Maximum cross-entropy and prior distributions.

Han-Ping Li

MAXIMUM CROSS-ENTROPY AND PRIOR DISTRIBUTIONS

Han-Ping LI
Université Louis Pasteur, Strasvbourg, France
and Wuhan University, China

3 mars 1998

Résumé

A maximum cross-entropy method for determining the reference prior distribution, which represents partial information or complete ignorance, is proposed and studied in this paper, several properties and interpretations are given. Some sufficient conditions for a prior to be the reference distribution or $\epsilon-$ reference distribution are given, these sufficient conditions turn out to be also necessary within some convex class of prior distributions. Besides some theoretical results, attention is also paid to the applications to Bernoulli model, Binomial model, Poisson process, Translation-scale model and Elliptic model. In the case of group models, we find left Haar measure as the reference prior distribution.

Key Words: Maximum cross-entropy, Ignorance prior, Reference prior, Jeffreys prior, Left Haar measure.

AMS Codes: 62A15, 62F15, 94A17, 62A05

1 INTRODUCTION

In studying a statistical problem from Bayesian point of view, one meets at once a serious problem: How to choose the prior distribution? If we possess a huge amount of information, which is hardly ever true, we could probably do it according to our individual degree of belief. But very often, we are asked to give an explicit form of the prior distribution on the basis of some given, but usually very limited and little reliable information at hand. The seeming arbitrariness and subjectivity of the prior distribution is a controversial aspect of Bayesian inference as a scientific methodology of statistical inference, despite its success. Thus, there is a real need to establish certain rules for selecting a specific prior distribution representing partial information or complete ignorance. A lot of procedures has been put forward to deal with this problem.

In nineteenth-century Bayes' applications, it was common to take the uniform prior distribution as a suitable representation of ignorance. However, this naive procedure leads to inconsistencies if applied to different parametrizations of the same problem. Later approaches to this problem are often based on different type of invariance requirements, as those of Jeffreys [13, 1946], [14, 1967], Hartigan [9, 1964], [10, 1971], Villegas [17, 1971], [18, 1977], [20, 1981] and Jaynes [12, 1978]. These approaches can hardly be applied to models lack of group structure and don't always formulate a unique, cogent prior distribution. Other approaches include the use of different forms of information arguments, such as those of Lindley [15, 1956], Jaynes [11, 1968], Good [8, 1969], Zellner [23, 1977], Bernardo [2, 1979], Chang & Eaves [4, 1990] and Eaves & Chang [6, 1992]. If an entropy functional $I(\Pi)$ is considered as an appropriate evaluation of a prior distributions Π in a given preamble class, then the Maximum entropy method is to say that greater $I(\Pi)$ is , better Π is. But the very often proposed and studied Shannon entropy, which is given by

$$I(\Pi) = \int log\left(\frac{d\Pi}{d\lambda}(\theta)\right)\frac{d\Pi}{d\lambda}(\theta)\lambda(d\theta),$$

is not always agreed to be an appropriate evaluation of the prior distributions, especially when Π has a continuous density function. Although it is the only "reasonable" appropriate measure of the information provided by the prior distributions in the discrete case, that is, $Card(\Theta) < \infty$, it loses this property otherwise. Furthermore, it depends on the dominating measure, and is not invariant for one to one transformations. Note that the comparison of the knowledge before and after the experiments makes it possible to discuss the amount of information provided by the experiment. Very interesting papers are Zellner's and Bernardo's along this direction. As will be pointed in the next section, Zellner's approach is not really an information-type one. Bernardo's approach is a very important step in the direction, there are several inadequate reasonings in the paper, although he declared that "

much attention to mathematical detail would be premature". For example, the expression $\int p(\theta)log(f(\theta)/p(\theta))d\theta$ (in his notations) is maximized by $f(\theta) \propto p(\theta)$ only if $f(\theta)$ doesn't depend on $p(\theta)$. So the formula (5) of Bernardo [2, 1979] needs an correct justification. Another problem is that $logp^*(\theta|\hat{\theta})$ (always in his notations) isn't bounded, we haven't in general even the weak convergence of $\int p^*(\theta|\hat{\theta})logp^*(\theta|\hat{\theta})d\theta$ although the asymptotical posterior density will concentrate around the maximum likelihood estimate $\hat{\theta}$ as the sample size trends to infinity. The simple Gaussian model can serve as a counter-example. So the formula (8) (13) etc. of Bernardo [2, 1979] need all further justifications.

In this paper, a maximum cross-entropy method for determining the reference prior distribution, which represents partial information or complete ignorance, is proposed and studied. Bernardo's approach, when all arguments are justified, can be view as an asymptotical version of ours. Several properties and interpretations are given in Section 2. Main results are collected together in section 3. Several sufficient conditions for a prior to be the reference distribution or $\epsilon-$ reference distributions are given. These conditions turn out to be also necessary within some convex class of prior distributions. Besides the theoretical results, attention is also paid to the applications to Bernoulli model, Binomial model, Poisson process, Translation-scale model and Elliptic model. In the case of Group models, we restablish the reference prior position for the left Haar measure.

2 CRITERION and INTERPRETATIONS

Let $(\Omega, \mathscr{B}(\Omega), \mathscr{P})$ be a statistical model, that is, $\mathscr{P}$ is a family of possible probabilities of a certain observand X over a space $(\Omega, \mathscr{B}(\Omega))$ where $\mathscr{P}$ can be either non-parametric family, or a parametric family: $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$. We shall be content to restraint in the parametric case, but the same method can also be used for choosing one reference prior distribution on non-parameter family, c.f. Ferguson [7, 1974]. An prior distribution is a distribution defined over $\mathscr{P}$ endowed with some $\sigma-$ algebra. In the parametric case, this is equivalent to an distribution defined over the parametric space Θ endowded with some $\sigma-$ algebra.

Let $(\Omega, \mathscr{B}(\Omega), \{P_\theta : \theta \in \Theta\})$ be a parametric statistical model, we look for how to determine a prior distribution Π^* in a given preamble class of candidates which represents complete ignorance or insufficient information about parameter θ on $(\Theta, \mathscr{B}(\Theta))$. In [23, 1977], Zellner introduced $G'(\Pi) = I(\Pi) - \int I(P(.|\theta)\Pi(d\theta)$ where $I(\Pi)$ and $I(P(.|\theta))$ are Shannon entropy for prior distribution Π and the

Shannon entropy for the distribution $P(.|\theta)$ respectively. Zellner considered the G' as a measure of the gain in information after having observed X, and for him, maximal date prior Π is that maximizes the gain in information. But, this expression is not a information-type one. Suppose that P_θ, Π are dominated by some σ -finite measure μ, λ respectively, $P_\Pi = \int P_\theta \Pi(d\theta)$, then

$$G'(\Pi) = \int\int log\left(\frac{\frac{dP_\theta}{d\mu}(x)\frac{d\Pi}{d\lambda}(\theta)}{\frac{d\Pi}{d\lambda}(\theta)\frac{d\Pi}{d\lambda}(\theta)}\right) P_\theta(dx)\Pi(d\theta).$$

Look at it more carefully, the numerator of the fraction in log expression is the joint density w.r.t. $\mu \times \lambda$ whereas the denominator is the square of the marginal density w.r.t. λ. This expression has no information-type sense at all. Inspired by him, we consider the following modified functional:

$$G(\Pi) = I(\Pi) - E_X I(\Pi(.|X)) \tag{1}$$

where $I(\Pi)$ and $E_X I(\Pi(.|X))$ are Shannon entropy for prior distribution Π and the expected conditional Shannon entropy for the posteriori distribution $\Pi(.|x)$ respectively. This functional was first used by Lindley to measure the information provided by models w.r.t. a given prior distribution.
Our goal is to search for the prior distribution Π^*, which serves as a reference, within a certain class of prior distributions such that the functional $G(\Pi)$ is maximal at the point $\Pi^* \in \mathscr{C}$.

Example 1(Bernoulli model)
Let X be a Bernoulli variable and P_θ give its distribution, that is,

$$P_\theta(X = 1) = \theta = 1 - P_\theta(X = 0),$$

and the class of prior distributions is given by

$$\mathscr{C} = \{Be(a,b)|Be(a,b)(d\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta, a \geq 0, b \geq 0\}$$

i.e. the class of all Beta priors (conjugate priors) Be(a,b) with $a \geq 0, b \geq 0$. It is obvious that $G(a,b) = G_{a,b} = I_1 + I_2$ where $I_1 = \frac{a}{a+b}\Psi(a+1) + \frac{b}{a+b}\Psi(b+1) - \Psi(a+b+1)$ with $\Psi(x) = \frac{d}{dx}log\Gamma(x)$ and $I_2 = \frac{b}{a+b}log\left(\frac{a+b}{b}\right) + \frac{a}{a+b}log\left(\frac{a+b}{a}\right)$.

It is straightforward that $I_2 \leq log2$, $I_2 = log2$ iff $a = b$; $I_1 \leq 0$, $I_1 = 0$ iff $a = b = 0$. Consequently the reference prior distribution in the class of all Beta priors Be(a,b) is given by Be(0,0): it is $\Pi(d\theta) = \theta^{-1}(1-\theta)^{-1}d\theta$ that maximizes G.

It is the locally invariant prior proposed by Hartigan [9, 1964]. It is also the ignorance prior proposed by Villegas [19, 1977], [20, 1981].

Before looking more examples, let us first establish some interpretations of this criterion.

Interpretation 1 *G is the difference between the Shannon entropy about θ before the observation X and the Shannon entropy about θ after having observed X, hence the diminution of the entropy due to the observation. If the Shannon entropy I is chosen as the uncertainty function of De Groot [5], which measures indetermination degree; then, $G(\Pi) = I(\Pi) - E_X(I(.|X)))$ is the expected information about θ provided by the model $(\Omega, \mathscr{B}(\Omega), \{P_\theta : \theta \in \Theta\})$, Lindley [15, 1956]. The maximization of the criterion means to search for a prior distribution such that the expected information about θ provided by the model is maximal.*

Interpretation 2 *Suppose that P_θ, Π are dominated by some σ -finite measure μ, λ respectively. Denote $P_\Pi = \int P_\theta \Pi(d\theta)$, and $\pi(\theta) = \frac{d\Pi}{d\lambda}(\theta)$, then*

$$G(\Pi) = \int\int log\left(\frac{\frac{dP_\theta}{d\mu}(x)\pi(\theta)}{\frac{dP_\Pi}{d\mu}(x)\pi(\theta)}\right) P_\theta(dx)\Pi(d\theta) \tag{2}$$

is precisely the cross-entropy. The maximization of this criterion means to search for the prior distribution Π such that the Kullback-Leibler divergence between the joint density $\frac{dP_\theta}{d\mu}(x)\pi(\theta)$ and the product marginal densities $\frac{dP_\Pi}{d\mu}(x)\pi(\theta)$ is maximal.

Interpretation 3 *Given $\{P_\theta : \theta \in \Theta\}$, with the prior distribution Π. If one knows only $\eta = T(\theta)$ and $\{\tilde{P}_\eta : \eta \in \Delta\}$ with $\tilde{P}_\eta(.) = E_\Pi P(.|T(\theta) = \eta)$ and $\Delta = T(\Theta)$, then*

$$G(\Pi) \geq G(\tilde{\Pi}(\eta)) \tag{3}$$

The maximization of this criterion means to search for most indeterminate prior distribution.

Interpretation 4 *From the dual point of view , we have always this inequality if the posteriori distribution* $\Pi(.|x)$ *is more informative than* $\tilde{\Pi}(.|x)$ *in the sense of Blackwell for any marginal law p(x).*

Here are some elementary properties:

First of all, we have here, in contrast to Shannon entropy, certain invariance properties, among others, G is parameter invariant, data invariant and context invariant. Having the second interpretation in mind, much more can be said.

According to the basic divergence inequality, $G(\Pi)$ is non-negative, G is nul iff $\Pi(d\theta) = \delta_{\theta_0}(\theta)$ for some θ_0 provided that P_θ is identifiable.
In fact, assume for instance that P_θ is dominated by some σ -finite measure $\mu, p(x|\theta) = \frac{dP_\theta}{d\mu}(x)$, then $\mu\{x \in \Omega | p(x|\theta) \neq p(x|\theta')\} > 0$, for $\theta \neq \theta'$. Hence $G(\Pi) = 0$ implies $\Pi(\Theta_0) = 1$ with $\Theta_0 = \{\theta \in \Theta | \int log(\frac{p(x|\theta)}{p_\Pi(x)} p(x|\theta)\mu(dx)) = 0\}$. If $\theta, \theta' \in \Theta_0$, we have $p(x|\theta) = p(x|\theta')\mu a.e.$, which in turn implies that $\Theta_0 = \{\theta_0\}$, $\Pi(d\theta) = \delta_{\theta_0}(\theta)$.
Thirdly, this functional is concave,(Lindley [15]), consequently, the maxima, if it exists, will be unique.

Moreover, the concavity implies that the map

$$\Pi \longrightarrow G(\Pi)$$

is continuous on $int(\mathscr{C})$. It may fail to be upper semi-continuous on $\mathscr{C}$. For assuming the existence of maxima within the compact class $\mathscr{C}$, one can relax G by $\tilde{G}$ defined as
$\tilde{G}(\Pi) = \lim_H inf\ G(\Pi)$ where infimum is taken over all upper semi-continuous $H \geq G$.

3 MAIN RESULTS

Several sufficient conditions on the functional G to be maximal are given. These conditions are also necessary within some convex class of prior distributions. More precisely, suppose $P_\theta << \mu$, define

$$p(x|\theta) = \frac{dP_\theta}{d\mu}(x),\ p_\Pi(x) = \int p(x|\theta)\Pi(d\theta) \tag{4}$$

$$H_{\Pi}(\theta) = \int log\left(\frac{p(x|\theta)}{p_{\Pi}(x)}\right) p(x|\theta)\mu(dx) \tag{5}$$

we have

Theorem 1 *If there is a prior distribution* Π^* *such that*

$$\int H_{\Pi^*}(\theta)\Pi^*(d\theta) \geq \sup_{\Pi\in\mathscr{C}} \int H_{\Pi^*}(\theta)\Pi(d\theta) - \epsilon \tag{6}$$

then

$$G(\Pi^*) \geq \sup_{\Pi\in\mathscr{C}} G(\Pi) - \epsilon. \tag{7}$$

Any prior distribution satisfies (3.7) is coined an $\epsilon-$reference prior distribution, it is called the reference prior distribution if $\epsilon = 0$.

If we define

$$\Theta^{\epsilon}_{\Pi} = \{\theta \in \Theta| \quad H_{\Pi}(\theta) \geq \sup_{\theta'} H(\theta') - \epsilon\} \quad \text{for} \quad \epsilon \geq 0 \tag{8}$$

we have a more applicable theorem as a consequence of the theorem 1:

Theorem 2 *If there is a prior distribution* Π^* *such that*

$$\Pi^*(\Theta^{\epsilon}_{\Pi^*}) = 1 \tag{9}$$

then Π^* *is a* $\epsilon-$ *reference prior distribution.*

Here is a simple application:

Example 1 :Bernoulli model (continued) .

We know that $P_\theta(X = 1) = \theta = 1 - P_\theta(X = 0)$.

If

$$\Pi(\theta \in B) = \Pi(1 - \theta \in B) \quad \forall \quad B \in \mathscr{B}(\Theta),$$

then

$$\begin{aligned} P_\Pi(X=0) &= \int P(X=0|\theta)\Pi(d\theta) \\ &= \int P(X=1|\theta)\Pi(d\theta) = P_\Pi(X=1) = \frac{1}{2} \end{aligned}$$

$$H(\theta) = (1-\theta)log(1-\theta) + \theta log\theta + log2 \leq log2$$

Take $\Pi[0,\delta] = \Pi[1-\delta,1] = \frac{1}{2}$, we have

$$\Pi(\Theta_\Pi^\epsilon) = 1,$$

with $\epsilon = -(\delta log\delta + (1-\delta)log(1-\delta))$, which give $\epsilon-$ reference prior distributions. Since $\lim_{\delta\to 0}\epsilon = 0$, we find that the reference prior distribution Π^* is given by

$$\Pi^*(\theta=0) = \Pi^*(\theta=1) = \frac{1}{2} \tag{10}$$

Theorem 3 *If there is a prior distribution* Π^* *such that*

$$\int H_{\Pi^*}(\theta)\Pi^*(d\theta) = \sup_{\Pi\in\mathscr{C}}\int H_{\Pi^*}(\theta)\Pi(d\theta) \tag{11}$$

then then Π^* *is the reference prior distribution. This condition is also necessary if the class of prior distributions is convex.*

Remember that

$$\Theta_\Pi^0 = \{\theta\in\Theta| \quad H_\Pi(\theta) = \sup_{\theta'} H_\Pi(\theta')\} \tag{12}$$

If $Pi(\Theta_\Pi^0) > 0$, we can define a hat operation on $\mathscr{C}$ that concentrates each Π on Θ_Π^0:

$$\hat{\Pi}(A) = \frac{\Pi(A\cap\Theta_\Pi^0)}{\Pi(\Theta_\Pi^0)} \tag{13}$$

Note that passing Π to $\hat{\Pi}$ is not necessrily favorable, that is, we don't always have $G(\Pi) \leq G(\hat{\Pi})$. For example, if $\Theta_\Pi^0 = \{\theta_0\}, \hat{\Pi} = \delta_{\theta_0}$, then $G(\hat{\Pi}) = 0 \leq G(\Pi)$. However,we have

Theorem 4 *If there is a prior distribution* Π^* *such that*

$$\Pi^*(\Theta_{\Pi^*}^0) = 1 \tag{14}$$

then Π^* is the reference prior distribution.

This condition is also necessary if the class of prior distributions is convex and closed under the hat operation .

If Π^* verifies the above condition (3.14), then

$$\begin{aligned}\frac{d\Pi^*}{d\lambda} &\propto \frac{d\Pi^*}{d\lambda} exp\left(\int log\left(\frac{p(x|\theta)}{p_{\Pi^*}(x)}\right) p(x|\theta)\mu(dx)\right) \quad \Pi^* \ a.e. \\ &= exp\left(\int log\left(\frac{d\Pi^*(\theta|x)}{d\lambda}\right) p(x|\theta)\mu(dx)\right)\end{aligned}$$

which is precisely the formula (6) of Bernardo [2, 1979].

Example 2:(Binomial model)

Let X be a binomial variable with the parameter θ, that is, $P_\theta(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}$.

If we define

$$\Pi^*(\theta = 0) = \Pi^*(\theta = 1) = p; \quad \text{and} \quad \Pi^*(\theta = \frac{1}{2}) = q \tag{15}$$

with

$$p = \frac{2^{\frac{n}{2^{n-1}-1}} - 2^{-n}}{2\left(2^{\frac{n}{2^{n-1}-1}} - 2^{-n}\right) + 1} \tag{16}$$

and

$$q = \frac{1}{2\left(2^{\frac{n}{2^{n-1}-1}} - 2^{-n}\right) + 1} \tag{17}$$

then

$$\begin{aligned}H_{\Pi^*}(\theta) &= n(\theta log\theta + (1-\theta)log\theta) - (1 - \theta^n - (1-\theta)^n)log(\frac{q}{2^n}) \\ &\quad - (1 - \theta^n - (1-\theta)^n)log(\frac{q}{2^n}) - (\theta^n + (1-\theta)^n)log\left(p + \frac{q}{2^n}\right)\end{aligned}$$

$$
\begin{aligned}
H_{\Pi^*}(0) - H_{\Pi^*}(\theta) &= -(1-\theta^n-(1-\theta)^n) log\left(\frac{p+\frac{q}{2^n}}{\frac{q}{2^n}}\right) \\
&\quad - n(\theta log\theta + (1-\theta)log\theta) \\
&= n(\theta log\tfrac{1}{\theta} + (1-\theta)log\tfrac{1}{(1-\theta)}) \\
&\quad - n\left(\tfrac{2^{n-1}}{2^{n-1}-1}\right) log2(1-\theta^n-(1-\theta)^n) \\
&\geq n\left(4log2\times\theta(1-\theta) - \left(\tfrac{2^{n-1}}{2^{n-1}-1}\right) log2(1-\theta^n-(1-\theta)^n\right)
\end{aligned}
$$

since

$$
\begin{aligned}
(\theta log\frac{1}{\theta} + (1-\theta)log\frac{1}{(1-\theta)}) &= \sum_{k=0}^{\infty}\frac{\theta(1-\theta)^k}{k} + \sum_{k=0}^{\infty}\frac{(1-\theta)\theta^k}{k} \\
&= (2+\sum_{k=2}^{\infty}\frac{1}{k}(\theta^{k-1}+(1-\theta)^{k-1})\theta(1-\theta) \\
&\geq \left(2+\sum_{k=2}^{\infty}\frac{1}{k}\left(\frac{1}{2}\right)^{k-2}\right)\theta(1-\theta) \\
&= \left(2+4\sum_{k=2}^{\infty}\frac{1}{k}\left(\frac{1}{2}\right)^{k}\right)\theta(1-\theta) \\
&= 4log2\times\theta(1-\theta).
\end{aligned}
$$

When n=2, we have

$$
\begin{aligned}
\left(\frac{2^{n-1}}{2^{n-1}-1}\right) log2\times(1-\theta^n-(1-\theta)^n) &= \left(\frac{2^{2-1}}{2^{2-1}-1}\right) log2\times(1-\theta^2-(1-\theta)^2) \\
&= 4log2\times\theta(1-\theta).
\end{aligned}
$$

When n=3, we have

$$
\begin{aligned}
\left(\frac{2^{n-1}}{2^{n-1}-1}\right) log2\times(1-\theta^n-(1-\theta)^n) &= 3\left(\frac{2^{3-1}}{2^{3-1}-1}\right) log2\times(1-\theta^3-(1-\theta)^3) \\
&= 4log2\times\theta(1-\theta).
\end{aligned}
$$

Therefore,

$$
H_{\Pi^*}(0) = H_{\Pi^*}(1) = H_{\Pi^*}(1/2) \geq H_{\Pi^*}(\theta) \text{ and } \Pi^*(\theta\in\{0,1/2,1\}=1. \qquad (18)
$$

So (3.14) holds for n=2, 3.

It is easy to check that

$(\theta log\frac{1}{\theta} + (1-\theta)log\frac{1}{(1-\theta})) \geq \left(\frac{2^{4-1}}{2^{4-1}-1}\right) log2(1-\theta^4-(1-\theta)^4)$

So (3.14) holds also for n=4. Therefore, (3.15), (3.16) and (3.17) give the reference prior distribution when n=2, 3 or 4.

Example 3 (Mixed Gaussian model)

If we have three Gaussian distributions:

$$P_\theta(dx) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(1-\theta)^2\}dx,\ \theta \in \{-5, 0, 0.5\}$$

and the priors are given by

$$\Pi(\theta = 0.5) = p, \Pi(\theta = 0) = q \text{ and } \Pi^*(\theta = -5) = 1 - p - q.$$

It can be checked that

$$p = 0.291044, \quad q = 0.216402$$

give the reference prior distribution Π^*.

Example 2 :Binomial model (continued).

We have $P_\theta(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k(1-\theta)^{n-k}$. Assume that

$\mathscr{C} = \{Be(a,b)|Be(a,b)(d\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta, a \geq 0, b \geq 0\}$ i.e. the class of all Beta priors (conjugate priors) Be(a,b) with $a \geq 0, b \geq 0$.

It turns out that the values of a and b of $Be(a,b)$ that maximizes G depend on the sample seize n. The following table gives the approximate values of a and b as a function of n:

n	1	2	3	4	5
a	0	0.0407691	0.0861015	0.117923	0.141884
b	0	0.0407691	0.0861015	0.117923	0.141884
n	6	7	8	9	10
a	0.160907	0.176573	0.189821	0.201249	0.211264
b	0.160907	0.176573	0.189821	0.201249	0.211264
n	20	30	60	100	150
a	0.271985	0.303157	0.348875	0.376715	0.395446
b	0.271985	0.303157	0.348875	0.376715	0.395446
n	300	500	1000	2000	10000
a	0.421688	0.436927	0.453060	0.465222	0.482686
b	0.421688	0.436927	0.453060	0.465222	0.482686

It seems that when n tends to infinity, a and b both tend to 0.5 which correspond to Jeffreys prior. See also the conjecture below.

We see that for small sample size, the problem gets hardly a simple solution. that is probablly why Bernardo [2, 1979] used somehow indirect, technical and asymptotical arguments. In fact, if we allow the sample size tend to infinity, things go more soundly. Suppose $P_\theta << \mu$, define

$$p(x_i|\theta) = \frac{dP_\theta}{d\mu}(x_i),\ p_\Pi(x_1, ..., x_n) = \int \Pi_{i=1}^n p(x_i|\theta)\Pi(d\theta) \tag{19}$$

$$H_\Pi^{(n)}(\theta) = \int log\left(\frac{\Pi_{i=1}^n p(x_i|\theta)}{p_\Pi(x_1, ..., x_n)}\right)\Pi_{i=1}^n p(x_i|\theta)\mu(dx_i), \tag{20}$$

$$\Theta_\Pi^{\epsilon,(n)} = \{\theta \in \Theta|\ H_\Pi^{(n)}(\theta) \geq \sup_{\theta'} H_\Pi^{(n)}(\theta') - \epsilon\} \text{ for } \epsilon \geq 0 \tag{21}$$

we have

Theorem 5 *If there is a prior distribution* Π^* *such that*

$$\lim_{n\to\infty} \Pi^*(\Theta_{\Pi^*}^{0,(n)}) = 1$$

then $\lim_{n\to\infty} G(\Pi^*) = \sup_{\Pi\in\mathscr{C}} \lim_{n\to\infty} G(\Pi)$.

As a consequence of this, we have a very simple solution to the discrete case:

Theorem 6 *If* $\Theta = \{\theta_1, ..., \theta_m\}$ *and the prior distribution* Π^* *is defined by*

$$\Pi^*(\theta = \theta_j) = \frac{1}{m}, \text{ for } 1 \leq j \leq m$$

then $\lim_{n\to\infty} G(\Pi^*) = \sup_{\Pi\in\mathscr{C}} \lim_{n\to\infty} G(\Pi)$.

Sometimes the parameter may be considered as an unknown element of some group of transformations of the sample space $(\Omega, \mathscr{B}(\Omega))$, in this case, we have

Theorem 7 *Suppose the parameter space* Θ *is a measurable compact group and* $P_\theta(B) = P(\theta^{-1}B), B \in \mathscr{B}(\Omega)$, *then the normalized Haar measure* Π^* *is the reference prior distribution, that is to say:* $G(\Pi^*) \geq G(\Pi)$ *for any prior distribution* Π.

$H_\Pi(\theta)$ is always non-negative for any (proper) prior distribution, but same can not be said for improper prior, as can be showed by the following:

Example 4

Let $\Omega = R$, $p(x|\theta) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}$, are the densities with respect to the Lebesgue measure, λ = Lebesgue measure. It is obvious that $p(x) = 1$ and $H_\lambda(\theta) = -\frac{1}{2}[log(2\pi\sigma^2) + 1]$, and so $G(\lambda) = -\infty$.

This is not surprising, because G is not a suitable measure for evaluate the improper prior. In fact, if we use $\frac{1}{c}\times$ Lebesgue measure , instead of Lebesgue measure as the prior, then we can choose c such that $G = \infty$. We shall overcome this difficulty by "normalize" improper prior as the following theorem.

In discussing the Villegas' paper [17, 1971], Hartigan [10, 1971] talked about the difficulty of choosing one from a family of relatively invariant prior measures, each of which is such that the corresponding posterior distributions are invariant under the same transformations which leave the sampling distribution model invariant. From our point of view, only the left Haar measure is to be chosen as reference by the following:

Theorem 8 *Suppose the parameter space* Θ *is a locally compact group,* $P_\theta(B) = P(\theta^{-1}B), B \in \mathscr{B}(\Omega)$, *and there are a series of sets* $\{K_n\}$ *such that* $K_n \uparrow \Theta$ *and* $0 < \Pi^*(K_n) < \infty$, *with a left Haar measure* Π^*. *Define proper prior distribution* Π_n^* *on* $(\Theta, \mathscr{B}(\theta))$ *by*

$$\Pi_n^*(A) = \frac{\Pi^*(A \cap K_n)}{\Pi^*(K_n)}$$

then we have $\lim_{n\to\infty} G(\Pi_n^*) \geq G(\Pi)$ *for any* Π *provided*

$$\int log\left(\frac{dP_\Pi}{dP_{\Pi^*}}(x)\right) P_\Pi(dx) > -\infty \tag{22}$$

where $P_\Pi(B) = \int P(\theta^{-1}B)\Pi(d\theta)$ *and* $P_{\Pi^*}(B) = \int P(\theta^{-1}B)\Pi^*(d\theta)$.

Note that (3.22) can be arranged to be true by replacing Π^* by $\Pi^* \times$ Constant. Here we form

Conjecture 1 *If the proper prior distribution* Π^* *is defined by* $\Pi^*(d\theta) \propto |I(\theta)|^{\frac{1}{2}} d\theta$, *then under some regular conditions, we have* $\lim_{n\to\infty} G(\Pi^*) = \sup_{\Pi\in\mathscr{C}} \lim_{n\to\infty} G(\Pi)$.

In the same spirit of theorem 8, we can form

Conjecture 2 *If the improper prior distribution* Π^* *is defined by* $\Pi^*(d\theta) \propto |I(\theta)|^{\frac{1}{2}} d\theta$, *then under some regular conditions, we have* $\lim_{n\to\infty} G(\Pi_n^*) = \sup_{\Pi\in\mathscr{C}} \lim_{n\to\infty} G(\Pi)$.

4 APPLICATIONS

Example 5 (Poisson process)

Let $X(t)$ be a Poisson process with parameter λ, that is, its distribution is given by

$$P_\lambda(X(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

One can show that the reference prior distribution is given by its Haar measure:

$$\Pi(d\lambda) = \lambda^{-1} d\lambda$$

It is the prior of ignorance proposed by Villegas [20, 1981].

Example 6 (The translation-scale model])
Let X be a variable with its density given by:

$$p(x|(\sigma,\mu)) = \Pi_{i=1}^{n} f(\tfrac{x_i-\mu}{\sigma}), \text{ and } \Theta = \{(\sigma,\mu)|\sigma > 0, \mu \in R\}.$$

Define

$$g_{(\sigma,\mu)}(x_1, ..., x_n) = (\sigma x_1 + \mu, ..., \sigma x_n + \mu),$$

then,

$$g_{(\sigma,\mu)^{-1}}(x_1, ..., x_n) = (\frac{x_1 - \mu}{\sigma}, ..., \frac{x_n - \mu}{\sigma})$$

and the reference prior distribution is given by its Haar measure:

$$\Pi(d\sigma) = \sigma^{-2} d\sigma d\mu.$$

c.f. Barndorff-Nielsen [1, 1988]. It is worth noting that it is precisely the prior called "inner prior" of Villegas , [19, 1977], [20, 1981].

It is also the prior of Jeffreys $\Pi(d\theta) \propto |I(\theta)|^{\frac{1}{2}} d\theta$.

Example 7(The elliptic model)
Let X be a variable of elliptic form with its distribution given by:

$$P(dx|\beta,\sigma) = \tfrac{\Gamma(n/2)}{\Pi^{n/2}} |\Sigma_0|^{-1/2} (\sigma)^{-n} F(\sigma^{-2}(x - A\beta)^t \Sigma_0^{-1}(x - A\beta)) dx$$

We know that its left Haar measure is given by:

$$\Pi(d\beta, d\sigma) = \sigma^{-(n+1)} d\beta d\sigma$$

it is also its reference prior distribution.

One can show that its Fisher information matrix is given by:

$$I(\beta,\sigma) = \begin{pmatrix} a\sigma^{-2} A^t \Sigma_0^{-1} A & 0 \\ 0 & b\sigma^{-2} \end{pmatrix} \tag{23}$$

so $detI(\beta,\sigma) = c \times \sigma^{-2(n+1)}$, see Burbea & Oller [3, 1988] for the detail. Hence the reference prior distribution is also the prior of Jeffreys: $\Pi(d\theta) \propto |I(\theta)|^{\frac{1}{2}} d\theta$.

There seems to be some connexion between the reference prior and minimax estimator under entropy loss, see Wieczorkowski, R. Zielinski, [21, 1992] and R.Yang M. C. [22, 1992].

5 SKETCH of PROOFS

Proof of Theorem 1:

Suppose that Π^* verifies the condition in the theorem, $\Pi^* \in \mathscr{C}$ is any prior distribution, then

$$\begin{aligned}
G(\Pi^*) - G(\Pi) &= \int H_{\Pi^*}(\theta)\Pi^*(d\theta) - \int H_{\Pi^*}(\theta)\Pi(d\theta) \\
&\quad + \int H_{\Pi^*}(\theta)\Pi(d\theta) - \int H_{\Pi}(\theta)\Pi(d\theta) \\
&\geq \int H_{\Pi^*}(\theta)\Pi(d\theta) - \int H_{\Pi}(\theta)\Pi(d\theta) - \epsilon \\
&= \iint log\left(\frac{p_\Pi(x)}{p_{\Pi^*}(x)}\right) p(x|\theta)\mu(dx)\Pi(d\theta) - \epsilon \\
&= \int log\left(\frac{p_\Pi(x)}{p_{\Pi^*}(x)}\right) p_\Pi(x)\mu(dx) - \epsilon \\
&\geq -\epsilon
\end{aligned}$$

the last inequality is due to the positivity of Kullback information.

Proof of Theorem 2:

It is easy to see that

$$\begin{aligned}
\int H_{\Pi^*}(\theta)\Pi^*(d\theta) &\geq \sup_{\theta'} H_{\Pi^*}(\theta') - \epsilon \\
&\geq \int H_{\Pi^*}(\theta)\Pi(d\theta) - \epsilon
\end{aligned}$$

Proof of Theorem 3:

It remains only to prove the necessary part:
Assume that G is maximal at the point $\Pi^* \in \mathscr{C}$, let $\Pi \in \mathscr{C}$ be another point, then $\Pi^\alpha = (1-\alpha)\Pi^* + \alpha\Pi \in \mathscr{C}$ also because of the convexity of $\mathscr{C}$. We have $\lim_{\alpha\to 0} \frac{G(\Pi^\alpha)-G(\Pi^*)}{\alpha} \leq 0$, that is, $\frac{d}{d\alpha} G(\Pi^\alpha)\big|_{\alpha=0} \leq 0$. Therefore,

$$\int H_{\Pi^*}(\theta)\Pi^*(d\theta) - \int H_{\Pi^*}(\theta)\Pi(d\theta) \geq 0.$$

Proof of Theorem 4:

Suppose $\Pi^* \in \mathscr{C}$ maximize G. $\mathscr{C}$ is closed under the hat operation implies that $\hat{\Pi}^* \in \mathscr{C}$. So

$$\int H_{\Pi^*}(\theta)\Pi^*(d\theta) - \sup_{\theta'} H_{\Pi^*}(\theta') = \int H_{\Pi^*}(\theta)\Pi^*(d\theta) - \int H_{\Pi^*}(\theta)\hat{\Pi}^*(d\theta) \geq 0.$$

But this is true iff $\Pi^* = \hat{\Pi}^*$ iff (3.14) holds.

Proof of Theorem 5:

The proof is similar to the sufficient part of the theorem 4.

Proof of Theorem 6:

See, Renyi [16, 1964]. See also Bernardo [2, 1979].

Proof of Theorem 7:

Suppose that $P << \mu$, define a map T_θ on $(\Omega, \mathscr{B}(\Omega))$ by $T_\theta : x \longmapsto \theta x$, then

$$\frac{dP_\theta}{d\mu T_\theta^{-1}}(t) = \frac{dP}{d\mu}(\theta^{-1}t), \quad \mu T_\theta^{-1} \; p.s.$$

It is not difficult to see that

$$\begin{aligned}
I_1 &= \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log\left(\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)\right) \mu T_\theta^{-1}(dt) \\
&= \int \frac{dP}{d\mu}(\theta^{-1}t) log\left(\frac{dP}{d\mu}(\theta^{-1}t)\right) \mu T_\theta^{-1}(dt) \\
&= \int \frac{dP}{d\mu}(\theta^{-1}T_\theta(x)) log\left(\frac{dP}{d\mu}(\theta^{-1}T_\theta(x))\right) \mu(dx) \\
&= \int \frac{dP}{d\mu}(x) log\left(\frac{dP}{d\mu}(x)\right) \mu(dx)
\end{aligned}$$

Denote the Haar measure on $(\Theta, \mathscr{B}(\Theta))$ by Π^*,

$$\begin{aligned} I_2 &= \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log\left(\frac{dP_{\Pi^*}}{d\mu}(t)\right) \mu T_\theta^{-1}(dt) \\ &= \int \frac{dP}{d\mu}(\theta^{-1}t) log\left(\frac{dP_{\Pi^*}}{d\mu}(t)\right) \mu T_\theta^{-1}(dt) \\ &= \int \frac{dP}{d\mu}(\theta^{-1}T_\theta(x)) log\left(\frac{dP_{\Pi^*}}{d\mu}(T_\theta(x))\right) \mu(dx) \\ &= \int \frac{dP}{d\mu}(x) log\left(\frac{dP_{\Pi^*}}{d\mu}(\theta x)\right) \mu(dx) \\ &= \int \frac{dP}{d\mu}(x) log\left(\frac{dP_{\Pi^*}}{d\mu}(x)\right) \mu(dx) \end{aligned}$$

because

$$\begin{aligned} \frac{dP_{\Pi^*}}{d\mu}(\theta x) &= \int \frac{dP}{d\mu}(\xi^{-1}\theta x)\Pi^*(d\xi) \\ &= \int \frac{dP}{d\mu}((\theta^{-1}\xi)^{-1}x)\Pi^*(d\xi) \\ &= \int \frac{dP}{d\mu}(\xi^{-1}x)\Pi^*(d\xi) = \frac{dP_{\Pi^*}}{d\mu}(x); \end{aligned}$$

the before-last equality being true since Π^* is left invariant. As a consequence of that,

$$\begin{aligned} H_{\Pi^*}(\theta) &= I_1 - I_2 = \text{Constant} \qquad (24) \\ &= \sup_{\theta\in\Theta} H_{\Pi^*}(\theta). \qquad (25) \end{aligned}$$

Note that when Π^* is an infinite left Haar measure, the above relation (5.23) is always true except the last equality (5.24), since this constant may fail to be positive. See also example 4.

Proof of Theorem 8:
Keep the same natations as in the theorem 8, denote the corresponding densities

by small letters, $\frac{d\Pi_n^*}{d\lambda}(\theta) = \pi_n^*(\theta)$ etc. , then

$$\begin{aligned}
G(\Pi_n^*) - G(\Pi) &= \left(\int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)\pi_n^*(\theta)}{\frac{dP_{\Pi_n^*}}{d\mu}(t)\pi(\theta)} \right) \mu T_\theta^{-1}(dt) \right. \\
&- \left. \int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \right) \\
&+ \left(\int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \right. \\
&- \left. \int \Pi(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_\Pi}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \right) \\
&+ \left(\int \Pi_n^*(d\theta) log \left(\frac{\pi(\theta)}{\pi_n^*(\theta)} \right) \right) \\
&= \sum_{i=1}^{3} I_i
\end{aligned}$$

We then have

$$\begin{aligned}
I_3 &= \int \Pi_n^*(d\theta) log \left(\frac{\pi(\theta)}{\pi_n^*(\theta)} \right) = \int \Pi_n^*(d\theta) log \Pi^*(K_n) \\
&= log \Pi^*(K_n) \longrightarrow +\infty
\end{aligned}$$

$$
\begin{aligned}
I_2 \quad &= \quad \int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \\
&- \quad \int \Pi(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \\
&= \quad \int \Pi(d\theta) \int \frac{dP}{d\mu}(\theta^{-1}t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \\
&- \quad \int \Pi(d\theta) \int \frac{dP}{d\mu}(\theta^{-1}t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \\
&= \quad \int \Pi(d\theta) \int \frac{dP}{d\mu}(\theta^{-1}t) log \left(\frac{\frac{dP_{\Pi}}{d\mu}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \\
&= \quad \int P_\Pi(dx) log \left(\frac{dP_\Pi}{dP_{\Pi^*}}(x) \right) \\
&> -\infty
\end{aligned}
$$

since

$$
H_{\Pi^*}(\theta) = \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu T_\theta^{-1}}(t)} \right) \mu T_\theta^{-1}(dt) = \text{ Constant}
$$

by (5.23).

$$
\begin{aligned}
I_1 &= \left(\int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)\pi_n^*(\theta)}{\frac{dP_{\Pi_n^*}}{d\mu}(t)\pi(\theta)} \right) \mu T_\theta^{-1}(dt) \right. \\
&- \left. \int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_\theta}{d\mu T_\theta^{-1}}(t)}{\frac{dP_{\Pi^*}}{d\mu}(t)} \right) \mu T_\theta^{-1}(dt) \right) \\
&= \int \Pi_n^*(d\theta) \int \frac{dP_\theta}{d\mu T_\theta^{-1}}(t) log \left(\frac{\frac{dP_{\Pi^*}}{d\mu}(t)\pi_n^*(\theta)}{\frac{dP_{\Pi_n^*}}{d\mu}(t)\pi(\theta)} \right) \mu T_\theta^{-1}(dt) \\
&= \int \Pi_n^*(d\theta) \int log \left(\frac{\int_\Theta \frac{dP}{d\mu}(\theta^{-1}t)\Pi^*(d\theta)}{\int_{K_n} \frac{dP}{d\mu}(\theta^{-1}t)\Pi^*(d\theta)} \right) P_\theta(dt) \\
&> 0.
\end{aligned}
$$

due to

$$
log \left(\frac{\int_\Theta \frac{dP}{d\mu}(\theta^{-1}t)\Pi^*(d\theta)}{\int_{K_n} \frac{dP}{d\mu}(\theta^{-1}t)\Pi^*(d\theta)} \right) > 0.
$$

I tender my thanks to Jean-Pierre Florens, Fabrice Gamboa, Elisabeth Gassiat and Michel Mouchart for the helpful discussions.

Références

[1] Barndorff-Nielsen O.E.(1988) Parametric Statistical Models and Likelihood. *Lecture Notes in Statist.*,**50**, Springer.

[2] Bernardo, J.M. (1979), Reference posterior distributions for Bayesian inference (with Discussion).
J.Roy. Statist. Soc. Ser. B. **41**, 113-147.

[3] Burbea J. & Oller J.M. (1988) The information metric for univariate lear elliptic models.
Statistics & Decisions **6**, 209-221.

[4] Chang T., Eaves D. (1990) Reference prior for the orbit in a group model.
Ann. Statist. **18**, No 4, 1595-1614.

[5] De Groot (1962) Uncertainty, information, and sequential experiments.
Ann. Math. Statist. **33**, 404-419.

[6] Eaves D., Chang T. (1992) Priors for ordered conditional variance and vector partial correlation.
J. Multivariate Anal., **41**, 43-55.

[7] Ferguson,T.S. (1974) Prior distributions on spaces of probability measures.
Ann. Statist. **2**, No 4, 615-629.

[8] Good I.J.(1969) What is the use of a distribution?
In Krishnaiah (edt.),*Multivariate Analysis* ,**II**, 183-203, New York: Academic Press.

[9] Hartigan J.A. (1964) Invariant prior distributions.
Ann. Math. Statist.,**35**, 836-845.

[10] Hartigan J.A.. (1971) Discussion of Villegas'paper.
In Godambe & Sprott (edt.),*Foundation of Statistical inference*, 415-416. Waterloo.

[11] Jaynes, E.T. (1968) Prior probabilities.
IEEE Trans. Systems, Science and Cybernetics.,**SCC-4**,227-291.

[12] Jaynes, E.T. (1978) Marginalization and prior probabilities.
In *Zellner, A. ed; Studies of Bayesian Statistics.*

[13] Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems.
Proc. R. Sco. London **A, 186**,453-461.

[14] Jeffreys, H. (1967) Theory of Probability (3rd edt.).
Oxford: Clarendon Press.

[15] Lindley,D.V. (1956) On a measure of the information provided by an experiment.

Ann. Math. Statist. **27**, 986-1005.

[16] Renyi, A. (1964). On the amount of information concerning an unknown parameter in a sequence of observations.

Publ. Math. Inst. Hung. Acad. Sci. **9A**, 617-625.

[17] Villegas C. (1971) On Haar priors.

In Godambe & Sprott (edt.),*Foundation of Statistical inference*, 409-414. Waterloo.

[18] Villegas C. (1977) Inner statistical inference.

J. Amer. Statist. Assoc., **72**, No 358, 453-458.

[19] Villegas C. (1977) On the representation of ignorance.

J. Amer. Statist. Assoc., **72**, No 359, 651-654

[20] Villegas C. (1981) Inner statistical inference II.

Ann. Statist., **9**, No 4, 768-776.

[21] Wieczorkowski, R. Zielinski, R.(1992) Minimax estimation of binomial probability with entropy loss function.

Statistics & Decisions, **10**, 39-44.

[22] Yang M. C. (1992). Ridge estimation of independent poisson means under entropy loss.

Statistics & Decisions, **10**, 1-23.

[23] Zellner, A. (1977) Maximal data information prior distributions.

In A. Aykac and C. Brumat (edt.) *New Developments in the Applications of Bayesian Methods.* 211-232, North-Holland. Amesterdam.

Han ping LI

IRMA, Université Louis Pasteur
8 rue René-Descartes
67084 Strasbourg
France

and

Department of Mathematics
Wuhan University
430072, Wuhan
P. R. of China