i=1}^{w} p_i + i)$th element is one and zero otherwise. Let a be a vector. Then, $a'Z_n$ asymptotically is distributed as $\mathcal{N}(0, a'\Sigma a)$. Also, if X is distributed as $\mathcal{N}_w(\mu, \Sigma)$, the q linear combinations $C_{(q\times w)}X_{(w\times 1)}$ are distributed as $\mathcal{N}_q(C\mu, C\Sigma C')$. Let $S = (0, I_w)$, and consider SZ_n. It is easy to see that

$$\sqrt{n}(\frac{1}{n}LL_{wn} - E_{*w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, S\Sigma S'),$$

where $S\Sigma S' = \Sigma_{LL}$. So

$$C\sqrt{n}(\frac{1}{n}LL_{wn} - E_{*w}) \xrightarrow{\mathcal{L}} \mathcal{N}_q(0, C\Sigma_{LL}C').$$

The calculation of the rejection region using the above multivariate normal distribution needs much more computational cost when the value of w is large. Therefore, we propose another statistics. We need to estimate Σ_{LL}. As a natural estimator, consider,

$$\hat{\Sigma}_{LL} = \left(\hat{C}ov(\log g_i^{\beta_{i*}}(Y_t|\cdot), \log g_j^{\beta_{j*}}(Y_t|\cdot))\right) = \left(\hat{\sigma}^{ij}_{LL}\right).$$

It is clear that

$$\sqrt{n}\Sigma_{LL}^{-1/2}(\frac{1}{n}LL_{wn} - E_{*w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_w).$$

Considering

$$\sqrt{n}\Sigma_{LL}^{-1/2}\hat{\Sigma}_{LL}^{1/2}\hat{\Sigma}_{LL}^{-1/2}(\frac{1}{n}LL_{wn} - E_{*w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_w),$$

since $\hat{\Sigma}_{LL}$ is non-singular, if it was also consistent for Σ_{LL}, then by the multivariate version of Slutsky Theorem, we would obtain,

$$\sqrt{n}\hat{\Sigma}_{LL}^{-1/2}(\frac{1}{n}LL_{wn} - E_{*w}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_w),$$

so,

$$\left\{C\sqrt{n}(\frac{1}{n}LL_{wn} - E_{*w})\right\}' \{C\hat{\Sigma}_{LL}C'\}^{-1} \left\{C\sqrt{n}(\frac{1}{n}LL_{wn} - E_{*w})\right\} \xrightarrow{\mathcal{L}} \chi^2_q.$$

The weak consistency, in the sense of $\hat{\Sigma}_{LL} \xrightarrow{\mathcal{P}} \Sigma_{LL}$, is easily established given a weakly consistent estimate of σ_{LL}^{ij}, denoted as $\hat{\sigma}_{LL}^{ij}$.

To test

$$\mathcal{H}_0 : \mu_1 = \mu_2, \mu_3 = \mu_4 = ... = \mu_w,$$

which is equal to

$$\mathcal{H}_0 : \Delta_{1,2} = 0, \Delta_{3,4} = 0, \Delta_{4,5} = 0, ..., \Delta_{w-1,w} = 0,$$

where $\Delta_{i,j} = \mu_i - \mu_{i+1}, \quad i = 1, 2, \ldots, w-1$, we consider,

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & ... & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Now testing

$$C \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

is equal to testing $\mathcal{H}_0$. For testing

$$\mathcal{H}_0 : \mu_1 = \mu_2 = ... = \mu_w,$$

we use,

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & ... & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & -1 \end{pmatrix}.$$

Sometimes, we want to estimate the difference of the $\mathcal{KL}$ risks, which leads to estimating the differences of the weighted difference of the maximized log- likelihood functions, i.e.,

$$\triangle(g_i^{\beta_{i*}}, g_j^{\beta_{j*}}) = \mathcal{KL}(h, g_i^{\beta_{i*}}) - \mathcal{KL}(h, g_j^{\beta_{j*}}),$$

will be estimated by $-n^{-1}(LL_n(i) - LL_n(j))$. What we really want to estimate is

$$\triangle(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}}) = \mathcal{EKL}\,(h, g_i^{\hat{\beta}_{in}}) - \mathcal{EKL}\,(h, g_j^{\hat{\beta}_{jn}}).$$

Using Akaike's idea, we obtain a simple estimator of $\triangle(g_i^{\hat{\beta}_{in}}), g_j^{\hat{\beta}_{jn}})$:

$$\hat{\triangle}(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}}) = -n^{-1}[LL_n(i) - LL_n(j) - (p_i - p_j)],$$

where $\mathcal{E}\left\{\hat{\triangle}(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}}) - \triangle(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}})\right\}$ is an $o(n^{-1})$. For $w = 2$, we consider the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$. Using Theorem 3.3 of Vuong (1989), when $g_i^{\beta_{i*}} \neq g_j^{\beta_{j*}}$ and obtain that

$$n^{1/2}\{\hat{\triangle}(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}}) - \triangle(g_i^{\hat{\beta}_{in}}, g_j^{\hat{\beta}_{jn}})\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \omega_*^2),$$

where $\omega_*^2 = var[\log \frac{g_i^{\beta_{i*}}(Y_t|\cdot)}{g_j^{\beta_{j*}}(Y_t|\cdot)}]$. An estimator of ω_*^2 is

$$n^{-1}\sum_{i=1}^{n}\left[\log\frac{g_i^{\hat{\beta}_{in}}(Y_t|\cdot)}{g_j^{\hat{\beta}_{jn}}(Y_t|\cdot)}\right]^2 - \left[n^{-1}\sum_{i=1}^{n}\log\frac{g_i^{\hat{\beta}_{in}}(Y_t|\cdot)}{g_j^{\hat{\beta}_{jn}}(Y_t|\cdot)}\right]^2,$$

which is used by Commenges et al. (2008) to compute a tracking interval for difference of $\mathcal{KL}$ risks of two rival models. A main problem in model selection is selecting a simple model in a set of equivalent rival models. Consider five models as $g_i^{\beta_i}, i = 1, 2, ..., 5$. There is a claim that $g_1^{\beta_1} = g_2^{\beta_2}$ and $g_3^{\beta_3} = g_4^{\beta_4} = g_5^{\beta_5}$. We may write this claim as $\mathcal{H}_0 : g_1^{\beta_1} = g_2^{\beta_2}, g_3^{\beta_3} = g_4^{\beta_4} = g_5^{\beta_5}$. Testing $\mathcal{H}_0$ is equivalent to testing

$$C\begin{pmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\\ \mu_5\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix},$$

where

$$C = \begin{pmatrix}1 & -1 & 0 & 0 & 0\\ 0 & 0 & 1 & -1 & 0\\ 0 & 0 & 0 & 1 & -1\end{pmatrix}.$$

Accepting $\mathcal{H}_0$ is equal to categorizing the rival models in sets $S_1 = \{g_1^{\beta_1}, g_2^{\beta_2}\}$ and $S_2 = \{g_3^{\beta_3}, g_4^{\beta_4}, g_5^{\beta_5}\}$.

6 Suitable Set Selection Using Quasi Cluster Approach

For k rival models, consider a divergence matrix as bellow,

$$\mathcal{D}_{\mathcal{M}} = \begin{array}{c} \mathcal{M}(1) \\ \mathcal{M}(2) \\ \mathcal{M}(3) \\ \vdots \\ \mathcal{M}(k) \end{array} \begin{pmatrix} 0 & & & & \\ \hat{\Delta}(g^{\hat{\beta}_{n2}}, g^{\hat{\beta}_{n1}}) & 0 & & & \\ \hat{\Delta}(g^{\hat{\beta}_{n3}}, g^{\hat{\beta}_{n1}}) & \hat{\Delta}(g^{\hat{\beta}_{n3}}, g^{\hat{\beta}_{n2}}) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\Delta}(g^{\hat{\beta}_{nk}}, g^{\hat{\beta}_{n1}}) & \hat{\Delta}(g^{\hat{\beta}_{nk}}, g^{\hat{\beta}_{n2}}) & \hat{\Delta}(g^{\hat{\beta}_{nk}}, g^{\hat{\beta}_{n3}}) & \ldots & 0 \end{pmatrix},$$

where $\mathcal{M}(j)$ indicate model j. To simplify the notation, we indicate $\hat{\Delta}(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$ by $\hat{\Delta}(ji)$, then the new presentation of $\mathcal{D}_{\mathcal{M}}$ is

$$\mathcal{D}_{\mathcal{M}} = \begin{array}{c} \mathcal{M}(1) \\ \mathcal{M}(2) \\ \mathcal{M}(3) \\ \vdots \\ \mathcal{M}(k) \end{array} \begin{pmatrix} 0 & & & & \\ \hat{\Delta}(21) & 0 & & & \\ \hat{\Delta}(31) & \hat{\Delta}(32) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\Delta}(k1) & \hat{\Delta}(k2) & \hat{\Delta}(k3) & \ldots & 0 \end{pmatrix},$$

where 0 indicates the observed value for $\hat{\Delta}(g^{\hat{\beta}_{ni}}, g^{\hat{\beta}_{ni}}) = \hat{\Delta}(ii)$. To select an admissible set of models, we consider a procedure. At the first stage of the procedure, we consider two models which have the lowest $\hat{\Delta}(g^{\hat{\beta}_{nj}}, g^{\hat{\beta}_{ni}})$. They merged to form the smallest set of rival models which are equally close to the true model. We will show it by

$$\hat{\Delta}_{(mt)}; \quad m, t \in \{1, 2, ..., k\}.$$

At the second stage we consider the divergence between this set and the $k-2$ remaining rival models as follows,

$$\hat{\Delta}_{(mt)r} = min\{\hat{\Delta}_{mr}, \hat{\Delta}_{tr}\} \quad \text{for} \quad r = 1, 2, ..., k \quad \& \quad r \neq m, t.$$

We may now form a new divergence matrix as

$$\mathcal{D}^1_{\mathcal{M}} = \begin{pmatrix} 0 & & & & \\ \hat{\triangle}(w(mt)) & 0 & & & \\ \hat{\triangle}(l(mt)) & \hat{\triangle}(lw) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\triangle}(z(mt)) & \hat{\triangle}(zw) & \hat{\triangle}(zl) & ... & 0 \end{pmatrix},$$

where the first row of this matrix shows the absolute value of divergence between $\hat{\triangle}_{(mt)}$ and all of the other rival models. Zero in the first row and the first column is $\hat{\triangle}_{(mt)(mt)}$. This procedure will continue until all of the rival models become a member of the set of models.

6.1 Simulation Study

To verify our approach and answer our question, we perform a simulation study. We consider the data generating proabilities (true model) as *Lognormal*(*LN*) and six rival models. These models are non-nested, and they are mis-specified. Some of them have the same domain and some of them are far from the true model. The models are

model 1; $\mathcal{M}(1)$: Lognormal, $\mathcal{LN}(\alpha_1 = 2, \beta_1 = \sqrt{.5})$;

and six rival models as,

model 2; $\mathcal{M}(2)$: Weibull, $\mathcal{W}(\alpha_2, \beta_2)$;

model 3; $\mathcal{M}(3)$: Gamma, $\mathcal{G}(\alpha_3, \beta_3)$;

model 4; $\mathcal{M}(4)$: Normal, $\mathcal{N}(\alpha_4, \beta_4)$;

model 5; $\mathcal{M}(5)$: Cauchy, $\mathcal{C}(\alpha_5, \beta_5)$;

model 6; $\mathcal{M}(6)$: Uniform, $\mathcal{U}(\alpha_6, \beta_6)$ and

model 7; $\mathcal{M}(1)$: F, $\mathcal{F}(\alpha_7, \beta_7)$.

The parameters of the rival models are estimated as the quasi maximum likelihood estimators. The *AIC* for this models are respectively,

$$310.462, \quad 319.629, \quad 315.481, \quad 352.244, \quad 331.866, \quad 387.756 \quad \text{and} \quad 400.617,$$

for $n = 50$ observation from a $\mathcal{LN}(2, \sqrt{.5})$ model. Matrix for these *AIC*'s is given by $\mathcal{D}_{\mathcal{M}}$ as,

$$\begin{array}{c} \\ \mathcal{M}(1) \\ \mathcal{M}(2) \\ \mathcal{M}(3) \\ \mathcal{M}(4) \\ \mathcal{M}(5) \\ \mathcal{M}(6) \\ \mathcal{M}(7) \end{array} \begin{pmatrix} 0 & & & & & & \\ \frac{|319.62-310.46|}{100} & 0 & & & & & \\ \frac{|315.48-310.46|}{100} & \frac{|315.48-319.62|}{100} & 0 & & & & \\ \frac{352.24-310.46|}{100} & \frac{|352.24-319.62|}{100} & \frac{|352.24-315.48|}{100} & 0 & & & \\ \frac{|331.86-310.46|}{100} & \frac{|331.86-319.62|}{100} & \frac{|331.86-315.48|}{100} & \frac{|331.86-352.24|}{100} & 0 & & \\ \frac{|387.75-310.46|}{100} & \frac{|387.75-319.62|}{100} & \frac{|387.75-315.48|}{100} & \frac{|387.75-352.24|}{100} & \frac{|387.75-331.86|}{100} & 0 & \\ \frac{|400.61-310.46|}{100} & \frac{|400.61-319.62|}{100} & \frac{|400.61-315.48|}{100} & \frac{|400.61-352.24|}{100} & \frac{|400.61-331.86|}{100} & \frac{|400.61-387.75|}{100} & 0 \end{pmatrix},$$

which is equal to,

$$\begin{array}{c} \\ \mathcal{M}(1) \\ \mathcal{M}(2) \\ \mathcal{M}(3) \\ \mathcal{M}(4) \\ \mathcal{M}(5) \\ \mathcal{M}(6) \\ \mathcal{M}(7) \end{array} \begin{pmatrix} 0 & & & & & & \\ 0.091 & 0 & & & & & \\ 0.050 & 0.041 & 0 & & & & \\ 0.418 & 0.326 & 0.368 & 0 & & & \\ 0.214 & 0.122 & 0.164 & 0.204 & 0 & & \\ 0.773 & 0.681 & 0.723 & 0.355 & 0.559 & 0 & \\ 0.902 & 0.810 & 0.851 & 0.484 & 0.688 & 0.129 & 0 \end{pmatrix} = \mathcal{D}_{\mathcal{M}}.$$

In the first step, based on the minimum absolute value of differences between $\mathcal{KL}$'s divergences of rival models, the minimum value is related to divergence between model 2 and model 3,

$$\hat{\Delta}_{(mt)} = \hat{\Delta}_{(23)} = 0.041.$$

Therefore, these two models belong to a same set of models. Now,

$$\hat{\Delta}_{(mt)r} = min\{\hat{\Delta}_{mr}, \hat{\Delta}_{tr}\} = \hat{\Delta}_{(23)r} = min\{\hat{\Delta}_{2r}, \hat{\Delta}_{3r}\} \quad \text{for} \quad r = 1, 4, 5, 6, 7.$$

We see that $\hat{\Delta}_{(23)r}$ for $r = 1, 4, 5, 6, 7$, are $\hat{\Delta}_{(21)}$, $\hat{\Delta}_{(24)}$, $\hat{\Delta}_{(25)}$, $\hat{\Delta}_{(26)}$ and $\hat{\Delta}_{(27)}$, respectively. Based on this computation, the divergence matrix will be,

$$\begin{array}{c} \\ \mathcal{M}(1) \\ \mathcal{M}(23) \\ \mathcal{M}(4) \\ \mathcal{M}(5) \\ \mathcal{M}(6) \\ \mathcal{M}(7) \end{array} \begin{pmatrix} 0 & & & & & \\ 0.050 & 0 & & & & \\ 0.418 & 0.326 & 0 & & & \\ 0.214 & 0.122 & 0.204 & 0 & & \\ 0.773 & 0.681 & 0.355 & 0.559 & 0 & \\ 0.902 & 0.810 & 0.484 & 0.688 & 0.129 & 0 \end{pmatrix} = \mathcal{D}^1_{\mathcal{M}}.$$

Investigation of the last divergence matrix shows that the model 1 (true model) will

belong to our admissible set of rival models. The second step divergence matrix is,

$$\begin{array}{c} \\ \mathcal{M}((23)1) \\ \mathcal{M}(4) \\ \mathcal{M}(5) \\ \mathcal{M}(6) \\ \mathcal{M}(7) \end{array} \begin{pmatrix} 0 & & & & \\ 0.326 & 0 & & & \\ 0.122 & 0.204 & 0 & & \\ 0.681 & 0.355 & 0.559 & 0 & \\ 0.810 & 0.484 & 0.688 & 0.129 & 0 \end{pmatrix} = \mathcal{D}^2_{\mathcal{M}}.$$

In this stage our admissible set of models will be

$$\mathcal{AS}_1 = \Big\{\{model\,2, model\,3\}\{model\,1\}\Big\}.$$

The minimum absolute value of differences between $\mathcal{KL}$'s divergences of rival models indicate that model 5 will attach to the $\mathcal{AS}_1$. Then,

$$\begin{array}{c} \\ \mathcal{M}(((23)1)5) \\ \mathcal{M}(4) \\ \mathcal{M}(6) \\ \mathcal{M}(7) \end{array} \begin{pmatrix} 0 & & & \\ 0.204 & 0 & & \\ 0.559 & 0.355 & 0 & \\ 0.668 & 0.484 & 0.129 & 0 \end{pmatrix} = \mathcal{D}^3_{\mathcal{M}}.$$

Based on our criterion, the minimum value is for model 6 and model 7. The new divergence matrix has a future as

$$\begin{array}{c} \\ \mathcal{M}(((23)1)5) \\ \mathcal{M}(4) \\ \mathcal{M}(67) \end{array} \begin{pmatrix} 0 & & \\ 0.204 & 0 & \\ 0.559 & 0.355 & 0 \end{pmatrix} = \mathcal{D}^4_{\mathcal{M}}.$$

Note that

$$\begin{aligned} \hat{\triangle}_{(mt)(rq)} &= min\{\hat{\triangle}_{(mt)r}, \hat{\triangle}_{(mt)q}\} = min\{\hat{\triangle}_{(((23)1)5)6}, \hat{\triangle}_{(((23)1)5)7}\} \\ &= \hat{\triangle}_{(((32)1)5)6} = 0.559, \end{aligned}$$

and for the other elements of new matrix is as before. This procedure shows that in this stage we have three subset of models,

$$\mathcal{AS}_1 = \Big\{ \underbrace{\underbrace{\{model\,2, model\,3\}}\{model\,1\}\{model\,5\}}\Big\},$$

$$\mathcal{AS}_2 = \Big\{\{model\,6, model\,7\}\Big\},$$

and

$$\mathcal{AS}_3 = \Big\{\{model\,4\}\Big\}.$$

If we continue our procedure, the next candidate to attach to our admissible set is normal density. The ordered *AIC*'s for these seven models is as bellow

$$\begin{aligned} &AIC(model\,1) < AIC(model\,3) < AIC(model\,2) < AIC(model\,5) \\ &< AIC(model\,4) < AIC(model\,6) < AIC(model\,7). \end{aligned}$$

This clustering of models seems reasonable. The models which belong to the $\mathcal{AS}_1$ have smaller *AIC* than the other ones. We have to focus on this set of models to select the best one. Selecting each of these models decreases the bias in model selection. In this discrimination between models for Cauchy density, $\mathcal{M}(5)$, we don't have "adequate reason to consider it as a member of the admissible set, AS_1. We can only say that for the proposed parameters for model 1, $\mathcal{M}(1)$, the Cauchy density is a candidate to describe the data. One may stop his search when he finds a reasonable set of models.

7 Conclusion

In this paper, we have proposed an approach to test whether competing models have expected relations. We have extended Vuong's (1989) results to various cases in non-nested situations. In many situations we have w rival models and a n-sample. For each model, we can compute the individual lack of fit or losses as $-\log(g^{\hat{\beta}_n}(Y_t|\cdot))$, $\quad i = 1,\ldots,n$. The mean corrected losses is equal to $1/(2n)$AIC(j). Assuming the multivariate distribution for the mean corrected losses, we consider models where we can put equality constraints on the mean of this multivariate distribution, which defines models for mean corrected losses, say m-models. Then, we can find the best m-models among the m-models assuming equality of the risks for some rival models. We may try m-models, m-model0: $\mu_1 = \mu_2 = \ldots = \mu_w$, m-model1: $\mu_1, \mu_2 = \ldots = \mu_w$, m-model(1-$w$): $\mu_1, \mu_2, \ldots, \mu_w$ and so on. For instance for m-model1, $-\log(g^{\hat{\beta}_{1n}}(Y_t|\cdot))$ has a certain expectation while $-\log(g^{\hat{\beta}_j n}(Y_t|\cdot))$, $\quad j = 2,\ldots,w$, all have the same expectation, so there are only two parameters for the mean. There are of course other parameters for the covariances. Katayama (2008), in a unpublished paper, has considered a deep mathematical and asymptotic study on the extension of Vuong's (1989) model selection test. As a part of work, Katayama considerd the equality of all rival models. In this work, we have considered a test statistic, which is a little different from Katayama's work. Therefore, we could consider many m-models and different tests. Also, we have

proposed an approach which helps us answer this unsolved problem in statistics: How can we select an admissible set of models which are more reasonable to consider as a rival set of models? This set of rival models leads us to decrease the bias in model selection and make more precise decision to describe the data at hand. This approach lets us consider a large set of models as the candidate set and return out some of them because of their large divergence from the true model.

References

Akaike, H. (1973), Information theory and an extension of maximum likelihood principle. Second International Symposium on Information Theory, Akademia Kiado, 267-281.

Atkinson, A.C. (1970), A method for discriminating between models. *Journal of the Royal Statistical Society B*, **32**, 323-344.

Barmalzan, G. and Sayyareh, A. (2011), The choice of an admissible sete of rival models. *Journal of Statistical Sciences*, **4**(2), 149-165.

Clarke, K., A. and Signorino, C. S. (2010), Discriminating methods: Tests for non-nested discrete choice models. *Political Studies*, **58**, 368-388.

Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J. and Bar-Hen, A. (2008), Estimating a difference of Kullback-Leibler risks Using a normalized difference of AIC. *The Annals of Applied Statistics*, **2**(3), 1123-1142.

Cox, D.R. (1961), Test of separate families of hypothesis. *proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 105 – 123.

Katayama, N. (2008), Portmanteau likelihood ratio tests for model selection, (http://www.economics.smu.edu.sg/femes/2008/169.pdf).

Kullback, S., Leibler, R. (1951), On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

Lorestan, H. and Sayyareh, A. (2017), Model selection using union-intersection principle for non nested models. *Communications in Statistics-Theory and Methods*, **46**(4), 1636-1649.

Pesaran, M. H. (1974), On the general test of model selection. *Review of Economic Studies*, **41**, 153-171.

Pesaran, M. H., and Deaton, A.S. (1978), Testing non-nested nonlinear regression models. *Econometrica*, **46**, 667-694.

Pho, K. H., Ly, S. Ly, S., and Lukusa, T. M. (2019), Comparison among Akaike information criterion, Bayesian information criterion and Vuong's test in model selection: A case study of violated speed regulation in Taiwan. *Journal of Advanced Engineering and Computation*, **3**(1), 293-303.

Sayyareh, A. Obeidi, R., and Bar-Hen, A. (2011), Empirical comparison of some model selection criteria. *Communication in Statistics-Simulation and Computation*, **40**, 72-86.

Sayyareh, A. (2012), Inference after separated hypotheses testing: An investigation for linear models. *Journal of Statistical Computation and Simulation*. **82**(9), 1275-1286.

Sayyareh, A. (2017), Non parametric multiple comparisons of non nested rival models. *Communications in Statistics-Theory and Methods*, **46**(17), 8369-8386.

Shimodiara, H. (1998), An application of multiple comparison techniques to model selection. *Annals of Institute Statistical Mathematics*, **50**(1), 1-13.

Shimodaira, H. (2001), Multiple comparisons of log-likelihoods and combining non-nested models with application to phylogenetic tree selection. *Communication in Statistics-Theory and methods*, **30**, 1751-1772.

Vuong, Q. H. (1989), Likelihood ratio tests for model selection and non-nested hHypotheses. *Econometrica*, **57**(2), 307-333.

Yanagihara, H., and Ohomoto, C. (2005), On distribution of AIC in linear regression models. *Journal of Statistical Planning and Inference*, **133**, 417-433.

White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-26.

White, H. (1982b), Regularity conditions for Cox's test of non-nested hypotheses. *Journal of Econometrics*, **19**, 301-318.

Zucchini, W. (2000), An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41-61.