A New Upper Bound for Kullback-Leibler Divergence

Abdolreza Sayyareh

Razi University, Kermanshah, Iran
asayyareh@razi.ac.ir

Abstract

Choosing a proper divergence measure is a common task for a statistician and probabilist. The Kullback-Leibler divergence is well known among the information divergence. We introduce an upper bound on the Kullback-Leibler divergence. We show that this upper bound is better than the $\chi^2$ bound. Based on the proposed bound, we find a two sided bound for Shannon’s entropy. A simulation study shows that the new bound works to choose a proper model among the rival models. It is illustrated on the logistic regression and some non-nested models.

Mathematics Subject Classification: 60E15, 62E99

Keywords: Convex functions, Kullback-Leibler divergence, $\chi^2$-distance, Measures of divergence, Shannon’s entropy

1 Preliminaries

One important aspect of statistical modeling 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. In the literature there are many criteria to evaluate the best model. Two of them are $\chi^2$ distance and Kullback-Leibler, say, $\mathcal{KL}$ divergence. There is an extensive literature regarding lower bounds on the $\mathcal{KL}$ in terms of variational distance, $V$. Most of it concentrates on bounds which are accurate when $V$ is close to zero. It seems the best known result is Pinsker’s inequality $\mathcal{KL} \geq \frac{1}{2} V^2$, due to [1] and [2] and so called because it was Pinsker, [3], who first showed that $D \geq c V^2$ for sufficiently small $c$. Kullback [4, 5] and [6] showed that $D \geq \frac{1}{2} V^2 + \frac{1}{36} V^4$ and Topson in his paper, [7], that $D \geq \frac{1}{2} V^2 + \frac{1}{36} V^4 + \frac{1}{270} V^6 + \frac{231}{340200} V^8$. Recently, Dragomir have contributed a lot of work providing different kinds of upper
bounds on the distance and divergence measures. In this work we aim to construct an upper bound on the Kullback-Leibler divergence better than the $\chi^2$ bound. Throughout this paper, let $\Omega$ denote a measurable space with $\sigma$-algebra $\mathcal{B}$. Let $\mathcal{M}$ be the space of all probability measures on $(\Omega, \mathcal{B})$. In what follows, let $\mu$ and $\nu$ denote two probability measures on $\Omega$. Let $h$ and $g$ denote their corresponding density functions with respect to a $\sigma$-finite dominating measure $\zeta$, then the Kullback-Leibler ($\mathcal{KL}$) [8] divergence based on $h$ and $g$ is defined as
\[
\mathcal{D}_I(\mu, \nu) = \mathcal{KL}(h, g) = \int_{A(\mu)} h \log \frac{h}{g} d\zeta
\]
where $A(\mu)$ is the support of $\mu$ on $\Omega$. This definition is independent of the choice of dominating measure $\zeta$. For $\Omega$ a countable space
\[
\mathcal{D}_I(\mu, \nu) = \sum_{\omega \in \Omega} \mu(\omega) \log \frac{\mu(\omega)}{\nu(\omega)}.
\]
The other usual distance between two densities is the $\chi^2$ distance defined as
\[
\mathcal{D}_{\chi^2}(\mu, \nu) = \int_{A(\mu) \cup A(\nu)} \frac{(h - g)^2}{g} d\zeta
\]
This metric assumes values in $[0, \infty]$. For a countable space $\Omega$ this reduce to
\[
\mathcal{D}_{\chi^2}(\mu, \nu) = \sum_{\omega \in A(\mu) \cup A(\nu)} \frac{(\mu(\omega) - \nu(\omega))^2}{\nu(\omega)}.
\]
This measure defined by Pearson, for some history, see, [9]. The $\chi^2$ distance has not some property of a metric but like the $\mathcal{KL}$ divergence, the $\chi^2$ distance between product measures can be bounded in terms of the distances between their marginals, see, [10]. Developping of some inequalities for Kullback-Leibler divergence and $\chi^2$ distance are introduced in [11] and [12]. Section 2 presents the statistical model and model selection concepts. Section 3 presents a relation between the $\mathcal{KL}$ divergence and the $\chi^2$ distance. Section 4 presents the main problem and introduced a new bound for the $\mathcal{KL}$ measure and illustrate that by two examples. Section 5 presents a simulation study in the framework of the logistic regression and some classical models to model selection. In section 6 we present an application of the new bound to construct two sided bound for Shannon’s entropy.
2 Statistical Models and Model Selection

2.1 Statistical families and statistical models

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 parameterization can be represented by a function from a set \(B\) with values in \(\mathcal{P}: \beta \rightarrow P^\beta\). This parameterization 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 parameterization of this family will called a parameterized statistical model. If we have two parameterized statistical models \(T = (P^\beta, \beta \in B)\) on \(\mathcal{F}_1\) and \(T' = (P^\gamma, \gamma \in \Gamma)\) 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 \((Y, T)\) of a random variable and a parameterized statistical model induce the parameterized family of distributions on \((\mathcal{R}, \mathcal{B})\): \(T_Y = (P^\beta_Y, \beta \in B)\).

Conversely, the pair \((Y, T_Y)\) induce \(T\) 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 \(P^*\). Model selection as apart of the statistical inference aims to approach \(P^*\).

**Definition 2.1** Model \(T\) is well specified if \(P^* \in T\) and is mis-specified otherwise. If it is well specified, then there is a \(\beta^* \in B\) such that \(P^{\beta^*} = P^*\).

2.1.1 Kullback-Leibler Risk

In decision theory, estimators are chosen as minimizing some risk function. The most important risk function is based on the Kullback-Leibler divergence [8]. Let a probability \(P'\) is absolutely continuous with respect to a probability \(P\) and \(\mathcal{F}_1\) a sub-\(\sigma\)-field of \(\mathcal{F}\) the loss using \(P'\) in place of \(P\) is the \(\mathcal{L}_\mathcal{F}^{\mathcal{P}/\mathcal{P}'} = \log \frac{dP}{dP'}\). Its expectation is

\[
E_P\{\mathcal{L}_\mathcal{F}^{\mathcal{P}/\mathcal{P}'}\} = KL(P, P'; \mathcal{F}).
\]

This is the Kullback-Leibler (\(KL\)) 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 \(P\) and \(P'\) respectively we have \(\frac{dP}{dP'} = \frac{f_Y(Y)}{g_Y(Y)}\) and the divergence of the distribution \(P'\) relative to \(P\) can be written as

\[
KL(P, P') = \int \log \frac{f_Y(y)}{g_Y(y)} f_Y(y) d(y).
\]
We have that $KL(P, P'; \mathcal{F}) = KL(P, P')$ if $\mathcal{F}$ is the $\sigma$-field generated by $y$ on $(\Omega, \mathcal{F})$. Based on continuity arguments, we take $0 \log \frac{1}{r} = 0$ for all $r \in \mathbb{R}$ and $t \log \frac{1}{r} = \infty$ for all non-zero $t$. Hence $KL$ divergence takes its value in $[0, \infty]$. The $KL$ divergence is not a metric, but it is additive over marginals of product measures, see, [13]. $KL(h, g^\beta) = 0$ implies that $h = g^\beta$.

2.2 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 $h(.)$. A model is assumed as possible explanation of $Y$, represented by $(g) = \{g(y; \beta), \beta \in B\} = (g^\beta(\cdot))_{\beta \in B}$. This function is known but its parameter as $\beta \in B$ is unknown. The aim is to ascertain whether $(g)$ can be viewed as a family contained $h(.)$ or has a member which is a good approximate for $h(.)$. The log-likelihood loss of $g^\beta$ relatively to $h(.)$ for observation $Y$ is $\log \frac{h(Y)}{g^\beta(Y)}$. The expectation of this loss under $h(.)$, or risk, is the $KL$ divergence between $g^\beta$ and $h(.)$ as

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

Let $\hat{Y} = (Y_1, Y_2, ..., Y_n)$ be identically and independently distributed random variables from unknown density $h(.)$. Two rival models are assumed as possible explanation of $Y$, represented by $(f^\gamma(\cdot))_{\gamma \in \Gamma} = \{f(y; \gamma), \gamma \in \Gamma\}, \Gamma \subset \mathcal{R}^q$ and $(g^\beta(\cdot))_{\beta \in B} = \{g(y; \beta), \beta \in B\}, B \subset \mathcal{R}^p$. 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 $(f^\gamma(\cdot))_{\gamma \in \Gamma}$ and $(g^\beta(\cdot))_{\beta \in B}$ if any can be viewed as a family contained $h(.)$ or has a member which is a good approximate for $h(.)$. As we see, there is no trivial null hypothesis.

**Definition 2.2** (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 $\beta^* \in B$ such that $g^{\beta^*} = h$; otherwise it is misspecified.

We assume that there is a value $\beta_0 \in B$ which minimizes $KL(h, g^\beta)$. If the model is well specified $\beta_0 = \beta^*$; if the model is misspecified, $KL(h, g^\beta) > 0$. The
Quasi Maximum Likelihood Estimator (QMLE), $\hat{\beta}_n$, is a consistent estimator of $\beta_0$, see, [14]. 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,[15]. 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 [16].

### 3 Relation Between $\mathcal{KL}$ Divergence and $\chi^2$ Distance

The following theorem gives an upper bound for $\mathcal{KL}$ divergence based on the $\chi^2$ distance, see, [11].

**Theorem 3.1** The $\mathcal{KL}$ divergence, $D_T(\mu, \nu)$, and the $\chi^2$ distance, $D_{\chi^2}(\mu, \nu)$, satisfy

$$D_T(\mu, \nu) \leq \log[D_{\chi^2}(\mu, \nu) + 1].$$

In particular $D_T(\mu, \nu) \leq D_{\chi^2}(\mu, \nu)$, with equality if and only if $h = g$.

**Proof:** Since log is a concave function, we have

$$E \left( \log\frac{h}{g} \right) \leq \log E \left( \frac{h}{g} \right),$$

then

$$D_T(\mu, \nu) \leq \log \left\{ \int_{\Omega} \frac{h}{g} \, d\zeta \right\},$$

But

$$\int_{\Omega} \frac{(h-g)^2}{g} \, d\zeta = \int_{\Omega} \frac{h^2}{g} \, d\zeta - 1,$$
thus
\[ \mathcal{D}_F(\mu, \nu) \leq \log[\mathcal{D}_{\chi^2}(\mu, \nu) + 1] \leq \mathcal{D}_{\chi^2}(\mu, \nu). \quad (1) \]

### 4 Main Problem

One problem with \( KL \) criterion is that its value has no intrinsic meaning. Investigators commonly display big numbers, only the last digits of which are used to decide which is the smallest. If the specific structure of the models is of interest, because it tells us something about the explanation of the observed phenomena, it may be interesting to measure how far from the truth each model is. But this may not be possible. To do this we can quantify the difference of risks between two models, see, [15] and [17]. The Kullback-Leibler risk takes values between 0 and \(+\infty\) where we tends to some value between 0 and 1 for risks. We may relate the \( KL \) values to relative errors. We will make errors by evaluating the probability of an event \( \varepsilon \) using a distribution \( g, P_g(\varepsilon) \), rather than using the true distribution \( h, P_h(\varepsilon) \).

\[ \text{RE}(P_g(\varepsilon), P_h(\varepsilon)) = (P_h(\varepsilon))^{-1}(P_g(\varepsilon) - P_h(\varepsilon)). \]

To obtain a simple formula relating \( KL \) to the error on \( P_h(\varepsilon) \), we consider the case \( P_h(\varepsilon) = 0 \) and \( g/h \) constant on \( \varepsilon \) and \( \varepsilon^c \). In that case we find \( \text{RE}(P_g(\varepsilon), P_h(\varepsilon)) = \sqrt{2KL(h, g)} \), the approximation being valid for a small \( KL \) value. For \( KL \) values of \( 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1} \) we find that \( \text{RE}(P_g(\varepsilon), P_h(\varepsilon)) \) is equal to 0.014, 0.045, 0.14 and 0.44 errors that we may qualify as negligible, small, moderate and large, respectively. We propose Constructing a set of reasonable rival models is of interest in theory and applications [18] and [19]. As another insight, there is an established relationship between the total variation, Helinger, Kullback-Leibler and \( \chi^2 \) metrics and it is defined as follows:

\[ 0 \leq \sup_{\alpha \in A} | \int dP - dP' | \leq \left( \int [\sqrt{dP} - \sqrt{dP'}]^2 \right)^{1/2} \leq \sqrt{KL(P, P')} . \]

Indeed, we know that

\[ KL(P, P') \leq \log \left( \int \left( \frac{P^2}{P'} - P \right) d\zeta + 1 \right) . \]

Each metric will approach 0 when \( P \sim P' \).

In this work we want to improve the upper bound for the \( KL \) divergence. The better bound will help to find the better estimated model.

#### 4.1 A New Upper Bound for \( KL \) Divergence

To evaluate the \( KL \) loss it would be interesting to establish some upper bound. We propose a shorter bound for \( KL \). We wish to approach the true model.
A new upper bound for Kullback-Leibler divergence

A shorter bound about the KL criterion will help to achieve an appropriate set of models. Therefore, our search will be limited to determining optimal model as a member of the admissible set of models. The following theorem is of interest:

**Theorem 4.1** If \( h \) and \( g \) are two density functions, then

\[
\mathcal{KL}(h, g) \leq \int_{A(\mu)} \left( \frac{h}{g} \right)^h - h \, d\zeta \leq D_{\chi^2}(h, g)
\]

with equality for \( h(.) = g(.) \). This bound is better than the bound introduce in (1)

Proof: We recall the known inequality \( \log x \leq x - 1 \) with \( x = \frac{h}{g} \). Note that

\[
\log \left( \frac{h}{g} \right)^h \leq \left( \frac{h}{g} \right)^h - 1.
\]

Taking integral on both sides, we have

\[
\mathcal{KL}(h, g) \leq \int_{A(\mu)} \left( \frac{h}{g} \right)^h - \int_{A(\mu)} d\zeta = \int_{A(\mu)} \left( \frac{h}{g} \right)^h d\zeta - 1.
\]

We need to show that this bound is better than the bound which is given in (1). In fact we need to show that

\[
\left( \frac{h}{g} \right)^h - 1 \leq \frac{h^2}{g} - h.
\]

Consider the inequality \( x^\rho - 1 \leq \rho(x - 1); \ y > 0, \ \rho \in [0,1] \), with \( x = \frac{h}{g} \) and \( \rho = h \). Then \( \left( \frac{h}{g} \right)^h - 1 \leq h \left( \frac{h}{g} - 1 \right) \). It means that

\[
\int_{A(\mu)} \left( \frac{h}{g} \right)^h - h \, d\zeta \leq \int_{A(\mu)} h \left( \frac{h}{g} \right) d\zeta - 1 \leq D_{\chi^2}(h, g)
\]

which complete the proof.

In discrete case it is easy to see that

\[
\int_{A(\mu)} \left( \frac{h}{g} \right)^h - h \, d\zeta \leq \log \int_{A(\mu)} h \left( \frac{h}{g} \right) d\zeta
\]

implies that

\[
\mathcal{KL} \leq \int_{A(\mu)} \left( \frac{h}{g} \right)^h - h \, d\zeta \leq \log (D_{\chi^2}(h, g) + 1)
\]
In continuous if \( dx \) is an infinitely small number, the probability that \( X \) is included within the interval \((x, x + dx)\) is equal to \( g(x)d(x) \) or

\[
g(x)dx = P(x < X < x + dx)
\]

which let us consider a large \( g(x) \) with very small \( dx \). In this case with \( g(x) \in [0, 1] \) the proposed bound works. It is important to judge whether the values within a bound for \( KL \) loss corresponds to large or small loss. Sometimes we may do that using the relative error. Otherways we will search for a model with small \( KL \) loss. The given bound will help to characterize the more appropriate models among all rival models.

**Example 4.2:**

The random variable \( X \) has an exponential distribution if it has a probability density function of the form

\[
h(x; \sigma, \theta) = \sigma^{-1} \exp \left( -\frac{x - \theta}{\sigma} \right), \quad x > \theta > 0; \sigma > 0.
\]

We take \( \theta = 0 \). This density introduce a family as

\[
\mathcal{H} = \{h(x; \sigma); x > 0; \sigma > 0\}.
\]

We know that for \( \sigma^{-1} < 1 \) and for some \( x > 0 \) the exponential density is belong to \((0, 1]\). Consider \( h(x; \sigma_t) \in \mathcal{H} \) as the true density and \( h(x; \sigma_c) = g(x; \sigma_c) \in \mathcal{H} \) as the candidate density. For each fixed \( \sigma_t \) we consider some \( \sigma_c \) and compare the new bound and the bound based on the \( \chi^2 \) distance for Kullback-Leibler divergence. The result of simulation is given in Table 1 and Figure 1. The \( KL \) loss of the exponential density hathe same order as the integral of the relative error. For example the value of the \( KL \) \( (h, g) \) for \((\sigma_t, \sigma_c) = (0.9, 0.3), (0.9, 0.6), (0.9, 0.8)\) is equal to the \( 0.6317, 0.0928, 0.0083 \) respectively, which are of order \( 10^{-1}, 10^{-2} \) and \( 10^{-3} \). The similar value for proposed bound has the same order as the \( KL \) loss. But, the order of the \( \chi^2 \) bound is greater than the \( KL \) loss. So, under conditions which mentioned in Theorem 4.1 the new bound is better than the \( \chi^2 \) bound. It indicates that the new bound is more precise than the known \( \chi^2 \) bound to model choice. Only in three situations, when the rival model has a parameters which are far from the parameters of the true model, the first \( \chi^2 \) bound is better than the proposed bound.
Table 1- Comparison between the proposed bound and the bound based on the $\chi^2$ distance for Kullback-Leibler divergence.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\mathcal{KL}(h, g)$</th>
<th>$\int (\frac{h}{g} - h) d\zeta$</th>
<th>$\log { \int \frac{h^2}{g} - h d\zeta + 1 }$</th>
<th>$\int \frac{h^2}{g} - h d\zeta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 0.1$</td>
<td>1.3092</td>
<td>1.5498</td>
<td>1.5612</td>
<td>3.7647</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 0.3$</td>
<td>0.4319</td>
<td>0.6317</td>
<td>0.5878 *not satisfied</td>
<td>0.8000</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 0.6$</td>
<td>0.0721</td>
<td>0.0928</td>
<td>0.1187</td>
<td>0.1250</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 0.8$</td>
<td>0.0067</td>
<td>0.0083</td>
<td>0.0124</td>
<td>0.0125</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 1.1$</td>
<td>0.0213</td>
<td>0.0260</td>
<td>0.0506</td>
<td>0.0519</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 1.2$</td>
<td>0.0457</td>
<td>0.0548</td>
<td>0.1178</td>
<td>0.1250</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 1.3$</td>
<td>0.0767</td>
<td>0.0918</td>
<td>0.2344</td>
<td>0.2642</td>
</tr>
<tr>
<td>$\sigma_t = 0.9, \sigma_c = 1.4$</td>
<td>0.1137</td>
<td>0.1359</td>
<td>0.3691</td>
<td>0.4464</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.3$</td>
<td>0.1236</td>
<td>0.1289</td>
<td>0.1744</td>
<td>0.1905</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.5$</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.7$</td>
<td>0.0635</td>
<td>0.0707</td>
<td>0.1744</td>
<td>0.1905</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.9$</td>
<td>0.2122</td>
<td>0.2358</td>
<td>1.0217</td>
<td>1.7778</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.1$</td>
<td>0.8309</td>
<td>1.0558</td>
<td>1.0217 *not satisfied</td>
<td>1.7778</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.2$</td>
<td>0.3163</td>
<td>0.3817</td>
<td>0.4463</td>
<td>0.5625</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.3$</td>
<td>0.1105</td>
<td>0.1289</td>
<td>0.1744</td>
<td>0.1905</td>
</tr>
<tr>
<td>$\sigma_t = 0.5, \sigma_c = 0.4$</td>
<td>0.0231</td>
<td>0.0263</td>
<td>0.7138</td>
<td>1.0417</td>
</tr>
</tbody>
</table>
Figure 1: Comparison between the KL, the new bound and the $\chi^2$ bound

5 Simulation study

5.1 Logistic Model

We consider a simulation study where we have to select between different logistic regression models. We consider i.i.d. sample of size $n$ of triples $Y_i, x_{1i}, x_{2i}, i = 1, 2, ..., n$, from the following distribution. The conditional distribution of $Y_i$ given $(x_{1i}, x_{2i})$ was logistic with $\text{logit}[f_{Y|X}(1|x_{1i}, x_{2i})] = 1.5 + x_{1i} + 3x_{2i}$ where $f_{Y|X}(1|x_{1i}, x_{2i}) = P_h((Y = 1|x_{1i}, x_{2i}))$, the marginal of $(x_{1i}, x_{2i})$ were bivariate normal with zero expectation and variance equal to the identity matrix. We consider $h$ as true model and consider two nonnested families of parametric densities $\mathcal{G} \in \mathcal{M}$ and $\mathcal{F} \in \mathcal{M}$ as $\mathcal{G} = (g^\beta(.)|_{\beta \in B}, \mathcal{F} = (f^\gamma(\cdot))_{\gamma \in \Gamma}$ as rival models, where

$$\mathcal{G} = \{g(\cdot, \beta): \mathcal{R} \rightarrow \mathcal{R}^+; \beta \in B \subseteq \mathcal{R}^d\} = (g^\beta(.)|_{\beta \in B}.$$
We considered model \((g^\beta(.))_{\beta \in B}\) as
\[
\logit[p_{Y|X}(1|x_{1i}, x_{2i})] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},
\]
which was well specified and the mis-specified model
\[
(f^\gamma(.))_{\gamma \in \Gamma} = \logit[f_{Y|X}(1|x_{1i}, x_{2i})] = \gamma_0 + \sum_{k=1}^{2} \gamma_k x_{1ik} + \gamma_3 x_{2i},
\]
where \(x_{1ik}\) were dummy variables indicating in which categories \(x_{1i}\) fell, the categories were defined using terciles of the observed distribution of \(x_1\), and this was represented by two dummy variables \(x_{1i1}\) indicating whether \(x_{1i}\) fell in the first tercile or not, \(x_{1i2}\) indicating whether \(x_{1i}\) fell in the second tercile or not. Using reduced model, we can use the \(KL\) divergence for conditional models. The precise estimate of \(KL\) \((h, f^{\gamma_0})\), \(\int_{A(\mu)}[(\frac{h}{f^{\gamma_0}})^h - 1] d\zeta\), \(\log(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta + 1)\) and \(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta\), by the empirical value of the three criteria computed on \(10^5\) replicas. The \(\gamma_0\) is a value which minimizes \(KL\) \((h, f^\gamma)\). Their values are \(KL\) \((h, f^{\gamma_0}) = 0.0049, \int_{A(\mu)}[(\frac{h}{f^{\gamma_0}})^h - 1] d\zeta = 0.0056, \log(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta + 1) = 0.0064\) and \(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta = 0.0065\). It seen that \(KL\) \((h, f^{\gamma_0}) \leq \int_{A(\mu)}(\frac{h}{f^{\gamma_0}})^h d\zeta - 1 \leq \log(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta + 1) \leq \int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta\).

For the well-specified model \(G\), we know that \(KL\) \((h, g^{\beta_0}) = 0\). We can compute a precise estimate of the expected value of \(KL\) \((h, f^{\gamma_0})\), \(\int_{A(\mu)}[(\frac{h}{f^{\gamma_0}})^h - 1] d\zeta\), \(\log(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta + 1)\) and \(\int_{\Omega}(\frac{h^2}{f^{\gamma_0}} - h) d\zeta\), by replacing the parameters by their estimates. We generate 1000 replications from the above model, we find the estimates of the criteria as \(0.0058, 0.0111, 0.0140, 0.0141\) and \(0.0009, 0.0103, 0.0116, 0.1165\) for \(n = 200\) and \(n = 500\), respectively. By these values, the expected order of the criteria is confirmed again. Then the new bound will give the better choice between the rival models.

### 5.1.1 Classical Models

Consider the data generating probability as \(Weibull(\alpha, \beta), W_t\), and two rival models as \(Gamma(\eta, \theta), G_r\), as a mis-specified model and the \(Weibull(3.2, 4.2), W_r\). The gamma and the Weibull models are non-nested. We generate \(10^4\) Monte-Carlo data sets of sample sizes \(n = 100\). For each iteration of the given sample size, we compute the maximum likelihood estimates of the parameters of the gamma density, the \(KL\) divergence between the true model and the rival models, the \(\chi^2\) distance and the proposed criterion. Table 2 and Figure 2 show the results of the simulation for
\[
(\alpha, \beta) = \{(1.9, 1.2), (2.3, 1.4), (2.5, 1.6), (3, 3), (3, 4)\}
and \((\hat{\eta}, \hat{\theta})\).

Table 2- Results of simulation for model selection based on the proposed bound for gamma distribution as the rival model.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>(1.9, 1.2)</th>
<th>(2.3, 1.4)</th>
<th>(2.5, 1.6)</th>
<th>(3, 3)</th>
<th>(3, 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(KL(W_t, W_r))</td>
<td>0.3941</td>
<td>0.5090</td>
<td>0.4771</td>
<td>0.1453</td>
<td>0.0000</td>
</tr>
<tr>
<td>(KL(W_t, G_r))</td>
<td>1.4343</td>
<td>1.6873</td>
<td>1.8751</td>
<td>2.3355</td>
<td>2.1982</td>
</tr>
<tr>
<td>(\int f\left(\frac{h_{W_t}}{g_{G_r}}\right)^{h_{W_t}} h_{W_t} d\zeta)</td>
<td>3.9264</td>
<td>3.9352</td>
<td>3.9296</td>
<td>3.9554</td>
<td>4.0552</td>
</tr>
<tr>
<td>(\log\left{\int f\left(\frac{h_{W_t}}{g_{G_r}}\right)^{h_{W_t}} h_{W_t} d\zeta + 1\right})</td>
<td>6.7181</td>
<td>4.5223</td>
<td>10.2651</td>
<td>9.4141</td>
<td>6.5211</td>
</tr>
</tbody>
</table>

In Table 2, we denote the true density and the rival model by \(h_{W_t}\) and \(g_{G_r}\) respectively. The variability of \(\chi^2\) criteria is greater than the variability of new bound and always the upper based on the proposed bound is smaller than that for the known \(\chi^2\) bound.

Figure 2: Results of simulation based on the proposed bound for Weibull and lognormal distribution in non-nested case

6 Application

Example 6.1:

Let \(h(x), g(x)\), for \(x \in \Upsilon\) be two probability mass functions (instead of \(\mu\) and \(\nu\), respectively), \(g(x) = \frac{1}{\text{Card}(\Upsilon)}\) and \(\text{Card}(\Upsilon)\) is finite. Taking into account
A new upper bound for Kullback-Leibler divergence

that

$$KL(h, g) = \sum_{x \in \Upsilon} h(x) \log h(x) - \log \frac{1}{\text{Card}(\Upsilon)},$$

which reduce to

$$KL(h, g) = \log \text{Card}(\Upsilon) - \sum_{x \in \Upsilon} h(x) \log \frac{1}{h(x)},$$

where $\sum_{x \in \Upsilon} h(x) \log \frac{1}{h(x)}$ is known as the Shannon’s entropy. In addition

$$D_{\chi^2}(h, g) = \text{Card}(\Upsilon) \sum_{x \in \Upsilon} h^2(x) - 1.$$

The proposed upper bound for $KL$ divergence will be

$$\sum_{x \in \Upsilon} \left( \left[ \frac{h(x)}{g(x)} \right]^{h(x)} - 1 \right) = \sum_{x \in \Upsilon} [\text{Card}(\Upsilon) h(x)]^{h(x)} - \sum_{x \in \Upsilon} 1.$$

We aim to show that

$$\sum_{x \in \Upsilon} [\text{Card}(\Upsilon) h(x)]^{h(x)} - \sum_{x \in \Upsilon} 1 \leq \text{Card}(\Upsilon) \sum_{x \in \Upsilon} h^2(x) - 1. \quad (2)$$

Because of the condition $0 < h < 1$ we obtain that $\text{Card}(\Upsilon) h^{h(.)} \leq \text{Card}(\Upsilon) h^2(.)$, on the other hand $\sum_{x \in \Upsilon} 1 \geq 1$ which imply that (2) holds.

6.1 Two sided bound for Shannon’s entropy

In Theorem 4.1 we saw

$$D_{\chi^2}(\mu, \nu) \leq \log[D_{\chi^2}(\mu, \nu) + 1] \leq D_{\chi^2}(\mu, \nu),$$

on the other hand [11] show that

$$D_{\chi^2}(\mu, \nu) \geq \frac{1}{2} \left( \sum_{x \in \Upsilon} |h(x) - g(x)| \right)^2,$$

this two inequality state that

$$0 \leq \frac{1}{2} \left( \sum_{x \in \Upsilon} |h(x) - g(x)| \right)^2 \leq D_{\chi^2}(\mu, \nu) \leq D_{\chi^2}(\mu, \nu).$$

Following Example 4.1 and using Theorem 4.1, we have

$$0 \leq \frac{1}{2} \left( \sum_{x \in \Upsilon} |h(x) - g(x)| \right)^2 \leq D_{\chi^2}(\mu, \nu) \leq \int_{\mathcal{A}(\mu)} \left( \left( \frac{h}{g} \right)^h - h \right) d\xi.$$
After simplification we will find a bound for Shannon’s entropy, 

\[ \mathcal{H}(X) = \sum_{x \in \Upsilon} h(x) \log \frac{1}{h(x)}, \]

as 

\[ \mathcal{H}(X) \in \left[ \log |\Upsilon| - \sum_{x \in \Upsilon} [|\Upsilon|h(x)]^{h(x)} - \sum_{x \in \Upsilon} 1, \ \log |\Upsilon| - \frac{1}{2} (\sum_{x \in \Upsilon} |h(x) - g(x)|)^2 \right] \]

where \(|\Upsilon| = \text{Card}(\Upsilon)|.\]

7 Discussion

In this paper we attempted to a better upper bound for the \(KL\) divergence which has been widely used in the field of model selection, information theory and statistical signal processing. An important problem in model selection is; how select an admissible set of the rival models. The better bound for \(KL\) divergence helps to construct a more precise set of rival models to make a better decision. Naturally this bound lets to construct a bound for Shannon’s entropy. This bound will be comparable with the bounds which introduced by the others statistician and probabilists, for example, see, [11].

References


Received: April, 2011