مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۱۳، شماره‌ی ۱، بهار و تابستان ۱۳۹۵، صص ۵۹–۸۸

J. Statist. Res. Iran **13** (2016): 59–88
DOI: 10.18869/acadpub.jsri.13.1.4

Admissible Set of Rival Models based on the Mixture of Kullback-Leibler Risks

Abdolreza Sayyareh

K. N. Toosi University of Technology

Received: 12/14/2015 Approved: 8/31/2016

Abstract. Model selection aims to find the optimum model. A good model will generally yield good results. Herein lies the importance of model evaluation criteria for assessing the goodness of a subjective model. In this work we want to answer to this question that, how could infinite set of all possible models that could have given rise to data, be narrowed down to a reasonable set of statistical models? This paper considers a finite mixture of the known criterion to the model selection problem to answer to the question. The aim of this kind of criteria is to select an reasonable set of models based on a measure of closeness. We demonstrate that a very general class of statistical criterion, which we call that finite mixture Kullback-Leibler criterion, provides a way of rival theory model selection. In this work we have proposed two types of coefficients for the mixture criterion, one based on the density and another one based on the risk function. The simulation study and real data analysis confirme the proposed criteria.

Keywords. Akaike information criterion; Kullback-Leibler divergence; nonnested model selection.

MSC 2010: 62F07; 62F03.

1 Introduction

Selection of competing rival models lies at the heart of scientific research. Since there exists alternative models explaining the same phenomena, researchers can often increase the plausibility of their theory by rival models. Model selection is recognized as an integral part for statistical inference and then, the concept of the distance between probability distributions is of fundamental importance in statistical decision. The most important problem in model selection arises from the phenomenon as selection bias, Zucchini (2000). If one begins with a very large collection of rival models, then we can be sure that the selected model will have an high maximum likelihood. The selection bias can be expected to be less if we begin with a small set of rival models. But then we have the problem of how to select the few models that go into this set. After selecting the set, the best model from the set will be the postulated model. So we select a good model with a controlled bias. The challenge of testing a theory against its rivals is that the competing theories are often non-nested and the non-nested models require special procedures. A measure of affinity of several distributions has been used by Matusita (1966, 1967). One important aspect of statistical modelling is evaluating the fit of the chosen model. Evaluating the fit of the model to a set of data may take many forms of the divergence between the true model and the selected model. Fisher (1922) pioneered modern frequentist statistics as a model based approach to statistical induction anchored on the notion of a statistical model, formalized by:

$$\mathcal{G} = \{g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \tau\}$$

where $\beta_i \in B_i$, $i = 1, \ldots, m$ and $\tau \geqslant 0$. Indeed, several crucial foundational problems reverberate largely unresolved to this day:
i) How could infinite set of all possible models that could have given rise to data y_0, be narrowed down to a single statistical model?
ii) How could the adequacy of a statistical model be tested a posteriori?
iii) What is the role of substantive information in statistical modelling?
Question (i) is handled by separating the problem into two stage where, a broader family of models within this family is chosen. The problem raised in (ii) is treated by trading goodness of fit against parsimony, and the issue in (iii) is often handled by using substantive information in selecting the family of models. In the literature there are many criteria to evaluate the best model. To introduce the results, suppose a statistician is required

to estimate the true and unknown density, $h(\cdot)$ by $g_{ij}^{\beta_i}(\cdot)$. So he receives a reward $\log g_{ij}^{\beta_i}(y)$. Thus, by selecting $g_{ij}^{\beta_i}(\cdot)$ the statistician expects a return of $\int_y h(y)g^{\beta}(y)\,dy$. How should $g_{ij}^{\beta_i}(\cdot)$ be selected? Suppose the risk of using $g_{ij}^{\hat{\beta}_{in}}(Y)$ as a estimator of h(y), denoted by $R_i(h(Y), g_{ij}^{\hat{\beta}_{in}}(Y))$. Let $R_k(h,\cdot) = R(h(Y), g_{kj}^{\hat{\beta}_{kn}}(Y))$ with probability π_k, $k = 1,\ldots,m$, $\pi_k \geqslant 0, \sum_{i=1}^{m}\pi_i = 1$, where for simplicity we suppose $m = 2$. The expected risk then becomes $\int_y(\pi_1 R_1(h,\cdot)+\pi_2 R_2(h,\cdot))h(y)\,dy$. According to the maximum entropy principal of Akaike (1973), the goodness of a particular model $g_{ij}^{\beta_i}(\cdot)$, as the predictive distribution of a random response, Y, with true density h, is measured by the Kullback-Leibler (1951) risk,

$$\mathcal{KL}(h(y), g^{\beta}(y)) = \int \log\frac{h(y)}{g^{\beta}(y)}h(y)\,dy.$$

Note that, 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. In what follows, let μ and ν denote two probability measures on Ω. Let h and g denote their corresponding density functions with respect to a σ-finite dominating measure ζ, then the Kullback-Leibler divergence based on h and g is defined as

$$\mathcal{KL}(h,g) = \int_{A(\mu)} h\log\frac{h}{g}\,d\zeta$$

where $A(\mu)$ is the support of μ on Ω. This definition is independent of the choice of dominating measure ζ. It is rarely possible to assume that an underlying distribution is exactly characterized by a proposed statistical model $g^{\beta}(y)$. It is reasonable to assume that the proposed statistical model lies in a close proximity to the true underlying distribution. Eguchi and Copas (1998) defined a neighborhood enveloping of model as

$$N_h(\epsilon) = \cup_{\beta\in B}\{g^{\beta} : \mathcal{KL}(h(y), g^{\beta}(y)) \leqslant \epsilon\},$$

where $\epsilon \geqslant 0$. Instead of ϵ we want propose a number r to extend this idea to a set of reasonable models. The r represents the magnitude of resemblance between the best model and rival model $g_{ij}^{\beta_i}(\cdot)$, $i = 1,\ldots,m$. The π_k might not be known and their roles are purely conceptual and the assumption of their existence alone is enough to lead us to an appropriate risk function. In

our work, the importance of each rival model is expressed by π_k's and so by r. The nature of our criterion means that a greater value of π_k for a particular $\mathcal{KL}$ risk corresponds to greater importance or resemblance to the better rival model. Cavanaugh (1999) proposed a model selection criterion which serves as an asymptotically unbiased estimator of a variant of the symmetric divergence between the true model and a fitted model. Commenges et al. (2008) have recently considered the differences of the $\mathcal{KL}$ risks between two rival models. Sayyareh (2012a) have studied inference based on the tracking interval for censored data. Sayyareh et al. (2011) and Sayyareh (2012b) compare some tests and criteria to model selection. A typical approach to the problem of selecting models of differing complexity is to choose the model with the minimum AIC. This paper examines a common scenario in which there is more than one candidate model. A problem is to introduce a procedure for testing more than two non-nested models against one another. It seems that multiple model tests are not necessary given the availability of classical tests for pairwise model comparison. On the other hand we may reduce the number of the rival models based on the their distance from the true model. We have focused on the finite mixture of the Kullback-Leibler divergence criteria to selecting the best set of the rival models which adopted to address both of model uncertainty and model fidelity. To illustrate our approach, consider the case where we have m (possibly non-nested) rival models to be compared. Each model is a family of distributions which indexed by their parameters. In each family may exists some members which are good estimators for the true model. We want to construct a set contains these kind of distributions. In general the m distributions may have several interesting properties and their probability density functions can take different shapes. Although, most of the times the m distribution functions may provide a similar data fit, but still it is desirable to select the more nearly and simple models. The literature on non-nested hypothesis testing in statistics was pioneered by Cox (1961, 1962) and Atkinson (1970), this subject applied by Pesaran (1974) and Pesaran and Deaton (1978). The analysis of non-nested regression models was considered by Davidson and MacKinnon (1981, 2002), Fisher and McAleer (1981). Vuong (1989) considered the hypothesis testing when two competing models are nested, overlap and non-nested. His approach is based on the asymptotic distribution of difference of log-likelihood functions for two competing models. Shimodaira (1998) and Shimodaira (2001) has considered the sampling error of AIC in multiple comparisons and has constructed a set of good models rather than choos-

ing 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) tests the equivalence of two models, the Akaike (1973) information criterion (AIC) is introduced to select the best model under parsimony. This paper is organized as bellow. Zheng et al. (2004), have considered Kullback-Leibler information to the selection of a robust model for testing problems when the true model is unknown, and examined the robustness properties of statistical tests. They have considered two measures based on the Kullback-Leibler information as minimax and maximin approach. Based on these criteria we are able to define a family of admissible genetic models and obtain the corresponding two optimality criteria to select robust models. Their approach is an original work in nested models and especially in a parametric family of densities. In this paper, we have considered an admissible set of non-nested and misspecified nonnegative models which is different in idea and criteria from the Zheng et al. (2004). Section 2 presents the statistical models, $\mathcal{KL}$ risk and the model selection concepts. Section 3 presents the main results to construct the set of admissible distributions based on the proposed finite mixture $\mathcal{KL}$ criterion based on the density and risk function. Sections 4 and 5 present simulation and real data analysis studies, respectively.

2 Statistical Models and $\mathcal{KL}$ Risk

Consider $(\eta, \mathcal{A})$ as a measurable space and $\mathcal{P}$ a subset of probabilities on it. Such a subset is called a family of probabilities. We may parameterized this family. A parametrization can be represented by a function from a set B with values in $\mathcal{P} : \beta \to P^{\beta}$. This parametrization can be denoted by $T = (P^{\beta}; \beta \in B)$. Then we have $\mathcal{P} = \{P^{\beta}; \beta \in B\}$. We call T and $\mathcal{P}$ the statistical families.

A family of probabilities on the sample space of an experiment $(\Omega, \mathcal{F})$ can be called a statistical model and a parametrization of this family will called a parameterized statistical model. If we have two parameterized statistical models $\mathcal{T} = (\mathbf{P}^{\beta}, \beta \in B)$ on $\mathcal{F}_1$ and $\mathcal{T}' = (\mathbf{P}^{\gamma}, \gamma \in \Gamma)$ on $\mathcal{F}_2$ specify the same statistical models if $\mathcal{F}_1 = \mathcal{F}_2$ and they specify the same family of probability on $(\Omega, \mathcal{F}_1)$. The pair $(Y, \mathcal{T})$ of a random variable and a parame-

terized statistical model induce the parameterized family of distributions on $(\mathcal{R}, \mathcal{B}) : T_Y = (P_Y^{\beta}, \beta \in B)$. Conversely, the pair (Y, T_Y) induce $\mathcal{T}$ if $\mathcal{F}_1 = \mathcal{F}$. In that case we may describe the statistical model by (Y, T_Y). Two different random variables Y and X induce two generally different parameterized families on $(\mathcal{R}, \mathcal{B})$, T_Y and T_X. Assume that there is a true, generally unknown probability $\mathbf{P}^*$. Model selection as apart of the statistical inference aims to approach $\mathbf{P}^*$.

Definition 1. Model $\mathcal{T}$ is well specified if $\mathbf{P}^* \in \mathcal{T}$ and is mis-specified otherwise. If it is wellspecified, then there is a $\beta^0 \in B$ such that $\mathbf{P}^{\beta_0} = \mathbf{P}^0$.

In decision theory, estimators are chosen as minimizing some risk function. The most important risk function is based on the Kullback-Leibler divergence. Let a probability $\mathbf{P'}$ is absolutely continuous with respect to a probability $\mathbf{P}$ and $\mathcal{F}_1$ a sub-σ-field of $\mathcal{F}$ the loss using $\mathbf{P'}$ in place 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

$$E_{\mathbf{P}}\{\mathcal{L}_{\mathcal{F}}^{\mathbf{P}/\mathbf{P'}}\} = KL(\mathbf{P}, \mathbf{P'}; \mathcal{F}).$$

This is the Kullback-Leibler ($\mathcal{KL}$) 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 we have $\frac{d\mathbf{P}}{d\mathbf{P'}|\mathcal{F}} = \frac{f_Y(Y)}{g_Y(Y)}$ and the divergence of the distribution $\mathbf{P'}$ relative to $\mathbf{P}$ can be written as

$$\mathcal{KL}(P, P') = \int \log \frac{f_Y(y)}{g_Y(y)} f_Y(y)\, dy.$$

We have that $KL(\mathbf{P}, \mathbf{P'}; \mathcal{F}) = \mathcal{KL}(P, P')$ if $\mathcal{F}$ is the $\sigma - field$ generated by y on $(\Omega, \mathcal{F})$. Based on continuity arguments, we take $0 \log \frac{0}{r} = 0$ for all $r \in I\!R$ and $t \log \frac{t}{0} = \infty$ for all non-zero t.

2.1 Model Selection

Model selection is the task of choosing a model with the correct inductive bias, which in practice means selecting family of densities in an attempt to create a model of optimal complexity for the given data. Suppose a collection of data. Let $\mathcal{M}$ denote a class of these rival models. Each model $\mathcal{G} \in \mathcal{M}$ is considered as a set of probability distribution functions for the data. In this framework we do not impose that one of the candidate models $\mathcal{G}$ in $\mathcal{M}$ is a correct model. A fundamental assumption in classical hypothesis

testing is that h belongs to a parametric family of densities i.e. $h \in \mathcal{G}$. To illustrate model selection, let Y be a random variable from unknown density $h(\cdot)$. A model is assumed as possible explanation of Y, represented by $(g) = \{g^\beta(y), \beta \in B\} = (g^\beta(\cdot))_{\beta \in B}$. This function is known but its parameter as $\beta \in B$ is unknown. The aim is to ascertain whethear (g) can be viewed as a family contained $h(\cdot)$ or has a member which is a good approximate for $h(\cdot)$. The $\log -$likelihood loss of g^β relatively to $h(\cdot)$ for observation Y is $\log \frac{h(Y)}{g^\beta(Y)}$. The expectation of this loss under $h(\cdot)$, or risk, is the $\mathcal{KL}$ divergence between g^β and $h(\cdot)$ as

$$\mathcal{KL}(h, g^\beta) = E_h \left\{ \log \frac{h(Y)}{g^\beta(Y)} \right\}.$$

Hence $\mathcal{KL}$ divergence takes its value in $[0, \infty]$. The $\mathcal{KL}$ divergence is not a metric, but it is additive over marginals of product measures. $\mathcal{KL}(h, g^\beta) = 0$ implies that $h = g^\beta$. Let $\bar{Y} = (Y_1, Y_2, \ldots, Y_n)$ be identically and independently distributed random variables from unknown density $h(\cdot)$. Two rival models are assumed as possible explanation of Y, represented by $(f^\gamma(\cdot))_{\gamma \in \Gamma} = \{f^\gamma(y), \gamma \in \Gamma\}, \Gamma \subset \mathcal{R}^q$ and $(g^\beta(\cdot))_{\beta \in B} = \{g^\beta(y), \beta \in B\}, B \subset \mathcal{R}^p$. These functions are known but their parameters as $\gamma \in \Gamma$ and $\beta \in B$ are unknown. The aim is to ascertain which of the two alternatives $(f^\gamma(.))_{\gamma \in \Gamma}$ and $(g^\beta(.))_{\beta \in \mathcal{B}}$ if any can be viewed as a family contained $h(\cdot)$ or has a member which is a good approximate for $h(\cdot)$. As we see, there is no trivial null hypothesis.

Definition 2. (i) (f) and (g) are nonoverlapping if $(f) \cap (g) = \emptyset$; f is nested in (g) if $(f) \subset (g)$; (g) is well specified if there is a value $\beta_0 \in B$ such that $g^{\beta_0} = h$; otherwise it is misspecified.

We assume that there is a value $\beta_* \in B$ which minimizes $\mathcal{KL}(h, g^\beta)$. 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 $(1982a, 1982b)$. The most plausible view about the statistical hypothesis is that all models are idealization of reality, and non of them is true. But if all models are false, then the two types of errors never arises. One response to say that the null hypothesis may be approximately true, so, in which case rejecting the null hypothesis does count a mistake. Or does it? Selecting the alternative hypothesis can have more serious consequences. But we consider the alternative hypothesis to construct suitable test to model selection. It leads us to measure how far from the truth each model under null and alternative hypotheses is. This may not

be possible, but we can quantify the difference of risks between two models, see Commenges et al. (2008). The problem of testing hypothesis belonging to the same parametric family, also known as testing nested hypotheses. In classical approach, the null hypothesis is obtained as a simplified version of the alternative model. Well-known classical procedures such as those based on the likelihood ratio, Wald, and Lagrange-multiplier principal are available for testing hypotheses. When hypotheses do not belong to the same parametric family, a different approaches is necessary, since some classical procedures can not be applied. A comparison for some tests and criteria to non-nested model selection to find the optimum model is given in Sayyareh et al. (2011).

2.2 Affinity of Several Models

The distance between distributions plays an important role in statistics, and thus far, the statistician treated decision rules based on the distance in many occasions. As the affinity of several distributions, let $g_{1j}^{\beta_1}(y_j), \ldots, g_{mj}^{\beta_m}(y_j)$ be rival density functions. The affinity of these m models is

$$\rho_m(G_1, \ldots, G_m) = \int_{\chi} \{g_{1j}^{\beta_1}(y_j) g_{2j}^{\beta_2}(y_j) \ldots g_{mj}^{\beta_m}(y_j)\}^{1/m} \, d\mu.$$

Matusita (1967) shows that

$$0 \leqslant \rho_m^m(G_1, \ldots, G_m) \leqslant \rho_{m-1}^{m-1}(G_{i_1}, \ldots, G_{i_{m-1}}) \leqslant \ \ldots \ \leqslant \rho_2^2(G_{i_s}, G_{i_t}) \leqslant 1$$

where $(i_1, \ldots, i_{m-1})$ is any set of $(1, 2, \ldots, m)$ and

$$(i_s, i_t) \subset (i_1, i_2, \ldots, i_{m-1}) \subset (1, 2, \ldots, m).$$

Let $g_{ij}^{\beta_i}(y_j)$ be the density determined by a sample from G_i. When the problem is to decide whether or not the m rival models are equal, taking some value ρ_0, we decide that the m models are equal when $\rho_m(G_1, \ldots, G_m) \geqslant \rho_0$ and we decide that they are not identical when $\rho_m(G_1, \ldots, G_m) \leqslant \rho_0$. It is know that, if one begins with a large set of rival models, the selected model have a high maximum likelihood term. It is a overfitting effect. On the other hand, if we begin with a small set of rival models, the section bias can be expected to be less. But then we have the problem of how to select the few models that go into this set. When the number of rival models increase, the affinity measure decrease. So we tend to reject the identity of

the rival models. But the problem is that, we may reject the identity of models because of some irrelevant models. The propose of this paper is to reduce the number of the rival models at the first and then select the best one.

3 Admissible Set of Rival Models based on the Mixture of $\mathcal{KL}$ Risks

The classical tests for model selection are Cox's test (1961, 1962), Vuong's test (1989) and the criterion is the set of information criteria like Akaike information criterion, (1973), AIC. For m rival models we need to $m(m-1)/2$ pairwise tests. Here, the problem is not the number of tests, rather the problem is what happens to the type I error. If the tests are mutually independent and the probability of type I error is α, the probability of at least one type I error in k test is $1-(1-\alpha)^k$. A comparison of five rival models require ten tests. The probability of at least one type I error would be about 0.4 if the tests are independent. The tests are not independent as they are performed on the same data. The actual error rate cannot computed under these conditions. The maximum error rate in this situation is $k \times \alpha$, where k is again the number of individual tests that would have to be performed, see Miller (1981). Thus, a comparison of five rival models with $\alpha = 0.05$ would have a maximum error rate of $10 \times 0.05 = 0.50$. So the large number of the rival models caused the large value of the error rate. Then we should to reduce the number of the rival models. The goal of any model selection philosophy are the notations of model uncertainty and model fidelity. The $\mathcal{KL}$ risk is adopted to address both of these concerns. We consider the problem of selecting the best estimation or approximation for true model among some parameterized family of models. The $\mathcal{KL}$ risk is again employed as a measure of the fidelity of a rival model to the truth. The problem that the AIC aims to solve is the following: we have observed n samples and wish to learn something about the true model. In particular, we have a set of rival models which we may fit to the data. If we choose too simple model then the predictions of future data will be affected by the bias present due to the limitation of the model; in contrast, if we choose an overly complex model then the increased variance in the parameter estimates will lead to poor predictions. The AIC aims to select the model from the rival models that best trades off these two sources of error to give good prediction. We

use this criterion to select the suitable members of the different rival models which have the equal distance in $\mathcal{KL}$ sense from the true data generating density. Consider a finite number of m different statistical models as

$$\mathcal{G} = \{g_{ij}^{\beta_i}(\cdot); \beta_i \in B_i,\ j = 1, \ldots, \dim B_i = d_i \quad i = 1, \ldots, m\} = (g_{ij}^{\beta_i}).$$

Note that each member of $\mathcal{G}$ has a $\mathcal{KL}$ divergence as $\mathcal{KL}(h, g_{ij}^{\beta_i})$. Sometimes it is more reasonable to consider the risk $\mathcal{E}_h\{\log \frac{h(Y)}{g^{\hat{\beta}_n}(Y)}\}$ that we call the expected Kullback-Leibler risk and that we denote by $\mathcal{EKL}(h, g^{\hat{\beta}_n}(Y))$, $\hat{\beta}_n$ is the (quasi) maximum likelihood estimator of β. The expected Kullback-Leibler risk is the sum of the mis-specification risk plus the statistical risk:

$$\mathcal{EKL}(h, g^{\hat{\beta}_n}) = \mathcal{EKL}(h, g^{\beta_*}) + \mathcal{EKL}(g^{\beta_*}, g^{\hat{\beta}_n}).$$

If g is wellspecified we have $\mathcal{EKL}(h, g^{\beta_*}) = 0$. We cannot estimate $\mathcal{KL}(h, g_{ij}^{\beta_i})$ 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, AIC, as:

$$AIC(i) = -2LL_n(i) + 2p_i,$$

where

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

is the maximized log-likelihood function for model $(g_{ij}^{\beta_i})$ and p_i is the number of estimated parameter in model i. AIC was revisited by Linhart and Zucchini (1986) who showed that

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

where

$$B_{g_{ij}} = \mathcal{E}_h \left\{ \left[\frac{\partial \log g_{ij}^{\beta_i}(Y)}{\partial \beta_i} \right] \left[\frac{\partial \log g_{ij}^{\beta_i}(Y)}{\partial \beta_i'} \right] \Bigg|_{\beta_{i*}} \right\}$$

and

$$A_{g_{ij}} = -\left\{ \mathcal{E}_h \left[\frac{\partial^2 \log g_{ij}^{\beta_i}(Y)}{\partial \beta_i \partial \beta_i'} \Bigg|_{\beta_{i*}} \right] \right\}.$$

We also have:

$$\mathcal{EKL}(h, g_{ij}^{\hat{\beta}_{in}}) = F(h) - \mathcal{E}_h\{n^{-1}LL_n(i)\} + \frac{1}{n}Tr(B_{g_{ij}}A_{g_{ij}}^{-1}) + o_p(n^{-1}). \quad (2)$$

Akaike information criterion follows from (2) by multiplying by $2n$ deleting constant term $F(h)$ replacing second term in the right by $n^{-1}LL_n(i)$ and replacing $Tr(B_{g_{ij}}A_{g_{ij}}^{-1})$ by p_i. The term $\frac{1}{n}Tr(B_{g_{ij}}A_{g_{ij}}^{-1})$ 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 $B_g = A_g$. 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_{i*}, is known as pseudo true value of parameter. These pseudo-true values are defined by

$$\beta_{i*} = \arg\max_{\beta_i \in B_i} \mathcal{E}_h \left\{ \frac{1}{n} \sum_{j=1}^{n} \log g_{ij}^{\beta_i}(Y_j) \right\}.$$

In each models $(g_{ij}^{\beta_i}(\cdot))$ exists a density which based on some properties could be a good model to approximate the true model. Sometimes we have not enough reasons to select a model against the another one as a good approximate. We consider a finite mixture of these good models as $\sum_{j=1}^{n} \alpha_i \log g_{ij}^{\beta_{ij}}(\cdot)$. The problem is that how one can decide on the good models and their coefficients in the mixture model. One of these models, is the best one, and we expect the mixture gives the largest coefficient to this model. In the following Theorem, we aim to find these kind of coefficients based on the risk function and the information criterion. Model has a $\mathcal{KL}$ risk as $\mathcal{KL}_i = E_h\{\log \frac{h(Y)}{g_{ij}^{\beta_i}(Y)}\}$. The finite mixture of these risks is given by

$$\sum_{i=1}^{m} \alpha_i E_h \left\{ \log \frac{h(Y)}{g_{ij}^{\beta_i}(Y)} \right\}$$

The proposed finite mixture $\mathcal{KL}$ criterion then selects a set of the rival models which are belong to the

$$\mathcal{G} = \{g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \eta\}$$

where η is defined as $\eta = \mathcal{KL}(h, \sum_{i=1}^{m} \alpha_i \log g_{ij}^{\beta_i}(\cdot))$ and $\mathcal{G}_i$ is the class of ith model.

Definition 3. Admissible Set Induced by $\mathcal{KL}$ Infinite Mixture Criterion: A set contains some models g^{β} such that $\mathcal{KL}(h, g^{\beta}) < \zeta$, is called admissible set induced by $\mathcal{KL}$ infinite mixture criterion. Furthermore

$$\mathcal{G} = \{g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \zeta\}$$

is the admissible set induced by $\mathcal{KL}$ infinite mixture criterion.

3.1 A Probabilistic Setup for Model Uncertainty

Suppose a set of m models $\mathcal{M} = \{M_1, \ldots, M_m\}$ are under considerations for data Y, and that under M_i, Y has density $g^{\beta}(Y|M_i)$. We consider a probability $p(M_i)$ to each model. The model selection problem becomes that of finding the model in $\mathcal{M}$ that generated the data or the nearest one to the true model. The probability that M_i was this model is the

$$p(M_i|Y) = \frac{p(Y|M_i)p(M_i)}{\sum_{i=1}^{m} p(Y|M_i)p(M_i)}$$

which reduce to $p(h|Y) = 1$ when the models are wellspecified and $h(\cdot)$ is the true model. A simple and popular choice is the uniform probability $p(M_i) = 1/m$ which is noninformative in the sense of favoring all models equally. Under this selection

$$p(M_i|Y) = \frac{p(Y|M_i)}{\sum_{i=1}^{m} p(Y|M_i)}. \tag{3}$$

Although uniform over models, it will typically not be uniform on model characteristics such as model size. A more subtle problem occurs in setup where many models are very similar and only a few are distinct. In such case, $p(M_i) = 1/m$ will not assign probability uniformly to model neighborhoods and may bias the $p(M_i|Y)$ away from good models. The large value of $p(M_i|Y)$ indicate that the model M_i is a good estimate for the unknown true density h. Based on the posterior probabilities, pairwise comparison of models, say M_k and M_l, is summarized by posterior odds

$$\frac{p(M_k|Y)}{p(M_l|Y)} = \frac{p(Y|M_k)}{p(Y|M_l)} \times \frac{p(M_k)}{p(M_l)}.$$

The term $p(Y|M_k)/p(Y|M_l)$ is the Bayes factor, which update the prior odds $p(M_k)/p(M_l)$ to yield the posterior odds. The model posterior distribution is the fundamental object of interest for model selection. Insofar as the priors $p(M_i)$ provide an initial representation of model uncertainty, the posterior summarizes all the relevant information in the data Y. By treating $p(M_i|Y)$ as a measure of truth of model M_i, a natural and simple strategy for model selection is to choose the most probable M_i, the one for which $p(M_i|Y)$ largest, or, equivalently $p(Y|M_i)$ is the largest. A normalized version of this criterion is considered in the following subsection.

3.2 Admissible Set based on the Best Elements of the Rival Models

We expect the best model in the set of the rival models has the greatest chance to select as the model at hand. It reflects that, in the mixture of models we want assigned the greatest coefficient to the best rival model. To select the coefficients of a mixture of $\mathcal{KL}$ risks, we consider the suitable densities. Specifically, we use the latent variable $\mathcal{U}_i$ to represent the theory with which model i is consistent. Thus $\mathcal{U}_i$ can take one of m values, depending on which rival model is considered as a approximation to h. As we noted the finite mixture $\mathcal{KL}$ criterion then selects a set of the rival models which are belong to the

$$\mathcal{G} = \{g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \tau\}$$

where τ is defined as

$$\tau = \sum_{i=1}^{m} \pi_i \mathcal{KL}(h, g_{i*}^{\beta_{i*}}). \tag{4}$$

In this equation $\pi_i \propto p(\mathcal{U}_i = m)$ represents the relative frequency of model $g_{i*}^{\beta_{i*}}$ as a nearest model to the true one, where $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_i > 0$ for each i. We assume that there is a value $\beta_{i*} \in B_i$ which minimizes $\mathcal{KL}(h(\cdot), g_{ij}^{\beta_i})$ for $j = 1, \ldots, d_i$. If the model is wellspecified $\beta_{i0} = \beta_{i*}$, where β_{i*} minimizes $\mathcal{KL}(h, g_{ij}^{\beta_i})$. The MLE, $\hat{\beta}_{in}$, is a consistent estimator for β_{i*}.

Definition 4. Admissible Set Induced by $\mathcal{KL}$ Infinite Mixture Criterion: A set contains some models g^{β} such that $\mathcal{KL}(h, g^{\beta}) < \tau$, is called admissible set induced by $\mathcal{KL}$ infinite mixture criterion. Furthermore

$$\mathcal{G} = \{g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \tau\}$$

is the admissible set induced by $\mathcal{KL}$ infinite mixture criterion.

An estimation for $\mathcal{KL}(h, g_{i*}^{\beta_{i*}})$ is

$$\begin{aligned}\hat{\mathcal{KL}}(h, g_{i*}^{\beta_{i*}}) &= \hat{F}(h) + AIC(i) = \hat{F}(h) - 2LL_n(i) + 2p_i \\ &= \hat{F}(h) - 2\sum_{j=1}^{n} \log g_{i*}^{\hat{\beta}_{in}}(Y_j) + 2p_i.\end{aligned}$$

So

$$\hat{\tau} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} \hat{\pi}_i\{\hat{F}(h) - 2\sum_{j=1}^{n} \log g_{i*}^{\hat{\beta}_{in}}(Y_j) + 2p_i\}.$$

The entropy, $F(h) = \mathcal{E}_h\{\log h(Y)\}$, cannot be correctly estimated. We may estimate it as $\frac{1}{n}\sum_{j=1}^{n} \hat{h}_k(Y_j)$ where $\hat{h}_k(\cdot)$ is a kernel estimate of the true density. From Definition 3 and (3),

$$\mathcal{G} = \left\{ g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant \sum_{i=1}^{m} \pi_i \mathcal{KL}(h, g_{i*}^{\beta_{i*}}) \right\}$$

which defines a new version of admissible set as

$$\mathcal{G} = \left\{ (g_{ij}^{\beta_i} \in \mathcal{G}_i \mid F(h) - \mathcal{E}_h\{\log g_{ij}^{\beta_i}\}) \leqslant \sum_{i=1}^{m} \pi_i [F(h) - \mathcal{E}_h\{\log g_{i*}^{\beta_{i*}}\}] \right\}$$

which implies

$$\mathcal{G} = \left\{ (g_{ij}^{\beta_i} \in \mathcal{G}_i \mid -\mathcal{E}_h\{\log g_{ij}^{\beta_i}\}) \leqslant \sum_{i=1}^{m} \pi_i [-\mathcal{E}_h\{\log g_{i*}^{\beta_{i*}}\}] \right\}$$

or

$$\mathcal{G} = \left\{ g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \sum_{i=1}^{m} \pi_i \mathcal{E}_h\{\log g_{i*}^{\beta_{i*}}\} \leqslant \mathcal{E}_h\{\log g_{ij}^{\beta_i}\} \right\}.$$

The last set indicate that we select the jth element of the ith family of rival densities for admissible set if $\sum_{i=1}^{m} \pi_i \mathcal{E}_h\{\log g_{i*}^{\beta_{i*}}\} \leqslant \mathcal{E}_h\{\log g_{ij}^{\beta_i}\}$. We may estimate this inequality by

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} \hat{\pi}_i\{2\log g_{i*}^{\hat{\beta}_{in}}(Y_j) - 2p\} \leqslant \frac{1}{n}\sum_{j=1}^{n} \log g_{ij}^{\beta_i}(Y_j). \tag{5}$$

So

$$\mathcal{G} \cong \left\{ g_{ij}^{\beta_i} \in \mathcal{G}_i \mid \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} \hat{\pi}_i\{2\log g_{i*}^{\hat{\beta}_{in}}(Y_j) - 2p\} \leqslant \frac{1}{n}\sum_{j=1}^{n} \log g_{ij}^{\beta_i}(Y_j) \right\},$$

where p is the number of estimated parameters in model $\mathcal{G}_i$. The marginal likelihood, $p(Y|M_i) \propto p(M_i|Y) \simeq g_{i*}^{\beta_{i*}},\ i = 1, \ldots m$ is the fundamental object of interest for model selection, which summarizes all the relevant information in the data Y and provides a complete data representation of model uncertainty. By treating $p(M_i|Y)$ as a measure of truth of model M_i, a natural and simple strategy for model selection is to choose the most probable M_i, the one for which $p(M_i|Y) \propto p(Y|M_i)$ is the largest. We normalize the marginal likelihood as

$$\pi_k = \frac{p(Y|M_k)}{\sum_{i=1}^{m} p(Y|M_i)} = \frac{g_{i*}^{\beta_{i*}}(Y_i)}{\sum_{i=1}^{m} g_{i*}^{\beta_{i*}}(Y_i)}$$

thus

$$\hat{\pi}_k = \frac{g_{i*}^{\hat{\beta}_{in}}(Y_i)}{\sum_{i=1}^{m} g_{i*}^{\hat{\beta}_{in}}(Y_i)}.$$

For $m = 1$ we have $\pi_i = \frac{g_{i*}^{\beta_{i*}}(y)}{\sum_{i=1}^{m} g_{i*}^{\beta_{i*}}(y)} = 1$ and τ reduce to $\tau = \mathcal{KL}(h, g_{i*}^{\beta_{i*}})$ which is an information criterion when we are faced with one family of densities.

3.3 Admissible Set based on the Risk Function

Let $g_{i*}^{\beta_{i*}}(\cdot)$ is the best model to approximate the true density and y^* is the unique mode of $h(\cdot)$. When y^* is near to β_{i*} behave as a point in B_i. Consequently, the divergence between β_{i*} and y^* and so the divergence between $\hat{\beta}_{in}$ and y^* must be small. So, $\hat{\beta}_{in}$ as the mode of the best rival model to estimate the true density, h, must be near to y^*. It indicates that, we expect the risk function, $E_h\{(\hat{\beta}_{in} - y^*)^2\}$, should have a small value. On the other hand, the relative variation of $g_{ij}^{\beta_i}(\cdot)$ based on its parameter from β_{i*} to y^*

must be small. Assume the asymmetric squared-log error loss of the form

$$\mathcal{L}(g_{ij}^{\hat{\beta}_{in}}(Y), h(y^*)) = \left\{ \log \frac{g_{ij}^{\hat{\beta}_{in}}(Y)}{h(y^*)} \right\}^2 = \{\log g_{ij}^{\hat{\beta}_{in}}(Y) - \log h(y^*)\}^2.$$

This loss function is balanced in the sense that $\lim \mathcal{L}(g_{ij}^{\hat{\beta}_{in}}(Y), h(y^*))$ goes to infinity as $g_{ij}^{\hat{\beta}_{in}}(Y)$ approaches to zero or ∞. This loss function is convex for $g_{ij}^{\hat{\beta}_{in}}(Y)/h(y^*) \leqslant e$ and concave otherwise, but its risk function has unique minimum with respect to $g_{ij}^{\hat{\beta}_{in}}(Y)$.

In the following Theorem the difference between $\log g_{ij}^{\hat{\beta}_{in}}(y)$ and $\log h(y^*)$ introduces a criterion for model choice.

***Theorem* 1.** *Let $Y_1, \ldots, Y_n$ is an i.i.d. random sample from the true density $h(\cdot)$ and*

$$\mathcal{G} = \{g_{ij}^{\beta_i}(\cdot); \beta_i \in B_i,\ j = 1, \ldots, \dim B_i = d_i \quad i = 1, \ldots, m\} = (g_{ij}^{\beta_i}).$$

*A collection of m non-nested models, where $g_{ij}^{\beta_i} \in (g_{ij}^{\beta_i})$ for $i = 1, \ldots, m$ and $j = 1, \ldots, d_i$. Consider y^* as the unique mode of $h(\cdot)$. Then the squared-*log*-risk function is proportional to $E_h\{(y^* - \hat{\beta}_{in})^2$, i.e.*

$$E_h\{[\log g_{ij}^{\hat{\beta}_{in}}(Y) - \log h(y^*)]^2\} = I(\beta_{i*}) E_h\{(y^* - \hat{\beta}_{in})^2\}$$

Proof. Taylor expansions of $g_{ij}^{\hat{\beta}_{in}}(y)$ around β_{i*} and $g_{ij}^{\beta_{i*}}(y)$ around $h(y^*)$ are given as

$$\log g_{ij}^{\hat{\beta}_{in}}(y) = \log g_{ij}^{\beta_{i*}}(y) + \left. \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_i} \right|_{\beta_{i*}} (\hat{\beta}_{in} - \beta_{i*}) + o(1) \tag{6}$$

and

$$\log g_{ij}^{\beta_{i*}}(y) = \log h^*(y) + \lim_{|\beta_{i*}-y^*|<\epsilon} \left\{ \frac{\log g_{ij}^{\beta_{i*}}(y) - \log g_{ij}^{y^*}(y)}{\beta_{i*} - y^*} \right\} \{y^* - \beta_{i*}\} + o(1)$$

$$= \log h^*(y) + \lim_{|\beta_{i*}-y^*|<\epsilon} \left\{ \frac{\log g_{ij}^{y^*}(y) - \log g_{ij}^{\beta_{i*}}(y)}{y^* - \beta_{i*}} \right\} \{y^* - \beta_{i*}\} + o(1) \quad (7)$$

respectively. Using Fisher score algorithm in (4),

$$\log g_{ij}^{\beta_{i*}}(y) = \log h(y^*) - \left. \frac{\partial \log g_{ij}^{y^*}(y)}{y^*} \right|_{\beta_{i*}} \left\{ y^* - \hat{\beta}_{in} + I^{-1}(\beta_{i*}) \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_i} \right\} + o(1)$$

$$= \log h(y^*) - \left. \frac{\partial \log g_{ij}^{y^*}(y)}{\partial y^*} \right|_{\beta_{i*}} (y^* - \hat{\beta}_{in}) - \left. \frac{\partial \log g_{ij}^{y^*}(y)}{\partial y^*} \right|_{\beta_{i*}} I^{-1}(\beta_{i*}) \left. \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_{ij}} \right|_{\beta_{i*}} + o(1)$$

From (5) and (6) we have

$$\log g_{ij}^{\hat{\beta}_{in}}(y) - \log h(y^*) = \left. \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_i} \right|_{\beta_{i*}} I^{-1}(\beta_{i*}) \left. \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_i} \right|_{\beta_{i*}} - \left. \frac{\partial g_{ij}^{y^*}(y)}{\partial y^*} \right|_{\beta_{i*}} (y^* - \hat{\beta}_{in}) - \left. \frac{\partial \log g_{ij}^{y^*}(y)}{\partial y^*} \right|_{\beta_{i*}} I^{-1}(\beta_{i*}) \left. \frac{\partial \log g_{ij}^{\beta_i}(y)}{\partial \beta_i} \right|_{\beta_{i*}}$$

$$= - \left. \frac{\partial \log g_{ij}^{y^*}(y)}{\partial y^*} \right|_{\beta_{i*}} (y^* - \hat{\beta}_{in}).$$

So

$$E_h\{[\log g_{ij}^{\hat{\beta}_{in}}(Y) - \log h(y^*)]^2\} = E_h\left\{\left[-\frac{\partial \log g_{ij}^{y^*}(Y)}{\partial y^*}\Big|_{\beta_{i*}}(y^* - \hat{\beta}_{in})\right]^2\right\}$$
$$= E_h\left\{E_h\left\{\left[-\frac{\partial \log g_{ij}^{y^*}(Y)}{\partial y^*}\Big|_{\beta_{i*}}(y^* - \hat{\beta}_{in})\right]^2\right\}\Big|_{\hat{\beta}_{in}}\right\}$$
$$= E_h\{I(\beta_{i*})(y^* - \hat{\beta}_{in})^2\}$$
$$= I(\beta_{i*})E_h\{(y^* - \hat{\beta}_{in})^2\}.$$

□

Theorem 2. *An admissible set of rival models based on minimum risks is given by*

$$\mathcal{G} = \{g_{ij}^{\beta_{ij}} \in \mathcal{G}_i \mid \mathcal{KL}(h, g_{ij}^{\beta_i}) \leqslant r\}$$

where, $\mathcal{G}_i$ is the class of ith model and $r = \sum_{i=1}^m \alpha_i \mathcal{KL}(h, g_{i*}^{\beta_{i*}})$. Thus, the best (i.e. minimum variance) weighted average $\sum_{i=1}^m \alpha_i \mathcal{KL}(h, g_{i*}^{\beta_{i*}})$ with $\sum_{i=1}^m \alpha_i = 1$, and $E\{\sum_{i=1}^m \alpha_i \mathcal{KL}(h, g_{i*}^{\beta_{i*}})\} = \xi$ assign to $\mathcal{KL}(h, g_{ij}^{\beta_i})$ the weight

$$\alpha_i = \frac{\mathcal{L}^{-1}(g_{ij}^{\hat{\beta}_{in}}(Y), h(y^*))}{\sum_{k=1}^m \mathcal{L}^{-1}(g_{kj}^{\hat{\beta}_{kn}}(Y), h(y^*))} = \frac{I^{-1}(\beta_{i*})E_h^{-1}[\hat{\beta}_{in} - y^*]^2}{\sum_{i=1}^m I^{-1}(\beta_{i*})E_h^{-1}[\hat{\beta}_{in} - y^*]^2}.$$

Proof. The $\mathcal{KL}(h, g_{i*}^{\beta_{i*}})$ has an estimator as

$$AIC(i*) = -2\Sigma_{j=1}^n \log g_{ij}^{\beta_{i*}}(Y_{ij}) + 2p,$$

so

$$\sum_{i=1}^m \alpha_i \left\{-2\Sigma_{j=1}^n \log g_{ij}^{\beta_{i*}}(Y_{ij}) + 2p\right\} = -2\sum_{i=1}^m \sum_{j=1}^n \alpha_i \log g_{ij}^{\beta_{i*}}(Y_{ij}) + 2p.$$

The term $-2\log g_{ij}^{\beta_{i*}}(Y_{ij})$ has a variance as

$$\begin{aligned} Var(-2\log g_{ij}^{\beta_{i*}}(Y_{ij})) &= 4\mathcal{E}\left\{(\log g_{ij}^{\beta_{i*}}(Y_{ij}) - \mathcal{E}\{\log g_{ij}^{\beta_{i*}}(Y_{ij})\})^2\right\} \\ &\simeq 4\mathcal{E}\left\{(\log g_{ij}^{\beta_{i*}}(Y_{ij}) - \mathcal{E}\{\log h(Y^*)\})^2\right\} = 4\sigma^2 \end{aligned}$$

where $\sigma^2 = \mathcal{E}\left\{(\log g_{ij}^{\beta_{i*}}(Y_{ij}) - \mathcal{E}\{\log h(Y^*)\})^2\right\}$. Note that $\log g_{ij}^{\beta_{i*}}(Y_{ij})$ for $i = 1, \ldots, m$ are independent. So the best weighted average ξ assigns to $\log g^{\beta_{i*}}(Y_{ij})$ is

$$\frac{1/\sigma_i^2}{\sum_{i=1}^m 1/\sigma_i^2} = \frac{I^{-1}(\beta_{i*})E_h^{-1}[\hat{\beta}_{in} - y^*]^2}{\sum_{i=1}^m I^{-1}(\beta_{i*})E_h^{-1}[\hat{\beta}_{in} - y^*]^2}.$$

See Theorem 1 and the fact that, the best linear estimator for a common mean for independent random variables $T_1, \ldots, T_n$ with different variances σ_i^2 and common mean is given by $\sum_{i=1}^n w_i T_i$ with $w_i = \frac{1/\sigma_i^2}{1/\sum_{j=1}^n \sigma_j^2}$, $i = 1, \ldots, n$. □

This Theorem shows that the best model to approximate the true one is the model with minimum $I(\beta_{i*})E_h\{(y^* - \hat{\beta}_{in})^2\}$ or maximum $I^{-1}(\beta_{i*})E_h^{-1}\{(y^* - \hat{\beta}_{in})^2\}$. The mixture of the Kullback-Leibler risks for m rival models is given by $\mathcal{KL}_m = \sum_{i=1}^n \alpha_i \mathcal{KL}(h, g_{ij}^{\beta_{ij}}) = \sum_{i=1}^n \alpha_i E_h\{\log \frac{h(Y)}{g_{ij}^{\beta_{ij}}(Y)}\}$. To minimize $\mathcal{KL}_m$ we need to maximize the relevant part of $\mathcal{KL}$. So the suitable α_i is given by

$$\alpha_i = \frac{I^{-1}(\beta_{i*})E_h^{-1}\{(y^* - \hat{\beta}_{in})^2\}}{\sum_{i=1}^n I^{-1}(\beta_{i*})E_h^{-1}\{(y^* - \hat{\beta}_{in})^2\}}.$$

4 Simulation Study

Log-normal, Weibull and Gamma distributions are the three most popular distributions in statistics. They are the three most popular distributions for analyzing skewed lifetime data. We want to decide which members of rival models are belong to the admissible set of models. Words, whether or not the rival models have the same distance in $\mathcal{KL}$ sense from the true model. When we say two models are equivalent, it is not clear whether two models are mis-specified or both of them are well-specified. The second case will be worst when two models overlap. First, in the simulation study based on the

density, we change the rule of the true density. Each times we consider one of the three densities as the true model and the others as the rival models. In the simulation based on the information, we consider the data generating probabilities as $Log-normal(\mathcal{LN})$, and $Gamma(\Gamma)$ and $Weibull(\mathcal{W})$ as rival models. We generate 10^4 Monte-Carlo data sets of sample size $n = 80$. For each iteration of given sample size, we compute the proposed criteria.

4.1 Simulation Study for the Criterion based on the Best Elements of the Rival Models

For each iteration of given sample size, we compute the

$$L.H.S. = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m}\hat{\pi}_i\{2\log g_{i*}^{\hat{\beta}_{in}}(Y_j) - 2p\}$$

and

$$R.H.S. = \frac{1}{n}\sum_{j=1}^{n}\log g_{ij}^{\beta_i}(Y_j)$$

from inequality (5) Subsection 3.2 for each of the distributions under related models. For each hypotesized data generating model, the L.H.S. of (5) is constant but R.H.S based on the rival models will change. As a result, when we set the $\mathcal{LN}(2.00, 1.50)$ as the data generating density, the gamma densities with parameters $(2.02, 4.10)$ and $(1.10, 3.03)$ must be very close to the true model, whereas the gamma densities with parameters as $(2.20, 1.30)$ and $(2.02, 0.91)$ does not. In this situation, the Weibull density with parameters $(0.20, 1.48)$ is near to the true density but the Weibull densities with parameters $(0.90, 1.00)$, $(1.5, 3.80)$ and $(3.60, 1.00)$ do not, see Figure 1. The result of simulation is given in Table 1. Then the admissible set is $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$ where

$$\mathcal{G}_1 = \{\mathcal{LN}(2.90, 0.20), \mathcal{LN}(0.01, 1.96), \mathcal{LN}(0.01, 1.09), \mathcal{LN}(7.00, 5.00)\},$$

$$\mathcal{G}_2 = \{\Gamma(2.02, 4.10), \Gamma(1.10, 3.03)\}$$

and

$$\mathcal{G}_3 = \{\mathcal{W}(0.2, 1.48)\}.$$

When the data generating density is $\Gamma(2.00, 1.50)$, see column 4, the

Table 1. Computation of the $L.H.S.$ and $L.H.S.$ for Log-normal, Weibull and Gamma distributions.

	Data Generating Density					
	$\mathcal{LN}(2.00, 1.50)$		$\Gamma(2.00, 1.50)$		$\mathcal{W}(2.00, 1.50)$	
	L.H.S$= -11.6325$		L.H.S$= -6.3109$		L.H.S$= -5.9939$	
Rival Model	Parameters	R.H.S	Parameters	R.H.S	Parameters	R.H.S
	$(2.90, 0.20)$	-4.2885	$(3.20, 0.90)$	-7.5616	$(2.90, 0.50)$	-16.9864
$\mathcal{LN}$	$(0.01, 1.96)$	-4.5879	$(0.01, 1.96)$	-1.6557	$(0.01, 1.96)$	-1.7278
	$(0.01, 1.09)$	-5.9546	$(0.01, 1.09)$	-1.2477	$(0.01, 1.09)$	-1.2604
	$(7.00, 5.00)$	-5.1710	$(7.00, 5.00)$	-3.5090	$(7.00, 5.00)$	-3.5760
	$(2.02, 4.10)$	-5.7687	$(2.02, 4.10)$	-3.1901	$(2.02, 0.10)$	-8.1762
Γ	$(2.20, 1.30)$	-17.2842	$(2.20, 1.30)$	-1.6866	$(3.00, 1.00)$	-1.8186
	$(1.10, 3.03)$	-7.1760	$(1.10, 3.03)$	-1.5972	$(1.10, 3.03)$	-1.5871
	$(2.02, 0.91)$	-19.1164	$(6.00, 1.5)$	-8.1618	$(2.02, 0.90)$	-1.1494
	$(0.90, 1.00)$	-16.7495	$(0.90, 1.00)$	-1.3348	$(2.00, 1.50)$	-0.9904
$\mathcal{W}$	$(1.50, 3.80)$	-15.6586	$(1.50, 3.80)$	-1.8399	$(1.90, 1.40)$	-0.9914
	$(3.60, 1.00)$	-24.8525	$(3.60, 1.00)$	-12.6891	$(3.05, 3.80)$	-2.8584
	$(0.20, 1.48)$	-4.8633	$(1.67, 0.48)$	-4.9712	$(2.00, 1.48)$	-0.9895

admissible set is $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$ where

$$\mathcal{G}_1 = \{\mathcal{LN}(0.01, 1.96), \mathcal{LN}(0.01, 1.09), \mathcal{LN}(7.00, 5.00)\},$$

$$\mathcal{G}_2 = \{\Gamma(2.02, 4.10), \Gamma(2.20, 1.30), \Gamma(1.10, 3.03)\}$$

and

$$\mathcal{G}_3 = \{\mathcal{W}(0.90, 1.00), \mathcal{W}(1.50, 3.80), \mathcal{W}(1.670.48)\}.$$

On the other hand, we choose Weibull as the data generating density with the parameters that, set this density like the some gamma or Log-normal densities, see column 6, which introduce the admissible set as $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$ where

$$\mathcal{G}_1 = \{\mathcal{LN}(0.01, 1.96), \mathcal{LN}(0.01, 1.09), \mathcal{LN}(7.00, 5.00)\},$$

$$\mathcal{G}_2 = \{\Gamma(3.00, 1.00), \Gamma(1.10, 3.03), \Gamma(2.02, 0.91)\}$$

and

$$\mathcal{G}_3 = \{\mathcal{W}(2.00, 1.50), \mathcal{W}(1.90, 1.40), \mathcal{W}(3.05, 3.80), \mathcal{W}(2.00, 1.48)\}.$$

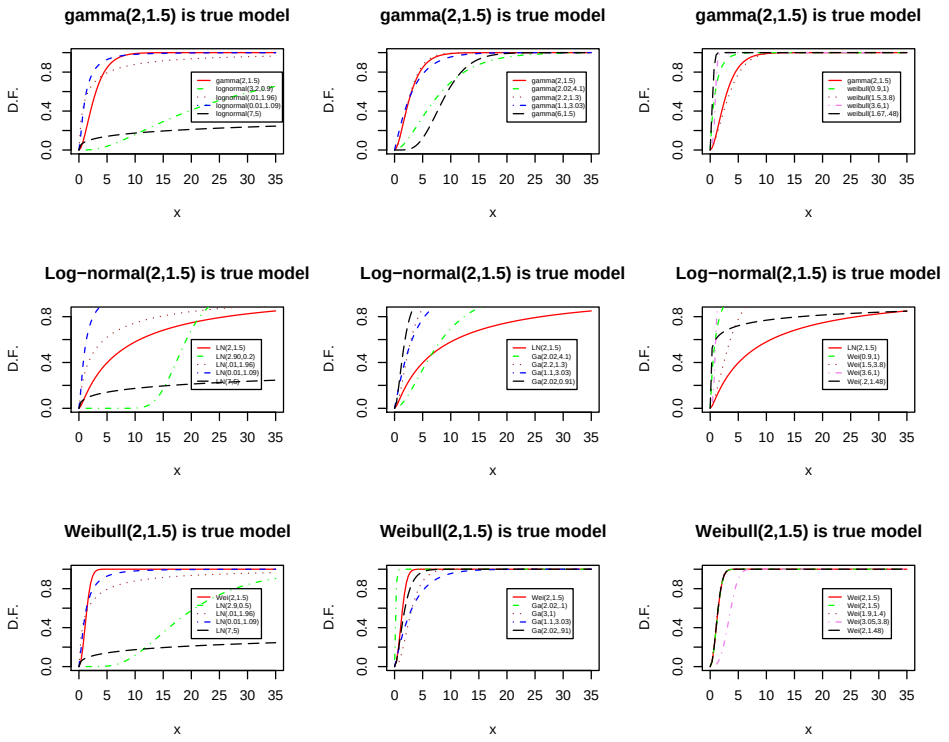

Figure 1. Comparison between true model and some rival models based on their distribution functions.

Considering the cumulative distribution functions of the true and rival models $\mathcal{LN}(\theta, \eta)$, $\mathcal{W}(a, b)$ and $\Gamma(\alpha, \beta)$ respectively, with respect to the value in the Table 1. The relative closeness of models is shown in Figure 1.

This figure shows that the selected models contains in the admissible set, are nearer to the true model in compare to the unselected rival models.

4.2 Simulation Study based on the Risk Function

In this situation we consider the coefficients as $\alpha_i \quad i = 1, 2, 3$, $\alpha_i \geqslant 0$ and $\sum_{i=1}^{3} \alpha_i = 1$. We consider two cases. First we consider some known parameters for the true and rival models. In the second case we estimate the parameters of the rival models under the true one. First, consider the true

Table 2. Kullback-Leibler Risks and coefficients of rival and mixed models for Log-normal, Weibull and Gamma distributions.

		$\mathcal{KL}$	Coefficient	
Mm	0.4123			
$\mathcal{LN}$	$(0.01, 1.00)$	0.0150	0.9306	$\mathcal{KL} < 0.4123$
$\mathcal{W}$	$(0.15.0.10)$	1.9907	9.9496×10^{-4}	$\mathcal{KL} > 0.4123$
	$(0.05, 0.40)$	1.8884	0.0051	$\mathcal{KL} > 0.4123$
	$(0.50, 0.12)$	0.4091	0.0365	$\mathcal{KL} < 0.4123$
Γ	$(0.90, 0.09)$	0.9613	0.0268	$\mathcal{KL} > 0.4123$
	$(0.09, 1.00)$	2.5893	3.9577×10^{-6}	$\mathcal{KL} > 0.4123$

model $\mathcal{LN}(0, 1)$ and six rival models as

$$\mathcal{L}(0.01, 1.00), \mathcal{W}(0.15, 0.10), \mathcal{W}(0.05, 0.40), \mathcal{G}(0.50, 0.12), \Gamma(0.90, 0.09), \Gamma(0.09, 1.00),$$

where the first rival model is wellspecified and the others are mis-specified.

So the admissible set of rival models based on the Theorem 2 is

$$\mathcal{G} = \{\mathcal{LN}(0.01, 1.00), \mathcal{W}(0.50, 0.12)\}$$

In the second case we fixed one parameter in each rival model and estimate the other parameters in each model under the log-normal model. Considering $\mathcal{LN}(\mu, \sigma^2)$, $\mathcal{W}(\alpha, \beta_0)$ and $\Gamma(\eta_0, \gamma)$, where β_0 and γ_0 are known and

$$\hat{\alpha} \to \alpha_{\mathcal{LN}} = \exp\left\{\mu + \frac{(\sigma^2)^{1/2}}{2}\right\},$$

$$\hat{\gamma} \to \gamma_{\mathcal{LN}} = \exp\{\mu + \frac{(\sigma^2)}{2}\}.$$

Set $\{\beta, \eta\} = \{(1.00, 3.00), (1.10, 1.30), (1.00, 2.00)\}$. Then

$$(\hat{\alpha}, \hat{\gamma}) = \{(1.65, 1.66)\}.$$

So we consider different situations under which we use known and estimated parameters for rival models to compute the given citerion in Theorem 2. The

Table 3. A comparison between rivals and mixed models for Log-normal, Weibull and Gamma distributions, with estimated parameters.

Model	Parameters	$\mathcal{KL}$	Coefficient	Comparison	Admissible set
Mm		0.0956			
$\mathcal{LN}$	$(5.78 \times 10^{-4}, 0.99)$	0.0012	0.9997	$\mathcal{KL} < 0.0956$	
$\mathcal{W}$	$(1.65, 1.00)$	0.0821	1.9764×10^{-4}	$\mathcal{KL} < 0.0956$	$\{\mathcal{LN}(5.78 \times 10^{-4}, 0.99),$
$\mathcal{G}$	$(3.00, 1.66)$	0.4927	6.6090×10^{-6}	$\mathcal{KL} > 0.0956$	$\mathcal{W}(1.65, 1.00)\}$
Mm		0.1397			
$\mathcal{LN}$	$(5.78 \times 10^{-4}, 0.99)$	0.0025	0.9997	$\mathcal{KL} < 0.1397$	
$\mathcal{W}$	$(1.65, 1.10)$	0.0919	1.8299×10^{-4}	$\mathcal{KL} < 0.1397$	$\{\mathcal{LN}(5.78 \times 10^{-4}, 0.99),$
$\mathcal{G}$	$(1.30, 1.66)$	0.5450	1.5498×10^{-4}	$\mathcal{KL} > 0.1397$	$\mathcal{W}(1.65, 1.10)\}$
Mm		0.1919			
$\mathcal{LN}$	$(5.78 \times 10^{-4}, 0.99)$	2.3714×10^{-4}	0.9997	$\mathcal{KL} < 0.1919$	
$\mathcal{W}$	$(1.65, 1.00)$	0.0796	2.0302×10^{-4}	$\mathcal{KL} < 0.1919$	$\{\mathcal{LN}(5.78 \times 10^{-4}, 0.99),$
$\mathcal{G}$	$(2.00, 1.66)$	0.3013	1.0182×10^{-4}	$\mathcal{KL} > 0.1919$	$\mathcal{W}(1.65, 1.00)\}$

result of simulation is given in Table 3. As we see, in the first row of Table 3, based on column 4 we accept the estimated $\mathcal{LN}$ and $\mathcal{W}$ as two members of the admissible set. It's because of their $\mathcal{KL}$'s related to the $\mathcal{KL}$ divergence of the mixed model, Mm, which is equal to 0.0956. For estimated $\mathcal{G}$ the $\mathcal{KL}$ is 0.4927 which is greater than 0.0956. So we do not accept this model as a member of the admissible set.

5 Real Data Analysis

In this section we analysis a set of real data to how the admissible set described above may be used in practice. We have deliberately kept the analysis as simple as possible in order to illustrate how our approach can be applied in practice. We consider admissible set selection based on the best elements of the rival models. The auto-mpg dataset concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes and the number of observations was 239.

1. mpg: continuous,
2. cylinders: multi-valued discrete,
3. displacement: continuous,
4. horsepower: continuous,

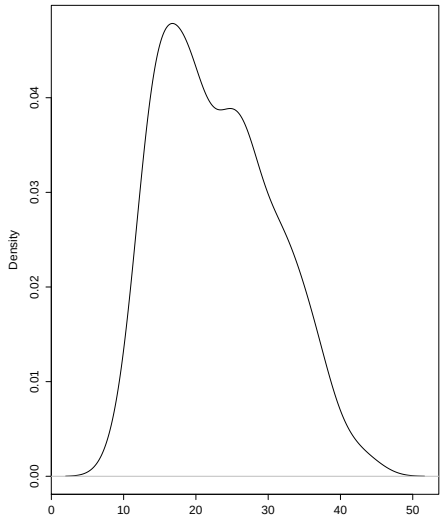

Figure 2. The city-cycle fuel consumption in miles per gallon.

5. weight: continuous,
6. acceleration: continuous,
7. model year: multi-valued discrete,
8. origin: multi-valued discrete,
9. car name: string (unique for each instance).

This data can be found in Itsm data libraries. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. Quinlan (1993). In Proceedings on the Tenth International Conference of Machine Learning. A graph of the auto-mpg dataset is displayed in Figure 2.

It suggests that a Normal mixture model might provide a reasonable model for given data. The EM estimation of parameters are

$$\hat{\mu} = (26.9829, 15.7559), \quad \hat{\sigma}^2 = (6.6482, 2.8301), \quad P = (0.6487, 0.3512).$$

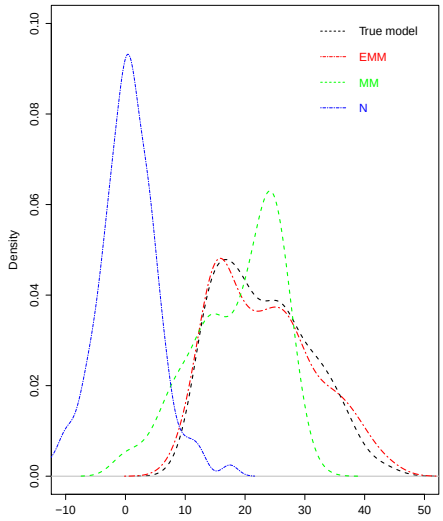

Figure 3. The estimated mixture model, mixture model and Normal curve.

We consider the estimated mixture model, EMM, the mixture model, MM, with parameters

$$\mu = (15, 25), \quad \sigma^2 = (6.6482, 2.8301), \quad P = (0.6487, 0.3512)$$

and Normal distribution, N(0,6.6482), as three competing models. The values of L.H.S. and R.H.S. of competing models are given in Table 4. Here we estimate Π as $\hat{\Pi}_k = 0.4980$ for estimated model, EMM, and $\hat{\Pi}_k = 0.5019$ for mixed model, MM. As we see, $-15.7407 < -3.3925$ and $-15.7407 < -4.0336$, so the models EMM and MM models are belong to the admissible set of models. On the other hand, $-17.2633 < -3.3925$ but $-17.2633 > -38.8590$ which means that only EMM model is belong to the admissible set. The Normal model is not a suitable model as a proposed model to describe the data. Figure 3 confirme these results.

Table 4. Comparison between rival models.

Competing models:	EMM	MM	EMM	N
R.H.S:	-3.3925	-4.0336	-3.3925	-38.8590
L.H.S:	-15.7407		-17.2633	

6 Conclusion

In many disciplines, there are competing explanations of the same phenomena, often characterized by alternative statistical models. When the null hypothesis is nested within the alternative, standard classical procedures can be utilized. But if the null and the alternative hypotheses belong to non-nested families of distributions, classical procedures cannot be applied directly. To testing the non-nested models we use some modified classical tests and criteria. A question which arises is that, how we select some rival family of models. In this paper we investigate some theoretical approach to answer to this essential question in model selection. This paper aims to say how we approach to some family of rival models and select a set of better rival models. This problem is an unsolved problem in model selection. When we have to propose some rival models, the main question is that, which models are suitable to consider as a competing models. It is not true that we set some complete family as the suitable rival models. In this work, we have considered the main question in model selection as, how could infinite set of all possible models for a population, be reduced to a smaller set of models, say admissible set. Then the problem is selecting the best model from the admissible set of models. In this work, we have focused on the finite mixture of the Kullback-Leibler divergence criteria to selecting the best set of the rival models which adopted to address both of model uncertainty and model fidelity. To illustrate our approach, consider the case where we have m (possibly non-nested) rival models to be compared. Each model is a family of distributions which indexed by their parameters. In each family may exists some members which are good estimators for the true model. We want to construct a set contains these kind of distributions. In general the m distributions may have several interesting properties and their probability density functions can take different shapes. Although, most of the times the m distribution functions may provide a similar data fit, but still it is desirable to select the more nearly and simple models. The further work could be based

on the cross validation approach to decide on the coefficient in the mixture.

Acknowledgement

I would like to thank Editor and referees for their valuable comments and suggestions that have greatly improved the presentation of the paper.

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.

Cavanaugh, J.E. (1999). A Large Sample Model Selection Criterion based on Kullback's Symmetric Divergence, *Statistics Probability Letter*, **42**, 333 − 343.

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 Annal of Applied Statistics*, **2**, 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.

Cox, D.R. (1962). Further Result on Tests of Separate Families of Hypothesis. *Journal of the Royal Statistical Society B*, **24**, 406 − 424.

Davidson, R. and MacKinnon, J.G. (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica*, **49**, 781 − 793.

Davidson, R. and MacKinnon, J.G. (2002). Bootstrap J Tests of Nonnested Linear Regression Models. *Journal of Econometrics*, **109**, 167 − 193.

Eguchi, S. and Copas, J. B. (1998). A Class of Local Likelihood Methods and Near- parametric Asymptotics. *J. R. Statist. Soc. B.* **60**, 709–724.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309 − 368.

Fisher, G.R. and McAleer, M. (1981). Alternative Procedures and Associated Tests of Significance for Non-Nested Hypotheses. *Journal of Econometrics*, **16**, 103 − 119.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*. **22**, 79 − 86.

Linhart H. and Zucchini W., (1986). *Model Selection.* Wiley, New York.

Matusitas, K. (1966). A Distance and Related Statistics in Multivariate Analysis. Multivariate Analysis (ed. P. R. Krishnaiah), 187 − 200. Akademic Press, New York.

Matusitas, K. (1967). On the Notion of Affinity of Several Distributions and some of its Applications. *Ann. Inst. Statist. Math.*, **19**, 181 − 192.

Miller, R.G., (1981). Simultaneous Statistical Inference. Springer-Verlag, New York.

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.

Quinlan, R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, University of Massachusetts, Amherst. Morgan Kaufmann, 236-243.

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. (2012a). Tracking Interval for Selecting between Non-Nested Models: An Investigation for Type II Right Censored Data . *Journal of Statistical Planning and Inference.* **142**, 3201 − 3208

Sayyareh, A. (2012b). Inference After Separated Hypotheses Testing: An Investigation for Linear Models. *Journal of Statistical Computation and Simulation.* **82**, 1275 − 1286.

Shimodiara, H. (1998). An Application of Multiple Comparison Techniques to Model Selection. *Annals of Institute Statistical Mathematics*, **50**, 1 − 13.

Shimodaira, H. (2001). Multiple Comparisons of Log-likelihoods and Combining Non-Nested Models with Application to Phylogenetic Tree Selection. *Communication in Statistics*, **30**, 1751 − 1772.

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, **57**, 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.

Zheng, G., Freidlin, B. and Gastwirth, J.L. (2004). Using Kullback-Leibler Information for Model Selection when the Data-Generating Model is Unknown: Applications to Genetic Problems. *Statistica Sinica*, **14**, 1021 − 1036.

Zucchini, W. (2000). An Introduction to Model Selection. *Journal of Mathematical Psychology*, **44**, 41-61.

Abdolreza Sayyareh

Faculty of Mathematics,

K. N. Toosi University of Technology,

Tehran, Iran.

email: *asayyareh@kntu.ac.ir*