A generalized predictive criterion for model selection

Mario Trottini and Fulvio Spezzaferri

Carnegie Mellon University, Pittsburgh, PA (USA)
and Università di Roma "La Sapienza" (Italy)

Abstract: Given a sample of n i.i.d observations generated from some unknown distribution F, assume that F belongs to one of two parametric models M_1, M_2, and that the estimation of the density of a future observation is of interest. San Martini and Spezzaferri (1984) proposed for this problem a predictive criterion based on the Kullback Leibler divergence between the estimate and the density of the future observation. In this paper a generalization of this criterion is presented both using a general class of divergences and relaxing the assumption that the true model belongs to M_1 or M_2.

Key words: Model Selection, Gaussian Process, Loss Function, α-divergences.

1. Introduction

Let X be a random variable with p.d.f $f(x)$, $x \in \mathcal{X}$, and let $\underline{x} = (x_1, x_2, \ldots, x_n)$ be a sample of n i.i.d observations generated from X. Let assume that $f(.)$ belongs to one of two parametric models:

$$f \in \mathcal{M} = \{M_1,\, M_2\},$$
$$M_i : \ f_i(x|\theta_i),\ \ \theta_i \in \Theta_i \subset \Re.$$

The problem that we are concerned with is that of choosing a model for predictive purposes. In particular if we denote with X_{n+1} the future observation and with $f_{x_{n+1}}(\cdot)$ its density, the problem that we want to face is that of choosing the model which gives the 'best' estimation of the density of the future observation $f_{x_{n+1}}(\cdot)$. This is a decision problem which first requires the choice of model M_i from $\mathcal{M}$ and then, assuming M_i to be the model, requires an estimation of $f_{x_{n+1}}(\cdot)$. Denote with $\hat{P}(\cdot)$ the generic estimate of $f_{x_{n+1}}(\cdot)$. The utility of $\hat{P}(\cdot)$ depends on an unknown 'state of the world' ω, $\omega \in \Omega$, and is measured by a given utility function that we denote by $u(\hat{P}(\cdot),\, \omega)$. Assigning a prior distribution on Ω, the standard Bayesian solution to this problem consists of two steps:

Step 1: finding the 'best' estimation of $f_{x_{n+1}}(\cdot)$ conditionally to each model M_i, say $\hat{P}_i(\cdot)$,

$$\hat{P}_i(x) = \arg \max_{P(\cdot)\in\mathcal{F}} E_{\omega|\underline{x},M_i}[u(P(\cdot),\, \omega)],$$

where $\mathcal{F}$ is the class of all density functions with support $\mathcal{X}$ and $E_{\omega|\underline{x},M_i}$ denotes the expectation with respect to the posterior distribution of ω if M_i is true;

Step 2: choosing the model which maximizes the expected utility, where the expected utility of M_i is defined as:

$$\bar{u}(M_i) = E_{\omega|\underline{x}}[u(\hat{P}_i(\cdot),\, \omega)]$$

and $E_{\omega|\underline{x}}$ denotes the expectation with respect to the posterior distribution of ω.

Given the nature of the problem that we are interested in, it seems natural to assume as state of the world the density of the future observation and defining the loss that we pay when we estimate the true density $f_{x_{n+1}}$ with some distribution P, as a measure of discrepancy between $f_{x_{n+1}}$ and P. In this paper we propose to use as measure of divergence one of the class of α-divergences discussed by Aly and Silvey (1966) and Csiszár (1967). This class of divergences contains many measures of divergence commonly used to compare distributions, as the Kullback Leibler divergence, the Hellinger distance or the Chi square divergence. In different problems the "natural " measure of divergence for the problem might be different. Depending on the problem, in fact, we might want to penalize those estimate P that give a poor estimate $f_{x_{n+1}}$ in the tails, or those which underestimate the variability of X_{n+1} or those which overestimate such variability. The class of the α-divergences in this respect seems to offer a class of loss function general enough to cover a large number of interesting applications. In addition, following the decisional framework described above the use the α-divergences as loss function leads to a predictive criterion, that we will refer to as *generalized predictive criterion*, which has a very nice interpretation. The expression of the predictive criterion is in fact the product of the posterior odds and a term that allow us to take into account the error that we make when we choose the wrong model. In Section 2 we derive the expression of the generalized predictive criterion and we show in Section 4 that it includes as a special case the predictive criterion of San Martini and Spezzaferri, that is obtained from the generalized predictive criterion using as α-divergence the Kullback Leibler divergence. The consistency and some other properties of the generalized predictive criterion are presented in Section 3. In Section 5 we discuss and outline solutions for several problems that arise with the generalized predictive criterion as: (i) the interpretation of the α-divergence as a function of α, (ii) the dependency of the criterion on the Bayes factor (when only weak prior information is available, the criterion inherits all the problems that usually affect the Bayes factor as indeterminacy and lack of robustness). The computation of the α-divergences although relevant is not addressed in this work. In Section 6 we show that the performances of the predictive criterion strongly depend on the assumption that the true model belongs to the class $\mathcal{M}$. However we do not know whether the true model belongs to $\mathcal{M}$, unless $\mathcal{M}$ is the class of all possible

models. Therefore we are forced to considerably enlarge the class $\mathcal{M}$. But the bigger is the class considered, the more complex are the models to compare, and smaller is the discriminatory power of the selection criterion. As a possible compromise between needs of enlarging $\mathcal{M}$ and choosing between simple models, in Section 6 we consider a bigger class of models, $\mathcal{M}'$ say, that contains $\mathcal{M}$ and we propose comparing the models in $\mathcal{M}$, assuming however that the true model belongs to $\mathcal{M}'$.
As an important special case we develop this procedure in Section 7 for the comparison of two normal linear models, generalizing a previous work of Gutiérrez-Penã in which the regression function is assumed to be a realization of a Gaussian process centered around a simple parametric model. Finally, an application of this procedure is shown in Section 8.

2. The generalized predictive criterion

Given two densities $g_1(\cdot)$, $g_2(\cdot)$, the α-divergences of $g_2(\cdot)$ from $g_1(\cdot)$ is defined as follows:

$$D_\alpha(g_2(\cdot)|g_1(\cdot)) = E_1(f_\alpha(\frac{g_2(\cdot)}{g_1(\cdot)}))$$

where E_1 denotes expectation with respect to $g_1(\cdot)$ and f_α is a convex function defined $\forall\, \alpha \in \Re$ as:

$$f_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2} \cdot (1 - z^{\frac{1+\alpha}{2}}) & \text{if } |\alpha| \neq 1 \\ z \cdot \log z & \text{if } \alpha = 1 \\ -\log z & \text{if } \alpha = -1 \end{cases}$$

Aly and Silvey (1966) and Csiszár (1967) introduced the class of the α-divergences, $\mathcal{D}_\alpha$, as a class of measure of discrepancy between distributions,

$$\mathcal{D}_\alpha = \{D_\alpha(\cdot|\cdot),\ \alpha \in \Re\}.$$

This class contains many measures of divergence commonly used for comparing distributions. For example $\alpha = 0$ corresponds to the Hellinger distance ($D_{\alpha=0}$=2·Hellinger distance), $\alpha = 3$ to the Chi-square divergence ($D_{\alpha=3}$=0.5·Chi-square divergence) and $\alpha = -1$ to the Kullback Leibler divergence.
Assume as state of the world ω the density of the future observation $f_{x_{n+1}}$. Therefore, $\Omega = \mathcal{M}$ and the prior distribution on Ω is defined by assigning both the prior probability π_i of M_i and the prior distribution $\pi_i(\theta_i)$ on Θ_i, i=1,2. Define now the loss that we pay for estimating the true density $f_{x_{n+1}}$ with some distribution P as α-divergence between P and $f_{x_{n+1}}$, $l(P(\cdot), f_{x_{n+1}}) = D_\alpha(P(\cdot)|f_{x_{n+1}})$. Using the decisional framework described above at the first step we have to find the 'optimal' estimation of $f_{x_{n+1}}(\cdot)$ under each model. In this case the 'optimal' estimation under model M_i, $i = 1, 2$ is that density, say $\hat{P}_{\alpha\, i}(\cdot)$, which is closest to M_i with respect to the α-divergence, that is the density which minimizes the expected α-divergence from $f_i(\cdot|\theta_i)$;

$$\hat{P}_{\alpha\, i}(\cdot) = \arg \min_{P_i(\cdot) \in \mathcal{F}} E_{\theta_i|\underline{x}}(D_\alpha(P_i(\cdot)|f_i(\cdot|\theta_i)),$$

where $E_{\theta_i|\underline{x}}$ denotes the expectation with respect to the posterior distribution of θ_i. From now on we will refer to $\hat{P}_{\alpha\, i}(\cdot)$ as *generalized Bayesian predictive distribution* or as *Bayesian α-predictive distribution.* It can be proved (Corcuera and Giummolé 1997) that:

$$\hat{P}_{\alpha\, i}(x) \;=\; \begin{cases} c_{\alpha\, i}\;\cdot\;[\int f_i(x|\theta_i)^{\frac{1-\alpha}{2}}\;\cdot\;\pi_i(\theta_i|\underline{x})d\theta_i]^{\frac{2}{1-\alpha}} & \text{if } |\alpha| \neq 1 \\ c_{1\, i}\;\cdot\;\exp\{\int \log f_i(x|\theta_i)\;\cdot\;\pi_i(\theta_i|\underline{x})d\theta_i\} & \text{if } |\alpha| = 1 \end{cases}$$

where $c_{1\, i}$ and $c_{\alpha\, i}$ are normalizing constants. As the previous formula shows, $\hat{P}_{\alpha\, i}(x)$ is a generalization of the standard Bayesian predictive distribution under model M_i. $\hat{P}_{\alpha\, i}(x)$ is in fact proportional to an L_p norm of the $f_i(x|\theta_i)$ with $p = \frac{1-\alpha}{2}$ and setting $\alpha = -1$ we find the standard Bayesian predictive distribution. Note also that there is an interesting correspondence between some values of α, some common measures of divergence and some characteristic values of p. In particular for $\alpha = -1$, which corresponds to the Kullback Leibler divergence, $\hat{P}_{\alpha\, i}(\cdot)$ is proportional to the *simple* average of the $f_i(x|\theta_i)$, for $\alpha \to 1$, which corresponds to the Shannon information, $\hat{P}_{\alpha\, i}(\cdot)$ is proportional to the *geometric* average of the $f_i(x|\theta_i)$, while for $\alpha = 3$, which corresponds to the Chi-square divergence, $\hat{P}_{\alpha\, i}(\cdot)$ is proportional to the *harmonic* average of the $f_i(x|\theta_i)$.
Now, going through the second step and minimizing the expected loss of each model, we have the following *generalized predictive criterion*: model M_1 is preferred to model M_2 if and only if:

$$C_\alpha \cdot B_{12} \cdot \frac{\pi_1}{\pi_2} \;>\; 1. \tag{1}$$

where C_α is given by:

$$C_\alpha = \frac{E_{\theta_1|\underline{x}}[D_\alpha(\hat{P}_{\alpha\, 2}(\cdot)|f_1(\cdot|\theta_1)) - D_\alpha(\hat{P}_{\alpha\, 1}(\cdot)|f_1(\cdot|\theta_1))]}{E_{\theta_2|\underline{x}}[D_\alpha(\hat{P}_{\alpha\, 1}(\cdot)|f_2(\cdot|\theta_2)) - D_\alpha(\hat{P}_{\alpha\, 2}(\cdot)|f_2(\cdot|\theta_2))]}, \tag{2}$$

- $\hat{P}_{\alpha,i}(\cdot)$ is the Bayesian α - predictive distribution under model M_i, $i = 1,\, 2$;
- π_i is the prior probability of model M_i being true, $i = 1,\, 2$;
- B_{12} is the Bayes factor in favor of M_1:

$$B_{12}(\underline{x}) = \frac{\int_{\Theta_1} f_1(\underline{x}|\theta_i)\cdot\pi_1(\theta_1)d\theta_1}{\int_{\Theta_2} f_2(\underline{x}|\theta_i)\cdot\pi_2(\theta_2)d\theta_2}.$$

Note that each term that appears in the left side of (1) has a specific interpretation. We have the prior odds ($\frac{\pi_1}{\pi_2}$) which represents a measure of the *pre-experimental evidence* in favor of M_1; the Bayes factor which represents the *experimental evidence* in favor of M_1 (see Section 5.2 for further comments on the Bayes factor); the ratio C_α, which takes into account, in selecting a model, the error that we make choosing the wrong model. To better understand the interpretation of this last term let us have a closer look at formula (2). Note first that given two random variables reciprocally absolutely continuous with density g_1 and g_2 the α-divergence of g_2 from g_1, might be interpreted as the error that we make when we approximate the density g_1 with the density g_2. In the numerator of (2), therefore the expectation

$$E_{\theta_1|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f_1(\cdot|\theta_1))],$$

might be interpreted as the error that we make when the true model belongs to M_1 but we choose the wrong model M_2 and report $\hat{P}_{\alpha\,2}(\cdot)$ as estimate of $f_{x_{n+1}}(\cdot)$. Similarly the expectation,

$$E_{\theta_1|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f_1(\cdot|\theta_1))]$$

might be interpreted as the error that we make when the true density is some $f_1(x|\theta_1)$ in M_1 but, ignoring the true value of θ_1 we report $\hat{P}_{\alpha\,1}(\cdot)$ as estimate of $f_1(x|\theta_1)$. The numerator of (2) therefore is nothing but the expected additional loss that we pay when the true model belongs to M_1 but we choose the wrong modelM_2 and we report $\hat{P}_{\alpha\,2}(\cdot)$ as inference about $f_1(x|\theta_1)$ instead of $\hat{P}_{\alpha\,1}(\cdot)$. Similarly the denominator of (2) can be seen as the expected additional loss that we pay when the true model belongs to M_2 but we choose the wrong model M_1 and we report $\hat{P}_{\alpha\,1}(\cdot)$ as inference about $f_2(x|\theta_2)$ instead of $\hat{P}_{\alpha\,2}(\cdot)$. Therefore if the two models have equal posterior probability model M_1 is preferred to model M_2 if and only if $C_\alpha > 1$, i.e if the additional loss that we pay when we choose incorrectly M_2 is bigger than the additional loss that we pay when we choose incorrectly M_1. It is interesting to note that the term C_α which take into account the error that we make when we choose the wrong model is expressed in terms of the same measure of divergence that we use to define the optimal distribution under each model. In this sense the generalize predictive criterion is coherent.

3. Properties of the generalized predictive criterion

In this section we explore the properties of the above generalized predictive criterion. All the properties that we describe are derived under the following assumptions:

1. For every pair of probability measures G_1 e G_2 in M_1 and M_2 respectively, we assume that G_1 and G_2 are reciprocally absolutely continuous.

2. Define $\mathcal{A}$ and $\mathcal{B}$ as follows:

   $\mathcal{A} = \{\alpha \in \Re | P_{\alpha\,1}, P_{\alpha\,2} \text{ are defined}\}$;

   $\mathcal{B} = \{\alpha \in \Re | D_\alpha(P|f_i(x|\theta_i)) < \infty \, \forall \text{ P with the same support of } f_i(x|\theta_i)\ i = 1, 2\}$.

   We assume :

   $$\alpha \in \mathcal{E} = \mathcal{A} \cap \mathcal{B};$$

3. The prior probability for the competiting models π_1 e π_2, are both strictly positive.

Under the following set of assumptions we have that:

- The generalized predictive criterion is consistent.

  Dim. The expression of the generalized predictive criterion is the product of three terms, the prior odds, the Bayes factor and the C_α. Assumptions 2 and 3 guarantee that both the prior odds and C_α are finite. The consistency of the generalize predictive criterion follows therefore from the consistency of the Bayes factor (Gelfand and Dey, 1994).

- Let $\hat{P}_\alpha(x)$ be the estimate of the density of the future observation based on the all available prior information, when we use as loss function an α-divergence. It is easy to show that

$$\hat{P}_\alpha(x) = \arg \min_{P(\cdot) \in \mathcal{F}} E_{\theta,\phi|\underline{x}}(D_\alpha(P(\cdot)|f(\cdot|\theta,\phi));$$

$$\hat{P}_\alpha(x) \;=\; C \cdot \{\sum_{j=1}^{2} P(M_j|\underline{x}) \cdot c_{\alpha j}^{\frac{1-\alpha}{2}} \cdot \hat{P}_{\alpha j}(x)^{\frac{1-\alpha}{2}}\}^{\frac{2}{1-\alpha}},$$

  where:

  1. C is a normalization constant;
  2. $\phi = (\phi_1, \phi_2)$ is a latent variable, $\phi \in \{\phi^1, \phi^2\}$, $\phi^1 = (1,0), \phi^2 = (0,1)$, such that $\phi = \phi^i$ iff the true model belongs to M_i (i=1,2);
  3. $f(x|\theta,\phi) = \sum_{i=1}^{2} f(x|\theta_i) \cdot \phi_i, \;\; (\theta,\phi) \in (\Theta_1 \times \Theta_2) \times \{\phi^1, \phi^2\}$;
  4. $c_{\alpha j}$ are the normalization constants that appear in the expression of $\hat{P}_{\alpha j}(x)$;
  5. $\mathcal{F}$ is the class of all densities with the same support of $f(x|\theta,\phi)$.

  Then $M_1 \succeq M_2 \iff$

$$D_\alpha(\hat{P}_{\alpha 1}(\cdot)|\hat{P}_\alpha(\cdot)) \leq D_\alpha(\hat{P}_{\alpha 2}(\cdot)|\hat{P}_\alpha(\cdot)).$$

This property shows the relationship between the model selection problem and the estimation problem. Given the prior information that the true model belongs either to M_1 or to M_2 the best estimate of the density of the future observation using as loss function an α-divergence is given by $\hat{P}_\alpha$. However if the choice of a model in $\mathcal{M}$ is required, the generalized predictive criterion choose the model in $\mathcal{M}$ which is closest to $\hat{P}_\alpha$ in terms of the α-divergence used.

4. San Martini and Spezzaferri Criterion

In this section we briefly presents San Martini and Spezzaferri predictive criterion and we show that it can be derived as a special case of the generalized predictive criterion if we use as measure of α-divergence the Kullback Leibler divergence ($\alpha = -1$).

Consider the decisional framework described in the introduction. Assuming as state of the world the future observation ($\omega = x_{n+1}$), and the logarithmic function,

$$u(P(\cdot), x_{n+1}) = A \cdot \log P(x_{n+1}) + B(x_{n+1}),$$

for some positive constant A and function $B(\cdot)$, as the utility associated to $P(\cdot)$ as predictive distribution of x_{n+1}, we obtain the predictive criterion of San Martini and Spezzaferri. In particular we have:

step 1: The optimal distribution conditionally to model M_i (i=1, 2) is the standard Bayesian predictive distribution

$$\hat{P}_i(x) = \int f_i(x|\theta_i) \cdot \pi_i(\theta_i|\underline{x}) d\theta_i;$$

step 2: denoting with $P^*(\cdot)$ the predictive distribution based on the whole available information

$$P^*(x) = \sum_{i=1}^{2} P(M_i|\underline{x}) \cdot \hat{P}_i(x),$$

where $P(M_i|\underline{x})$ is the posterior probability of M_i being true, $i = 1, 2$,we have: $M_1 \succeq M_2 \longleftrightarrow$

$$\int \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} \cdot \hat{P}^*(x) dx \ > \ 0. \tag{3}$$

In the M-closed case (see Bernardo and Smith, 1994) in which we assume that the true model is one of the models under comparison, this is one of the most popular predictive criterion. The next proposition shows that actually the criterion can be derived as a special case of the generalized predictive criterion that we described in Section 2.

Proposition 1.

Given two parametric models M_i, (i=1, 2), the predictive criterion of San-Martini e Spezzaferri is equivalent to the generalized predictive criterion using the Kullback Leibler divergence as α-divergence.

Proof: From (3) we have $M_1 \succeq M_2 \longleftrightarrow$

$$\int \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} \cdot P^*(x) dx \ > \ 0 \longleftrightarrow$$

$$
\begin{aligned}
&P(M_1|\underline{x}) \cdot \int \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} \cdot \hat{P}_1(x)dx + P(M_2|\underline{x}) \cdot \int \log \frac{P_1(x)}{P_2(x)} \cdot P_2(x)dx \; > \; 0 \longleftrightarrow \\
&P(M_1|\underline{x}) \cdot \int \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} \cdot \hat{P}_1(x)dx > P(M_2|\underline{x}) \cdot \int \log \frac{\hat{P}_2(x)}{\hat{P}_1(x)} \cdot \hat{P}_2(x)dx \; \longleftrightarrow \\
&\frac{P(M_1|\underline{x})}{P(M_2|\underline{x})} \cdot \frac{\int \log \frac{\hat{P}_1(x)}{\hat{P}_2(x)} \cdot \hat{P}_1(x)dx}{\int \log \frac{\hat{P}2(x)}{\hat{P}_1(x)} \cdot \hat{P}_2(x)dx} > 1
\end{aligned}
$$

where the last equivalence is obtained noting that the Kullback Leibler divergence is always non negative and is zero iff the two distributions are the same. Proposition 1 follows easily noticing that (i) using the Kullback Leibler divergenge ($\alpha = -1$) the generalized predictive distribution is nothing but the standard Bayesian predictive distribution, (ii) the posterior odds $P(M_1|\underline{x})/P(M_2|\underline{x})$ might be written as product of the Bayes factor, B_{12}, and the prior odds, $\frac{\pi_1}{\pi_2}$,(iii)

$$
\begin{aligned}
&\int \log \frac{\hat{P}_i(x)}{\hat{P}_j(x)} \cdot \hat{P}_i(x)dx = \\
&= \; \int\int [\log \frac{\hat{P}_i(x)}{f_i(x|\theta_i)} + \log \frac{f_i(x|\theta_i)}{\hat{P}_j(x)}] \cdot f_i(x|\theta_i) \cdot \pi_i(\theta_i|\underline{x}) d\theta_i dx = \\
&= \; \int [\int \log \frac{f_i(x|\theta_i)}{\hat{P}_j(x)} \cdot f_i(x|\theta_i)dx - \int \log \frac{f_i(x|\theta_i)}{\hat{P}_i(x)} \cdot f_i(x|\theta_i)dx] \cdot \pi_i(\theta_i|\underline{x})d\theta_i = \\
&= \; E_{\theta_i|\underline{x}}[D_{KL}(\hat{P}_j(\cdot)|f_i(\cdot|\theta_i)) - D_{KL}(\hat{P}_i(\cdot)|f_i(\cdot|\theta_i))]\,. \;\; \clubsuit
\end{aligned}
$$

5. Problems

The implementation of the generalized predictive criterion presents several problems: (i) the interpretation of the α-divergences, (ii) the dependency of the criterion on the Bayes factor; when only weak prior information is available, the generalized predictive criterion inherits all the problems that usually affect the Bayes factor (luck of robustness, indeterminacy).

In sections 5.1 and 5.2 we outline some solutions to these problems. Note also that the computation of the α-divergence is crucial for the implement of the generalized predictive criterion. Vajda (1987) shows some interesting results for the computation of the α-divergence between two distributions which belong two the same exponential family. In the more general and interesting case, however, when the competing models do not belong to an exponential family, or at least not to the same exponential family, the expression of the α-divergence cannot be derived in closed form and numerical approximation are needed even if we use conjugate priors. The problem of the computation of the α-divergence, however, is not addressed in this work.

5.1 Interpretation of the α-divergences

The generalized predictive criterion is derived under the assumption that the loss that we pay for estimating the true model with some distribution P can be adequately represented as an α-divergence between the true model and our generic estimate P, for some value of α in $\Re$. The interpretation of the α-divergences therefore plays a crucial role in the use of the generalized predictive criterion in applications.
In this Section we present a first attempt to interpret the α-divergences. In particular we consider two examples, that we think might represent a much larger class of problems for which the α-divergences seem to be a very general class of loss functions of straightforward interpretation. In the two examples we assume that both the "true" model and our generic estimate belong to the same parametric family. In the first example (example A) this is the family of normal distribution with known mean, while in the second example (example B) this is the family of normal distribution with known variance. The results found in the normal case (in terms of interpretation of α-divergences) also hold when the parametric family is the double exponential, the exponential and we believe that these results could be extended to all the scale families (regarding example A) and to all the location families (regarding example B).

Example A

Assume that the true distribution, P^*, is normal with mean μ and variance $\tilde{\theta}$,

$$P^*(x) = N(\mu, \tilde{\theta}),$$

but we report as estimate of P^* a normal distribution $P_\theta(x)$ with the same mean μ but with variance θ,

$$P_\theta(x) = N(\mu, \theta).$$

Using an α-divergence as a loss function the loss that we pay for estimating P^* with $P_\theta(x)$ that we denote with $l_\alpha(P_\theta(x), P^*(x))$ is given by:

- α in (-1,1),

$$l_\alpha(P_\theta(x), P^*(x)) =$$

$$= D_\alpha(P_\theta(x)|P^*(x)) = \frac{4}{1-\alpha^2} \cdot [1 - (\frac{2 \cdot (\frac{\tilde{\theta}}{\theta})^{\frac{1+\alpha}{2}}}{1 + \alpha \cdot \frac{\tilde{\theta}}{\theta} + 1 - \alpha})^{\frac{1}{2}}], \ \forall\, \theta, \tilde{\theta} > 0$$

- $\alpha < -1$

$$l_\alpha(P_\theta(x), P^*(x)) =$$

$$= D_\alpha(P_\theta(x)|P(^*(x)) = \begin{cases} \frac{4}{1-\alpha^2} \cdot [1 - (\frac{2 \cdot (\frac{\tilde{\theta}}{\theta})^{\frac{1+\alpha}{2}}}{1+\alpha \cdot \frac{\tilde{\theta}}{\theta} + 1 - \alpha})^{\frac{1}{2}}] & \text{if } \frac{\tilde{\theta}}{\theta} < \frac{\alpha-1}{\alpha+1} \\ \infty & \text{if } \frac{\tilde{\theta}}{\theta} \geq \frac{\alpha-1}{\alpha+1} \end{cases}$$

- $\alpha > 1$

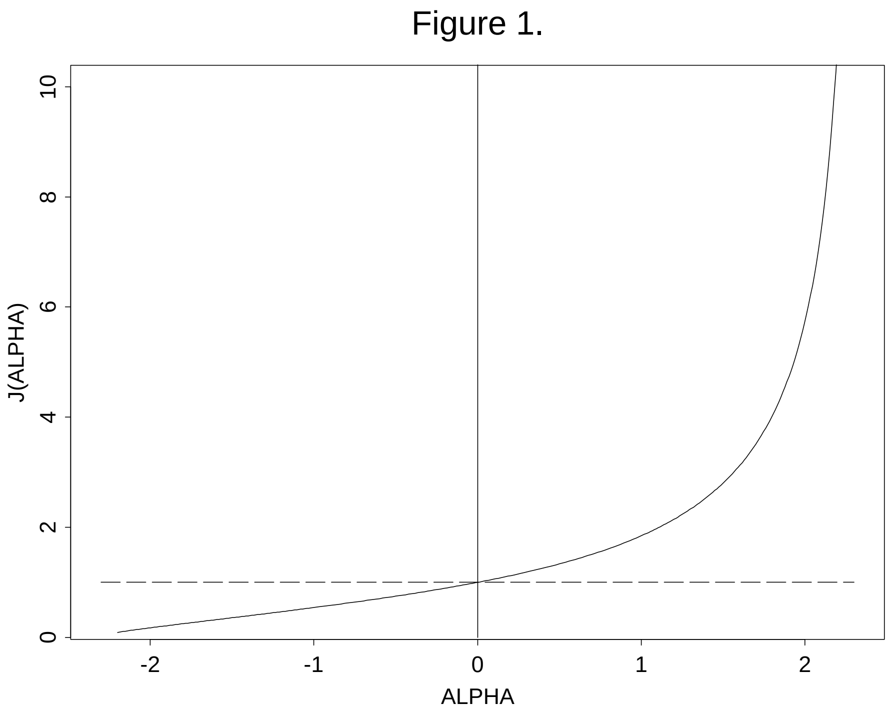
$$l_\alpha(P_\theta(x), P^*(x)) =$$

$$= D_\alpha(P_\theta(x)|P^*(x)) = \begin{cases} \infty & \text{if } \frac{\tilde\theta}{\theta} \leq \frac{\alpha-1}{\alpha+1} \\ \frac{4}{1-\alpha^2} \cdot [1 - (\frac{2\cdot(\frac{\tilde\theta}{\theta})^{\frac{1+\alpha}{2}}}{1+\alpha\cdot\frac{\tilde\theta}{\theta}+1-\alpha})^{\frac{1}{2}}] & \text{if } \frac{\tilde\theta}{\theta} > \frac{\alpha-1}{\alpha+1}. \end{cases}$$

First of all note that $D_\alpha(P_\theta(x)|P^*(x))$ represents a "natural" measure of the discrepancy between $P_\theta(x)$ and $P^*(x)$ since it depends only on the ratio $R = \tilde\theta/\theta$ and as we expect, it is an increasing function of $|\,1-\tilde\theta/\theta\,|$. In addition consider a positive constant k, with $k > 1$ and two arbitrary values for θ, $\theta_1 = k\tilde\theta$ and $\theta_2 = \tilde\theta/k$. The ratio J_α between the α-divergence of $P_{\theta_1}(x)$ from $P^*(x)$ and the α-divergence of $P_{\theta_2}(x)$ from $P^*(x)$,

$$J_\alpha = \frac{D_\alpha(P_{\theta_1}(x)|P^*(x))}{D_\alpha(P_{\theta_2}(x)|P^*(x))},$$

is nothing but the ratio between the loss that we pay when we *overestimate* $\tilde\theta$ by a factor k and the loss that we pay when we *underestimate* $\tilde\theta$ by a factor k. Figure 1 shows the behavior of J_α when α ranges in $(-2.3, 2.3)$.



If $\alpha = 0$ we have $J_\alpha = 1$, and we penalize in the same way overestimating and underestimating the scale parameter $\tilde{\theta}$ (symmetry). For $\alpha > 0$, instead, $J_\alpha > 1$ and we penalize more overestimating than underestimating $\tilde{\theta}$. Note that this penalty could be arbitrarily small (the closer α is to zero the closer we are to the symmetry case) or arbitrarily large (as α increases J_α diverges). Similarly for $\alpha < 0$ we have $J_\alpha < 1$ and we penalize more underestimating than overestimating $\tilde{\theta}$ and again this penalty could be arbitrarily small (for α close to zero) or arbitrarily large (for α which goes to minus infinity).

Example B

Suppose that the true distribution, P^* , is normal with mean $\tilde{\theta}$ and variance σ^2,

$$P^*(x) = N(\tilde{\theta}, \sigma^2),$$

but we report as an estimate of P^* a normal distribution $P_\theta(x)$ with the same variance σ^2 but mean θ,

$$P_\theta(x) = N(\theta, \sigma^2).$$

Using a generic α-divergence as loss function, the loss that we pay when we estimate P^* with $P_\theta(x)$ is given by:

$$l_\alpha(P_\theta(x), P^*(x)) =$$

$$= D_\alpha(P_\theta(x)|P^*(x)) = \begin{cases} \frac{4}{1-\alpha^2} \cdot [1 - \exp[\frac{\alpha^2-1}{8 \cdot \sigma^2} \cdot (\theta - \tilde{\theta})^2]] & \text{if } \alpha \neq -1, 1 \\ \frac{(\theta-\tilde{\theta})^2}{2 \cdot \sigma^2} & \text{if } \alpha = -1, 1 \end{cases}$$

Note that also in this case $D_\alpha(P_\theta(x)|P^*(x))$ represents a "natural" measure of the discrepancy between $P_\theta(x)$ and $P^*(x)$ since it depends only on $(\theta - \tilde{\theta})^2$ and it is an increasing function of $(\theta - \tilde{\theta})^2$. Figure 2 shows the behavior of $D_\alpha(P_\theta(x), P^*(x))$ when $\delta = (\theta - \tilde{\theta})$ ranges in $(-20, 20)$ and different values of α in (-1,1) are used.

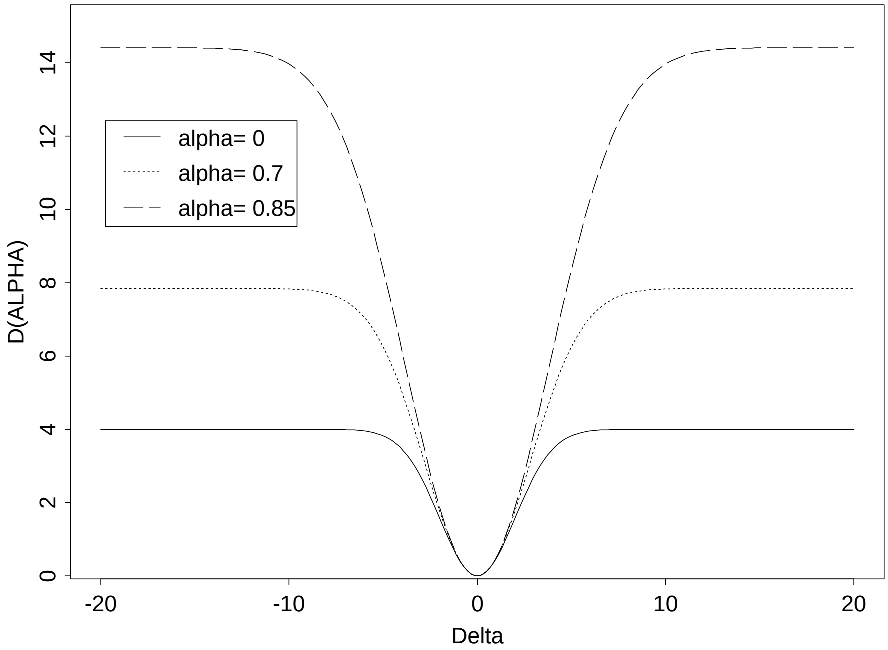
For values of α in $(-1, 1)$ (see Figure 2) we have a loss function with horizontal asymptote (4/(1-α^2)). The loss that we pay when we estimate P^* with $P_\theta(x)$ is an increasing function of $|\,\tilde{\theta} - \theta\,|$ and we can define a positive constant, r_α, such that the difference between the loss associated with two generic estimates $P_\theta(x)$ and $P_{\theta'}(x)$ for which $|\,\tilde{\theta} - \theta\,| > r_\alpha$, $|\,\tilde{\theta} - \theta'\,| > r_\alpha$, is negligible.

When instead $|\alpha| > 1$ we have a quite different picture (the cases $\alpha < -1$ and $\alpha > 1$ are equivalent since $D_\alpha(P_\theta(x)|P^*(x))$ is symmetric with respect to α). Figure 3 shows the behavior of $D_\alpha(P_\theta(x)|P^*(x))$ when $\delta = (\theta - \tilde{\theta})$ ranges in $(-5, 5)$ and different values of α in $(1, \infty)$ are used.

When $|\alpha|$ is greater than one (see Figure 3) the corresponding loss function is convex. In addition the bigger is α in absolute value, the more we penalize large deviations from the true value $\tilde{\theta}$.

For these two examples the class of α-divergences is a class of loss functions: (i) reasonable, as they are increasing function of quantities ($\tilde{\theta}/\theta$, $|\tilde{\theta} - \theta|$) which represent

Figure 2



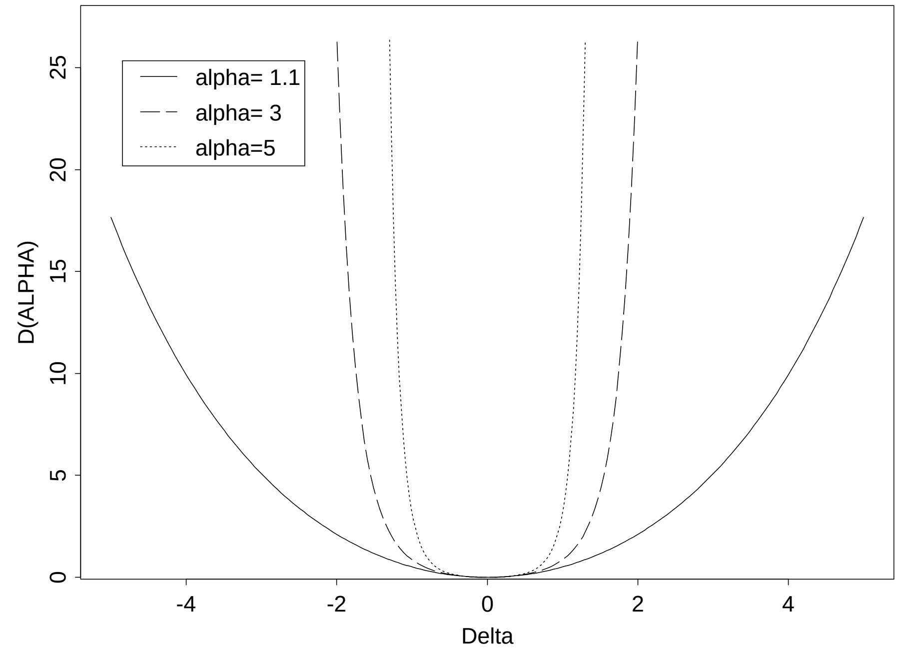
a natural measure of divergence between P_θ and $P_{\tilde{\theta}}$, (ii) easy to interpret, (iii) general enough, since, as we showed, different choices of α allow us to formalize very different opinions about the loss that we want to be paid for an estimate P_θ of the true density $P_{\tilde{\theta}}$.
Of course the most interesting case is the one in which we compare distributions that do not belong to the same parametric family. The α-divergences, in fact are used twice in the generalized predictive criterion: (1) to derive the "optimal " distribution conditionally to each model, (2) to evaluate the expected loss associated with each model. In (1) we have to solve a problem of minimization over the space of all possible distributions, and there is not reason at this stage to restrict our attention to a special set of parametric families as the scale families or the location families. The two examples that we presented, however, represents a first step in the interpretation of the α-divergences for their use in applications.

5.2 Generalize predictive criterion, Bayes factor and weak prior information

Consider two parametric models M_1, M_2

$$M_i : \; f_i(x|\theta_i), \;\; \theta_i \in \Theta_i \subset \Re \;\; i = 1, 2\,,$$

Figure 3



and let $\pi_1(\theta_1)$ e $\pi_2(\theta_2)$ the prior distributions on Θ_1 e Θ_2 respectively. Given a sample $\underline{x} = (x_1, x_2, \ldots, x_n)$ of n observations, the Bayes factor, $B_{12}(\underline{x})$ is defined as:

$$B_{12}(\underline{x}) = \frac{\int_{\Theta_1} f_1(\underline{x}|\theta_i) \cdot \pi_1(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(\underline{x}|\theta_i) \cdot \pi_2(\theta_2) d\theta_2}, \tag{4}$$

$B_{12}(\underline{x})$ represents the coefficient for which we have to multiply the prior odds in favor of M_1 to obtain the posterior odds and it can be interpreted as a measure of the experimental evidence in favor of model M_1. "Large" values of $B_{12}(\underline{x})$ supports the choice of model M_1 while "small" values of $B_{12}(\underline{x})$ supports the choice of model M_2 (see Jeffrey's (1961) for the calibration of the Bayes factor).

Several problem arise for the use of the Bayes factor when only weak prior information is available (see, for istance, O'Hagan, 1995 and Berger and Pericchi, 1996). In particular it can be shown that the Bayes factor is not well defined when at least one of the prior distributions for the parameters of the two competiting models is improper. Improper priors, in fact, are defined but a multiplicative constant and when we evaluate $B_{12}(\underline{x})$ these constants do not simplify.

A common approach in this case is to use partial Bayes factors (see Berger and Pericchi, 1996, and O'Hagan, 1995). The idea is to consider a partition of the original sample $\underline{x}$ in two subsamples $\underline{x}(l)$, $\underline{x}(n-l)$ of size l and $n-l$ respectively, $\underline{x} = (\underline{x}(l), \underline{x}(n-l))$,such that using $\underline{x}(l)$ as a *training sample* we obtain proper posterior distributions $\pi_i(\theta_i|\underline{x}(l))$,

$$\pi_i(\theta_i|\underline{x}(l)) = \frac{f_i(\underline{x}(l)|\theta_i) \cdot \pi_i(\theta_i)}{\int_{\Theta_i} f_i(\underline{x}(l)|\theta_i) \cdot \pi_i(\theta_i) d\theta_i} \quad i = 1, 2 .$$

The Bayes factor is then computed using the remaining part of the data, $\underline{x}(n-l)$, and as new prior distributions $\pi_i(\theta_i|\underline{x}(l))$, (i=1,2). The *partial Bayes factor* is therefore defined as:

$$B_{12}(l) = \frac{\int_{\Theta_1} f_1(\underline{x}(n-l)|\theta_1) \cdot \pi_1(\theta_1|\underline{x}(l)) d\theta_1}{\int_{\Theta_2} f_2(\underline{x}(n-l)|\theta_2) \cdot \pi_2(\theta_2|\underline{x}(l)) d\theta_2}. \tag{5}$$

Note that $B_{12}(l)$, is well defined even when the prior distributions are improper. In fact it is easy to show that formula (5) can be rewritten as follows:

$$B_{12}(l) = B_{12}(\underline{x})/B_{12}(\underline{x}(l))$$

and using improper priors the multiplicative constant which multiplies $B_{12}(\underline{x})$ and $B_{12}(\underline{x}(l))$ simplify.

In order to eliminate the dependency of $B_{12}(l)$ on the choice of the particular training sample Berger and Pericchi (1996) suggested to consider an average of the partial Bayes factor corresponding to all possible training samples with fixed size "l". These averages are called *intrinsic Bayes factors.* A simple alternative to the partial Bayes factor, which avoid to average over the space of all possible training samples, is the *fractional Bayes factor* proposed by O'Hagan (1995). The idea is to use, instead of a training sample, a fraction of the likelihood $f_i(\underline{x}|\theta_i)^{\frac{l}{n}}$. As for the partial Bayes factor, we use $f_i(\underline{x}|\theta_i)^{\frac{l}{n}}$ to obtain proper posterior distributions fo θ_i:

$$\frac{f_i(\underline{x}|\theta_i)^{\frac{l}{n}} \cdot \pi_i(\theta_i)}{\int_{\Theta_i} f_i(\underline{x}|\theta_i)^{\frac{l}{n}} \cdot \pi_i(\theta_i) d\theta_i} \quad , i = 1, 2 \ .$$

With the remaining part of the likelihood $f_i(\underline{x}|\theta_i)^{1-\frac{l}{n}}$, compute the Bayes factor, using as priors the above posteriors. The fractional Bayes factor is therefore defined as

$$\begin{aligned} B_{12}^F(l) &= \frac{\int_{\Theta_1} \ f_1(\underline{x}|\theta_1) \cdot \pi_1(\theta_1) d\theta_1}{\int_{\Theta_2} \ f_2(\underline{x}|\theta_2)\} \cdot \pi_1(\theta_2) d\theta_2} \cdot \frac{\int_{\Theta_2} \{f_2(\underline{x}|\theta_2)\}^{\frac{l}{n}} \cdot \pi_2(\theta_2) d\theta_2}{\int_{\Theta_1} \{f_1(\underline{x}|\theta_1)\}^{\frac{l}{n}} \cdot \pi_1(\theta_1) d\theta_1} \qquad (6) \\ &= B_{12}(\underline{x})/B_{12}^l \end{aligned}$$

where:

$$B_{12}^l = \frac{\int_{\Theta_1} \{f_1(\underline{x}(l)|\theta_1)\}^{\frac{l}{n}} \cdot \pi_1(\theta_1) d\theta_1}{\int_{\Theta_2} \{f_2(\underline{x}(l)|\theta_2)\}^{\frac{l}{n}} \cdot \pi_1(\theta_2) d\theta_2}.$$

From formula (1) we have that when only weak prior information is available the indeterminacy of the Bayes factor implies the indeterminacy of the generalized predictive criterion. However it is easy to show that the solutions proposed for the Bayes factor when only weak prior information is available, can be extended without difficulty to the generalized predictive criterion. This is shown in the following two propositions:

Proposition 2

Let $\phi = (\phi_1, \phi_2)$ a latent variable, such that $\phi = \phi^i$ if and only if the true model belongs to M_i (i=1, 2) with $\phi^1 = (1,0)$, $\phi^2 = (0,1)$.

Under the assumption that the true model belongs to one of the two competeting models, the likelihood function $f(\underline{x}|\theta,\phi)$ is given by:

$$f(\underline{x}|\theta,\phi) = \sum_{i=1}^{2} f(\underline{x}|\theta_i)\cdot\phi_i, \;\; (\theta,\phi)\in(\Theta_1\times\Theta_2)\times\{\phi^1,\phi^2\}.$$

Consider in addition a partition of the observed sample, $\underline{x} = (\underline{x}(l), \underline{x}(n-l))$, where $\underline{x}(l)$ is a proper training sample, *i.e:*

$$\int_{\Theta_i} f_i(\underline{x}(l)|\theta_i)\cdot\pi_i(\theta_i)d\theta_i < \infty \;\; i = 1,2.$$

Using $\underline{x}(l)$ to update the prior distributions for θ e ϕ and obtain proper posterior distributions $\pi(\theta|\underline{x}(l))$, $P(\phi|\underline{x}(l))$ and the remain part of the data $\underline{x}(n-l)$ to compute the generalized predictive criterion using as new priors $\pi(\theta|\underline{x}(l))$ and $P(\phi|\underline{x}(l))$ we have: $M_1 \succeq M_2 \longleftrightarrow$

$$\frac{E_{\theta_1|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f_1(\cdot|\theta_1)) - D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f_1(\cdot|\theta_1))]}{E_{\theta_2|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f_2(\cdot|\theta_2)) - D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f_2(\cdot|\theta_2))]}\cdot B_{12}(l)\cdot\frac{\pi_1}{\pi_2} > 1, \tag{7}$$

where $B_{12}(l)$ is the partial Bayes factor defined in (5) and $\hat{P}_{\alpha\,i}$ (i=1,2) are the Bayesian α - predictive distributions introduced in Section 2.

In order to eliminate the dependency of (7) on the particular training sample we can average (7) over all possible $K \le \binom{n}{l}$ proper training samples of fixed size l. The expression of the corresponding predictive criterion can be easily obtained replacing in (7) the partial Bayes factor with the intrinsic Bayes factor corresponding to the particular average used. Expression (7), in fact, depend on the training sample only through $B_{12}(l)$ whose average is an intrinsic Bayes factor.

Proposition 3

Use a fraction of the whole likelihood $f(\underline{x}|\theta,\phi)^{\frac{l}{n}}$ to obtain proper posteriors for θ e ϕ and the remaining part of the likelihood $f(\underline{x}|\theta,\phi)^{1-\frac{l}{n}}$ to compute the generalized *predictive criterion using as priors the above posteriors. We have that* $M_1 \succeq M_2 \longleftrightarrow$

$$\frac{E_{\theta_1|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f_1(\cdot|\theta_1)) - D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f_1(\cdot|\theta_1))]}{E_{\theta_2|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f_2(\cdot|\theta_2)) - D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f_2(\cdot|\theta_2))]} \cdot B_{12}^F(l) \cdot \frac{\pi_1}{\pi_2} > 1. \tag{8}$$

where $B_{12}^F(l)$ *is the fractional Bayes factor defined in (6).*

Note that (7), (8) have the same structure of (1) with the Bayes factor replaced by the partial Bayes factor and the fractional Bayes factor respectively.

6. A semiparametric approach

Both the expected loss of each model and the predictive criterion strongly depend on the assumption that the true model belongs to the class $\mathcal{M}$. It can be shown, for example, that when this assumption is not satisfied the criterion can choose asymptotically the model which is farthest from the true model with respect to the α-divergence used. Asymptotically, in fact, the generalized predictive criterion is equivalent to the Bayes factor. As the number of the observations increases the Bayes factor chooses with probability one the model in $\mathcal{M}$ which is closest to the true model with respect to the Kullback Leibler divergence (Dmochowski, 1995), and so does the predictive criterion. However in general the α-divergences do not induce the same ordering among models, and the model closest to the true model in terms of Kullback Leibler divergence might be the farthest from the true model with respect to another α-divergence and the incoherent behavior of the generalized predictive criterion is observed. In general we do not know whether the true model belongs to $\mathcal{M}$, unless $\mathcal{M}$ is the class of all possible models. Therefore we are forced to considerably enlarge the class $\mathcal{M}$. But the bigger is the class considered, the more complex are the models to compare, and the smaller is the discriminatory power of the selection criterion. As a possible compromise between needs of enlarging $\mathcal{M}$ and choosing between simple models, we consider a bigger class of models, $\mathcal{M}'$ say, and we propose comparing the models in $\mathcal{M}$, assuming however that the true model belongs to $\mathcal{M}'$. Suppose $\mathcal{M}'$ is defined as follows:

$$\mathcal{M}' = \{f(x|\theta), \theta \in \Theta\},$$

where the parameter space Θ might also have infinite dimension. Following the decisional framework described before the expected loss associated to each model is given by:

$$\bar{l}(M_i) = E_{\theta|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f(\cdot|\theta)) \; i = 1, 2,$$

and we have: $M_1 \succ M_2 \Leftrightarrow$

$$E_{\theta|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,1}(\cdot)|f(\cdot|\theta)) < E_{\theta|\underline{x}}[D_\alpha(\hat{P}_{\alpha\,2}(\cdot)|f(\cdot|\theta)),$$

where as usually, $E_{\theta|\underline{x}}$ denotes the expectation with respect to the posterior distribution of θ and $\hat{P}_{\alpha i}$ is the generalized predictive distribution for model M_i. As an important special case in the next Section we develop this procedure for the comparison of two normal linear models, generalizing a previous work of Gutiérrez - Penã.

7. Normal linear models

Assume that the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are generated from the regression model:

$$y_i = z(x_i) + \epsilon_i \;\; i = 1, 2, \ldots, n$$

where x_i are design points the $\{\epsilon_i\}$ are i.i.d. random variables with $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$, σ^2 known and z(x) is an unknown regression function. Suppose that given the prior information two possible explanations of the functional relationship between the expected value of the dependent variable y and the variable x, are proposed:

$$\begin{aligned} H_1 : \; z(x) &= h_1^t(x)\beta^{(1)}, \\ H_2 : \; z(x) &= h_2^t(x)\beta^{(2)}, \end{aligned}$$

where h_j represent vectors of smooth function of dimension q_j (j=1,2) and $\beta^{(j)} \in \mathcal{R}^{q_j}$ (j=1,2) represent the corresponding vectors of unknown parameters. The goal is to choose one of the two hypotheses for predictive purpose. In particular let us assume that we want estimate the density of Y when x takes a given value x^*. The hypotheses H_1 e H_2 about the regression function define corresponding hypotheses (models) about the distribution of Y at $x = x^*$:

$$\begin{aligned} M_1 : f_1(y|, x^*, \beta^{(1)}) &= N(h_1^t(x^*)\beta^{(1)}, \sigma^2), \beta^1 \in \mathcal{B}^{(1)}, \\ M_2 : f_2(y|, x^*, \beta^{(2)}) &= N(h_2^t(x^*)\beta^{(2)}, \sigma^2), \beta^2 \in \mathcal{B}^{(2)}. \end{aligned}$$

The elicitation of the models M_j is completed assigning prior distribution on the vectors of parameters $\beta^{(j)}$, that we denote with $\pi_j(\beta^{(j)})$ (j=1,2).

In the elicitacion process, however the models M_j (j=1,2) represent only a simplified representation of the phenomenon under study. In other words it is natural to think of M_j as a "representative" of a class of models $M_j^{'}$ "centered" around M_j. Instead of assuming that the distribution of y at x, belongs to $\mathcal{M} = \bigcup_{j=1}^{2} M_j$ we therefore assume that it belongs to the larger class $\mathcal{M}' = \bigcup_{j=1}^{2} M_j^{'}$. The idea is therefore to compare the simple models in $\mathcal{M}$, assuming however that the true model belongs to $\mathcal{M}'$.

In the case of normal linear models that we are examining the classes $M_j^{'}$ $(j = 1, 2)$ can be defined in a very natural way using the following hierarchical procedure:

I Stage: for each model $M_j^{'}$, conditionally on $z_j(\cdot)$, the distribution of Y at x is given by:

$$\tilde{f}_j(y|x, z_j(\cdot)) = N(z_j(x), \sigma^2) \;\; j = 1, 2;$$

II Stage: given $\beta^{(j)}$, the regression function $z_j(\cdot)$ is a realization of a Gaussian process centered around $h_j^t(\cdot)\beta^{(j)}$, (j=1,2):

$$z_j(\cdot)|\beta^{(j)} \sim GP(h_j^t(\cdot)\beta^{(j)},\, W_j(\cdot,\cdot)) \;\; j = 1, 2,$$

where $GP(h_j^t(\cdot)\beta^{(j)},\, W_j(\cdot,\cdot))$ denotes a Gaussian process with mean function $h_j^t(\cdot)\beta^{(j)}$ and covariance function $W_j(\cdot,\cdot)$ (j=1,2).

III Stage: normal prior distributions are assumed for the vectors of parameters $\beta^{(j)}$:

$$\pi_j(\beta^{(j)}) = N(b_j, B_j) \;\; j = 1, 2,$$

where $b_j \in \mathcal{R}^{q_j}$ and B_j is a $q_j \times q_j$ symmetric positive definite matrix.

The covariance functions W_j (j=1,2) control the degree of smoothness of the estimates of $z_j(\cdot)$. A possible choice is the stationary Gaussian form (O'Hagan 1978, 1992):

$$W_j(x^{'}, x'') = \tau_j^2 \cdot \exp -\gamma_j(x^{'} - x'')^2.$$

The smaller is γ_j the higher is the degree of smoothness. The use of a Gaussian process to formalize the prior uncertainty about the regression function has many advantages. First of all the choice of the mean function allow us to center the process around parametric models easy to interpret. In addition, with an appropriate choice of the covariance function, we can specify the degree of deviation from the regression function and we are able to include in the enlarged models $M_j^{'}$ (j=1,2) regression functions whose functional forms might be very complex but that still are in a neighborhood of a simple parametric model (the mean function). From a computational point of view, the use of Gaussian process greatly simplify the analysis. The posterior distribution of a Gaussian process, in fact, is still a Gaussian process. Integrating the process at stage II with respect to the priors $\pi_j(\beta^{(j)})$ defined on stage III we obtain that the distribution of $z_j(\cdot)$ is still a Gaussian process

$$z_j(\cdot) \sim GP(h_j^t(\cdot)b_j,\, W_j(\cdot,\cdot) + h_j^t(\cdot)B_j h_j(\cdot)) \; j = 1, 2,$$

and given the sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the posterior distribution of $z_j(\cdot)$ is a Gaussian process with mean function $\mu_j(\cdot)$ and covariance function $\Lambda_j(\cdot,\cdot)$ (O'Hagan, 1992),

$$z_j(\cdot)|\underline{y} \sim GP(\mu_j(\cdot), \Lambda_j(\cdot,\cdot)) \; j = 1, 2, \tag{9}$$

with:

$$\begin{aligned}
\mu_j(x) &= h_j^t(x)\hat{\beta}_j + w_j^t(x)(W_j(\underline{x}) + \sigma^2 I_n)^{-1}(\underline{y} - L_j\hat{\beta}_j) \;\; j = 1,2; \\
\Lambda_j(x^{'}, x^{''}) &= W_j(x^{'}, x^{''}) - w_j^t(x^{'})(W_j(\underline{x}) + \sigma^2 I_n)^{-1} w_j^t(x^{''}) + \\
&\quad + S_j^t(x^{'})(L_j^t(W_j(\underline{x}) + \sigma^2 I_n)^{-1} L_j + B_j^{-1})^{-1} S_j(x^{''}) \;\; j = 1,2;
\end{aligned}$$

where:

$$\begin{aligned}
\hat{\beta}_j &= (L_j^t(W_j(\underline{x}) + \sigma^2 I_n)^{-1} L_j + B_j^{-1})^{-1}(L_j^t(W_j(\underline{x}) + \sigma^2 I_n)^{-1}\underline{y} + B_j^{-1} b_j) \; j = 1,2; \\
S_j(x) &= h_j(x) - (L_j^t(W_j(\underline{x}) + \sigma^2 I_n)^{-1} w_j(x) \; j = 1,2;
\end{aligned}$$

I_n is the identity matrix of order n;

$$L_j = \begin{pmatrix} h_j^t(x_1) \\ h_j^t(x_2) \\ \cdot \\ \cdot \\ \cdot \\ h_j^t(x_n) \end{pmatrix}; \quad w_j(x) = \begin{pmatrix} W_j(x_1, x) \\ W_j(x_2, x) \\ \cdot \\ \cdot \\ \cdot \\ W_j(x_n, x) \end{pmatrix};$$

$$W_j(\underline{x}) = \begin{pmatrix} W_j(x_1,x_1) & W_j(x_1,x_2) & \cdot & \cdot & \cdot & W_j(x_1,x_n) \\ W_j(x_2,x_1) & W_j(x_2,x_2) & \cdot & \cdot & \cdot & W_j(x_2,x_n) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ W_j(x_n,x_1) & W_j(x_n,x_2) & \cdot & \cdot & \cdot & W_j(x_n,x_n) \end{pmatrix} \quad j = 1,2.$$

Note that from (9) it follows that the distribution of $z_j(\cdot)$ at $x = x^*$ is normal with mean $\mu_j(x^*)$ and variance $\Lambda_j(x^*, x^*)$,

$$z_j(x^*)|\underline{y} \sim N(\mu_j(x^*), \Lambda_j(x^*, x^*)) \;\; j = 1,2. \tag{10}$$

Similarly quite straightforward is the computation of the posterior probabilities of models $M_j^{'}$. Denoting with π_j the prior probability of the model $M_j^{'}$, (j=1,2) and with

$\pi(z_j(\underline{x}))$ the distribution of the random vector $(z(x_1), z(x_2), \ldots, z(x_n))$, we have:

$$\begin{aligned} P(M_j^{'}|\underline{y}) &\propto \pi_j \cdot \int \tilde{f}_j(\underline{y}|z_j)(\underline{x}) \cdot \pi(z_j(\underline{x}))dz_j(\underline{x}) \\ &= \pi_j \cdot N(L_j b_j, W_j(\underline{x}) + L_j B_j L_j^t + \sigma^2 I_n)(\underline{y}) \end{aligned} \tag{11}$$

where $N(\mu, \Sigma)(\underline{y})$ represents a multivariate normal density with mean vector μ and covariance matrix Σ evaluated at $\underline{y}$. Finally using Gaussian processes the expression of the generalized predictive criterion can be obtained in closed form.

Step 1 : The generalized predictive distribution of y at $x = x^*$ for the model M_j, is given by:

$$\hat{P}_{\alpha j}(y) = N(h_j^t(x^*)b_j^{'}, \sigma^2 + \frac{(1-\alpha)}{2} \cdot h_j^t(x^*)B_j^{'}h_j(x^*)),\ j = 1, 2;$$

where:

$$\begin{aligned} b_j^{'} &= B_j^{'} \cdot (\sigma^{-2} L_j^t \underline{y} + B_j^{-1} b_j); \\ (B_j^{'})^{-1} &= \sigma^{-2} L_j^t L_j + B_j^{-1}. \end{aligned}$$

Step 2 $M_1 \succ M_2 \Leftrightarrow$

$$\begin{aligned} &P(M_1^{'}|\underline{y}) \cdot E_{z_1(x^*)|\underline{y}}[D_\alpha(\hat{P}_{\alpha\,2}(y)|\tilde{f}(y|z_1(x^*))) - D_\alpha(\hat{P}_{\alpha\,1}(y)|\tilde{f}(y|z_1(x^*)))] > \\ &P(M_2^{'}|\underline{y}) \cdot E_{z_2(x^*)|\underline{y}}[D_\alpha(\hat{P}_{\alpha\,1}(y)|\tilde{f}(y|z_2(x^*))) - D_\alpha(\hat{P}_{\alpha\,2}(y)|\tilde{f}(y|z_2(x^*)))] \end{aligned} \tag{12}$$

where $P(M_j^{'}|\underline{y})$ and $E_{z_j(x^*)|\underline{y}}$ denote the posterior probability of $M_j^{'}$ and the expectation with respect to the posterior distribution of $z_j(x^*)|\underline{y}$ (j=1,2), defined in (11) and (10) respectively.

The following example shows an application of the generalized predictive criterion defined in (12).

8. Example

Assume that the observations are generated by the regression model M^*,

$$M^* : y = m^*(x) + \epsilon,\ \epsilon \sim N(0, 1),$$

where:

$$m^*(x) = 0.8 + x + 0.35 \cdot \sin(0.5 \cdot (x - 1)).$$

Using the notation introduced in the previous paragraph, let:

$$M_1 : y = m_1(x) + \epsilon,\ \epsilon \sim N(0,1),\ m_1(x) = 1 + x,$$

and

$$M_2 : y = m_2(x) + \epsilon,\ \epsilon \sim N(0,1),\ m_2(x) = 1.3 \cdot x + 0.1 \cdot x^2,$$

the parametric models that we want to compare and let

$$M_1^{'} : y = z_1(x) + \epsilon,\ \ z_1(x) \sim GP(m_1(x), W_1(\cdot,\cdot)),\ \ \epsilon \sim N(0,1),$$

and

$$M_2^{'} : y = z_2(x) + \epsilon,\ \ z_2(x) \sim GP(m_2(x), W_2(\cdot,\cdot)),\ \ \epsilon \sim N(0,1),$$

the corresponding enlarged models with:

$$W_1(x^{'}, x'') = W_2(x^{'}, x'') = 0.2^2 \cdot \exp\{-0.2 \cdot (x^{'} - x'')^2.$$

We want to compare models in $\mathcal{M} = \bigcup_{j=1}^{2} M_j$, assuming however that the true model belongs to $\mathcal{M}^{'} = \bigcup_{j=1}^{2} M_j^{'}$. Figures 4 and 5 show some trajectories simulated from the Gaussian processes centered around $m_1(x)$ and $m_2(x)$ respectively while figure 6 shows the behaviour of the true regression function $m^*(x)$ compared with the regression functions of M_1 and M_2. Note that while is very restrictive assuming that the regression function is exactly linear or quadratic it is much more reasonable to assume that the regression function is in a "neighborhood" of a linear or quadratic function.

Let us assume that the goal is to choose a model in $\mathcal{M} = \bigcup_{j=1}^{2} M_j$ in order to estimate the density of y at $x = -2$. and that the Hellinger distance ($\alpha = 0$) is a reasonable measure of divergence for the problem. With respect to this measure of divergence the best model is model M_1, since we have:

$$D_0(\hat{P}_{0,1}(y)|\hat{P}^*(y)) = 0.15 < 0.21 = D_0(\hat{P}_{0,2}(y), \hat{P}^*(y))$$

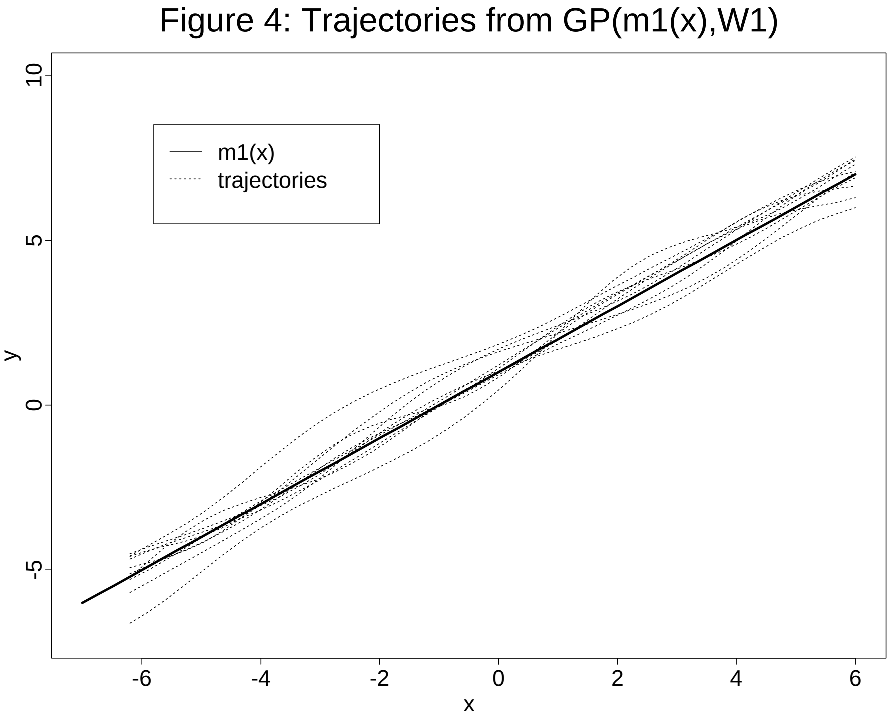
where:

- D_0 represent the Hellinger distance,
- $\hat{P}^*(y)$ is the true distribution of y at $x = -2$,

$$\hat{P}^*(y) = N(m^*(-2), 1) = N(-1.55, 1)$$

- $\hat{P}_{0,j}(y)$ is the generalized predictive distribution of y at $x = -2$ for model M_j, which in this case is simply given by:

$$\hat{P}_{0,j}(y) = N(m_j(-2), 1) = \begin{cases} N(-1,1) & \text{if } j = 1 \\ \\ N(-2.2,1) & \text{if } j = 2; \end{cases}$$



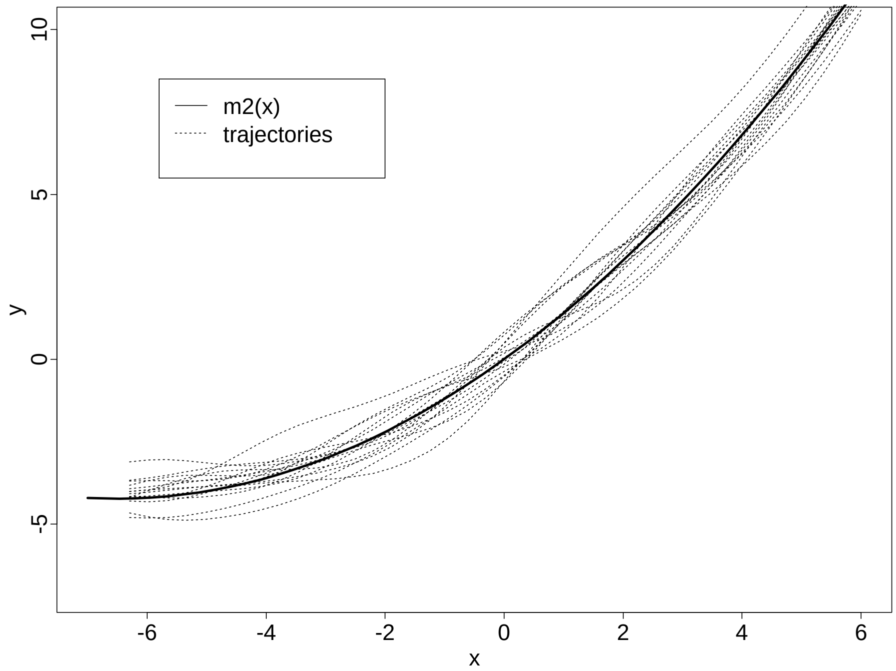
Given a random sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we can compute the expression of the generalized predictive criterion and compare with other standard criteria of model selection. Table 1 shows the values of the Bayes factor ($B_{12.semipar}$) and of the generalized predictive criterion ($GPC_{semipar}$) for $\alpha = 0$ and for three different sample sizes. Equal prior probability for M_1' and M_2' is assumed. For the generalized predictive criterion $GPC_{semipar}$ represents the difference between the expected loss for model M_1 and the expected loss for model M_2 (see (12)). Positive values of $GPC_{semipar}$ indicate that the model M_2 is chosen.

n	$B_{12.semipar.}$	$GPC_{semipar.}(\alpha = 0)$
10	0.70	0.16
20	0.73	-0.12
30	1.45	-0.38

Table 1

As shown in Table 1 already for n=20 the generalized predictive criterion chooses correctly model M_1 which is the closest to the true model with respect to the Hellinger distance. In Table 2 we compare the values of the Bayes factor and of the generalized

Figure 5: Trajectories from GP(m2(x),W2)



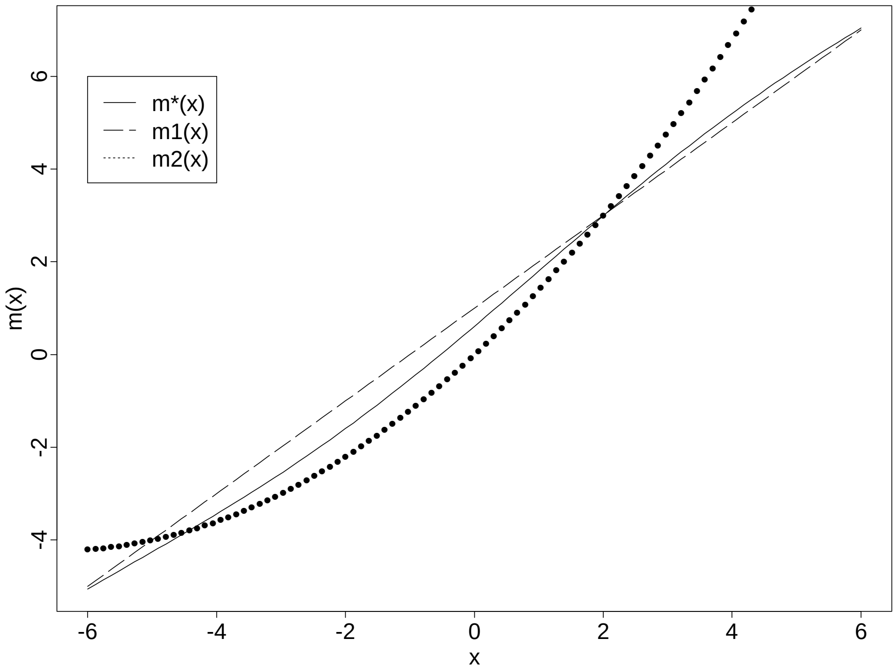
predictive criterion when we assume that the distribution of y at $x = -2$ belongs to $\mathcal{M} = M_1 \bigcup M_2$ with the corresponding values that we got following the semiparametric approach. With $B_{12.par}$ and $GPC_{par}(\alpha = 0)$ we denote respectively the Bayes factor and the generalized predictive criterion in the parametric case when we assume that the true model belongs to $\mathcal{M}$. Note that a value of $GPC_{par}(\alpha = 0)$ greater than one indicate that the model M_1 is selected.

As shown in Table 2 for n=10 there is no difference between the two approaches (semiparametric and parametric), both leading to choose the wrong model, M_2. For $n = 20$, however the generalized predictive criterion based on the semiparametric approach correctly chooses model M_1 while the generalized predictive criterion based on the parametric approach incorrectly chooses model M_2.

n	$B_{12.par}$	$GPC_{par}(\alpha = 0)$	$B_{12.semipar}$	$GPC_{semipar}(\alpha = 0)$
10	0.36	0.36	0.70	0.16
20	0.71	0.71	0.73	-0.12
30	24.13	24.13	1.45	-0.38

Table 2

Figure 6: m*(x) compared with m1(x) and m2(x)



The semiparametric approach that we described here generalizes a previous work of Gutiérrez-Penã. In this approach the author consider nested models and use as loss function the Kullback Leibler divergence. The class $\mathcal{M}'$ is defined assuming that the regression function is a realization of a Gaussian process centered around the larger of the two competiting models. We quite disagree with this choice. When the sample size is not very large centering around the larger model tends to favour the complex model and in general centering only on one model tends to favour that model. Incidentally in our example following the approach suggested by Gutiérrez-Penã for all the three samples that we considered we always end up choosing model M_1 if $\mathcal{M}'$ is defined using a Gaussian process centered around M_1, and model M_2 if $\mathcal{M}'$ is defined using a Gaussian process centered around M_2. One example, of course, is not enough to establish any general conclusion. However define $\mathcal{M}'$ as a mixture of Gaussian processes centered around the competiting models seems to us the most natural way to formalize the prior uncertainty about the regression function. Many generalizations of the predictive criterion presented here are possible. We have assumed, for example, that there are only two competeting models, that the error variance is known and that the goal is to estimate the density of y at $x = x^*$. Our results however can be easily extended to the case in which the models that we wish to compare are k, $k > 2$, the error variance is unkown and the joint distribution of y in m points x is of interest. Another important aspect that needs to be investigated is the asymptotic behaviour of the generalized predictive criterion based on the semiparametric approach. These topics as well the computation and interpretation of the α-divergences will be adressed in future works.

References

Ali, S. and Silvey, S.D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of Royal Statistical Society B, 28, 131-142.*

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association, 91, 109-122.*

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.*New York: Wiley.

Corcuera J.M. and Giummolé, F. (1997). A generalized Bayes rule for prediction. *The Scandinavian Journal of Statistics (to appear).*

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar. 2, 299-318.*

Dmochowshi, J.(1995). Properties of intrinsic Bayes factor . *Ph.D. Thesis.* Purdue University.

Gelfand , A.E. and Dey, D.K. (1994). Bayesian model selection: asymptotics and exact calculations. *Journal of Royal Statistical Society B, 56, 501-514.*

Gutiérrez-Peña, E. (1997). A Bayesian predictive semiparametric approach to variable selection and model comparison in regression. *Proceedings of the 51st Session of the ISI, Invited Papers, Book1. Instambul, Turkey, 17-29.*

Jeffreys, H. (1961). *Theory of Probability*(3rd ed.) London: Oxford University Press.

O' Hagan, A.(1978). Curve fitting and optimal designs for prediction. *Journal of Royal Statistical Society B, 40, 1-42.*

O' Hagan, A.(1992). Some Bayesian numerical analysis. In *Bayesian Statistics 4* (Bernardo, J.M., Berger,J.O., Dawid, A.P. and Smith, A.F.M., eds.). Oxford University Press, 345-363.

O' Hagan, A.(1995). Fractional Bayes factor for model comparison (with discussion). *Journal of Royal Statistical Society B, 57, 99-138.*

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of Royal Statistical Society B, 46, 296-303*

Vajda, I. (1987). *Theory of statistical inference and information.* Kluwer Academic Publisher, London.