Experimental Design to Maximize Information

P. Sebastiani* and H.P. Wynn†

*Department of Mathematics and Statistics
University of Massachusetts at Amherst, 01003 MA
†Department of Statistics, University of Warwick
Coventry, CV4 7AL

Abstract. This paper will consider different methods to measure the gain of information that an experiment provides on parameters of a statistical model. The approach we follow is Bayesian and relies on the assumption that information about model parameters is represented by their probability distribution so that a measure of information is any summary of the probability distributions satisfying some sensible assumptions. Robustness issues will be considered and investigated in some examples using a new family of information measures which have the log-score and the quadratic score as special cases.

INTRODUCTION

Suppose an experimenter has to choose an experiment ξ from a set Ξ. This choice has to be made in such a way to improve his knowledge about a statistical model with probability distribution $P_Y(\cdot|\theta,\xi)$. Suppose also that this experimenter is Bayesian, so that his knowledge about $P_Y(\cdot|\theta,\xi)$ is represented by a probability distribution $P_\Theta(\cdot)$ on the model parameters Θ. There are several approaches the experimenter can choose from and recent descriptions are in [1] or [2]. In this paper, we will focus attention to *the informational approach*, which can be described as follows. If the experimenter wishes to follow the informational approach, he will choose the experimental design which maximizes the average amount of information on the model parameters. This is the approach formulated by Lindley in [3] and uses the negative of Shannon entropy of the posterior distribution of the model parameters as uniquely acceptable measure of information. The criterion for choice of experiment then becomes to maximize the expected gain of information between prior and posterior distribution and reduces to minimizing the expected Shannon entropy of the posterior distribution. Thus, the informational approach requires one to measure the information provided by an experiment via the negative of Shannon entropy. This is the case if one wishes to follow, without modifications, the axiomatic development of the theory of measuring information as originated by Shannon in [4], see also [5]. A more general measure of information was proposed by Renyi in [6], by relaxing one of the axioms formulated by Shannon but, despite its generality, Renyi's entropy has not received great attention. This strong agreement on how to measure information should be set against the fact that there does not seem to be an agreement on the definition of information, see [7] for a recent discussion. We give a new definition of information about the parameter Θ. From this definition, we can then define a general measure of the gain of information provided by an experiment in terms of the expected

change of uncertainty from the prior to the posterior distribution of Θ. By using a new class of uncertainty measures introduced by Sebastiani and Wynn in [8], we derive a general class of criteria to design experiments, which have the informational approach as a special case.

INFORMATION AND ITS MEASURES

Let Θ be a random variable with support T and suppose we assign Θ the probability distribution

$$P(t) = p(\Theta \leq t) = \int_{-\infty}^{t} P(d\theta) = \int_{-\infty}^{l} p(\theta) d\theta$$

where $p(\theta)$ is the probability density function. Although uncertain, this is everything we know about Θ. Therefore, we call this the *representation of the information we have about* Θ. Within this definition, a measure of the information about Θ is a summary of its probability distribution. As minimal requirement, a good information measure should increase as the distribution becomes more concentrate. In general terms, we can also define the negative of an information measure an *entropy measure*. We will provide a general method to measure the information about Θ. We consider first an example.

Example 0.1 Suppose Θ is a binary variable, taking on value 1 with probability p and 0 with probability $q = 1 - p$. Thus, the probability distribution P of Θ is a representation of the information we have about the variable. Suppose now we define a scoring function, say $S(\theta, P)$, which assigns value $\log(p)$ to the probability p that the events $\Theta = 1$ occurs, and assigns value $\log(1-p)$ to the probability $1-p$ that the event $\Theta = 0$ occurs. Thus, the larger the value p, the larger the score assigned to it if the event $\Theta = 1$ occurs, so this function scores the quality of the information we have in terms of the predictive ability it provides. We can now define a measure of the information represented by the probability distribution P as, for example, a weighted average of the scores $\log(p)$ and $\log(1-p)$. Thus, for any λ in the interval $[0;1]$, we could define a measure of information as

$$I(P, \lambda) = \lambda \log(p) + (1-\lambda) \log(1-p)$$

It is now straightforward to show that $I(P, \lambda)$ is maximum when $\lambda = p$, and since we wish to find the best measure of information represented by the probability distribution P, it seems reasonable to use as measure of information the quantity $I_0(P) = p \log(p) + (1-p) \log(1-p)$. This quantity turns out to be the overall expected score and it is only a function of the probability distribution P. The function $I_0(P)$ is minimum when $p = 0.5$, so that the information we have about Θ does not allow us to discriminate between the two events $\Theta = 1$ and $\Theta = 0$. The maximum expected score is when either $p = 0$ or $p = 1$. In this case, prediction made with this probability distribution has no uncertainty. The idea of using the logarithm to score the predictive capability of a probability distribution is due to Good [9]. □

The negative of $I_0(P)$ is the Shannon entropy $\text{ENT}_0(\Theta)$ and, among others, has the following properties:

1. $\text{ENT}_0(\Theta)$ is a concave function of the probability distribution P.

2. When Θ is a discrete variable, $\text{ENT}_0(\Theta)$ is maximum if P is a uniform distribution.
3. If $\Theta = (\Theta_1, \Theta_2)$, then $\text{ENT}_0(\Theta)$ satisfies the following decomposition:

$$\text{ENT}_0(\Theta_1, \Theta_2) = \text{ENT}_0(\Theta_1) + E(\text{ENT}_0(\Theta_2|\Theta_1)). \tag{1}$$

Concavity is a consequence of $\text{ENT}_0(\Theta)$ being a weighted average score, in which the weights are chosen to minimize the entropy measure. Thus, in general, we have

$$\text{ENT}_0(\Theta) = -\int \log(p(\theta))p(\theta)d\theta \leq -\int \log(p(\theta))q(\theta)d\theta$$

for any distribution Q, with density function $q(\theta)$. Interestingly, under differentiability conditions, any concave function $H(\cdot)$ on the space of probability distributions for Θ can be seen as an expected score, with mixing distribution chosen to minimize the entropy measure. This is shown by Savage, see [10], or [11], in a decision-theoretic context, and the associated score is

$$S(\theta, P) = H(P) + R'(\epsilon)|_{\epsilon=0} \tag{2}$$

where $R(\epsilon)$ is the function defined as $H\{(1-\epsilon)P + \epsilon\delta_\theta\}$, and δ_θ is the point mass at θ. For example, in Shannon case, one can easily show that $R'(\epsilon) = -\log(p(\theta)) - \text{ENT}_0(\Theta)$ from which one can recover the score $S(\theta, P) = -\log(p(\theta))$. The second and third property resemble two of the three axioms required by Shannon in [4] to characterize a proper information measures. Property 2, above, is a desirable property of an entropy measure, as intuitively one expects an entropy measure to increase as the distribution becomes non-informative. In general, we can replace property 2 by the requirement that an entropy measure be monotone under contractive maps, so that the entropy decreases as the distribution becomes more concentrate. The third property requires that the overall entropy of a random vector Θ can be decomposed into a weighted sum of individual components. The decomposition (1) has proved to be particularly important in experimental design. Several authors, see [12, 13, 14, 15], have used this decomposition to develop maximum-entropy sampling criteria for experimental design. This decomposition, however, can be generalized to any function $H(\cdot)$, so that if $\Theta = (\Theta_1, \Theta_2)$, then, for some non-negative function g, we ask that

$$H(\Theta_1, \Theta_2) = H(\Theta_1) + E\{g(\Theta_1)[H(\Theta_2|\Theta_1)]\} = H(\Theta_2) + E\{g(\Theta_2)[H(\Theta_1|\Theta_2)]\} \tag{3}$$

One interesting consequence of decomposition (3) is that, when Θ_1 and Θ_2 are independent, then

$$H(\Theta_1, \Theta_2) = H(\Theta_1) + H(\Theta_2)E[g(\Theta_1)] = H(\Theta_2) + H(\Theta_1)E[g(\Theta_2)] \tag{4}$$

so that, when $g(\cdot) = 1$, as in Shannon case, the entropy measure of independent variables is additive:

$$\text{ENT}_0(\Theta_1, \Theta_2) = \text{ENT}_0(\Theta_1) + \text{ENT}_0(\Theta_2). \tag{5}$$

These generalizations of properties 1–3 of Shannon entropy are the rationale for the next definition.

Definition 1 (Entropy measure) *Given a variable Θ, we define a function $H(\cdot)$ which is:*

1. Concave *in the set of probability distribution for Θ;*
2. Monotone *under contractive maps;*
3. Decomposable *as in (3);*

an entropy measure.

In the next sections we will use this approach to define new measures of the gain of information provided by an experiment.

GENERALIZED INFORMATIONAL APPROACH

Suppose that y are the data produced by an experiment ξ and are generated from a parametric distribution, with density function $p(y|\theta,\xi)$, conditional on the parameter value θ and the experiment ξ. We suppose that Θ has prior distribution with positive density $q(\theta)$, independent of ξ. Data y are used to update the prior into posterior distribution, whose density function will be $p(\theta|y,\xi)$. The traditional approach, introduced by Lindley in [3] to measure the gain of information provided by an experiment ξ, uses the negative of Shannon entropy $-\text{ENT}_0(\Theta)$ as measure of information on the parameter Θ. Thus, the gain of information provided by the experiment ξ generating data y is $G_0(\xi,y) = \text{ENT}_0(\Theta) - \text{ENT}_0(\Theta|\xi,y)$. The expected gain of information provided by the experiment ξ is then

$$G_0(\xi) = E(G_0(\xi,Y)) = E[\text{ENT}_0(\Theta) - \text{ENT}_0(\Theta|\xi,y)]$$

where the expectation is with respect to the marginal distribution of Y. It is well known, see [3], that $G_0(\xi)$ is non-negative and $G_0(\xi) = 0$ when Θ and Y are independent. We seek a generalization of this which can use any entropy measure $H(\cdot)$ which obeys the property of concavity, monotony and decomposability. One possible interpretation of $G_0(\xi)$ is in terms of expected reduction of entropy of the distribution of Θ provided by the experiment. This interpretation suggests the generalization of the gain of information that we give next.

Definition 2 (Generalized Gain of Information) *Let $H(\cdot)$ be an entropy measure, as in Definition 1. Then, for some non-negative function g such that (3) holds, the quantity*

$$G(\xi) = E\{g(Y|\xi)[H(\Theta) - H(\Theta|Y,\xi)]\}$$

(the expectation being over the marginal distribution of Y) is a measure of the expected gain of information provided by the experiment ξ.

The function $G(\xi)$ is zero, whenever Θ and Y are independent, so that knowing Y does not change the distribution of Θ. Furthermore, by using Equation (3), we have that

$$G(\xi) = E[g(Y|\xi)]H(\Theta) + H(Y|\xi) - H(\Theta,Y|\xi)$$

and since $E[g(Y|\xi)]H(\Theta)+H(Y|\xi)$ equals $H(\Theta,Y|\xi)$ when Θ and Y are independent, then the expected gain of information is the difference between the overall entropy in the independent case, and the overall entropy in the dependent case. This result is a further support that $G(\xi)$ is an appropriate generalization of Shannon case. In fact, the expected gain of information $G_0(\xi)=E(\text{ENT}_0(\Theta)-\text{ENT}_0(\Theta|\xi,Y))$, can also be written as

$$\begin{aligned} G_0(\xi) &= \int p(\theta,y|\xi)\log(\frac{p(\theta,y|\xi)}{q(\theta)p(y|\xi)})d\theta dy \\ &= \text{ENT}_0(\Theta)+\text{ENT}_0(Y|\xi)-\text{ENT}_0(\Theta,Y) \end{aligned}$$

Thus, $G_0(\xi)$ is the difference between the overall Shannon entropy in the independent case, and the overall Shannon entropy in the dependent case. This difference is also known as cross-entropy of Θ and Y, or mutual information, see [16]. Compared to Shannon case, we note that the generalized expected gain of information is not necessarily positive, and in the next sections we will provide examples. It would seem a minimal requirement of an optimal experiment that

$$G(\xi)>0$$

and maximization of $G(\xi)$ (or a normalized version) is a suitable design criterion.

THE α-ENTROPY

Shannon entropy is a special case of a more general family of uncertainty measures, which was recently introduced by Sebastiani and Wynn [8]. They are a somewhat adapted version of the usual Renyi entropy, which is defined as $\text{ENT}(P)=(1-\alpha)^{-1}\log\int p(\theta)^{\alpha}d\theta$, see [6], and more recently [17].

Definition 3 (α-entropy) *Let P be a probability distribution for Θ in $\mathcal{P}$. We suppose that $\mathcal{P}$ contains only probability distributions with density function $p(\theta)$ in the space L_p, for $p\geq 0$, so that the integral $\int p(\theta)^{\alpha+1}d\theta$ exists and it is finite for any $\alpha>-1$. The α-entropy of P is*

$$\text{ENT}_{\alpha}(\Theta)=\int\frac{1-p(\theta)^{\alpha}}{\alpha}p(\theta)d\theta=\frac{1}{\alpha}\left[1-\int p(\theta)^{\alpha+1}d\theta\right]\quad \alpha>-1 \tag{6}$$

Sebastiani and Wynn [8] derive several properties of the α-entropy. Here, we are interested in those properties which characterize an entropy measure, according to Definition 1.

Theorem 1 *Let P be a probability distribution for Θ with positive density function $p(\theta)$, and suppose P belongs to the concave set of distributions $\mathcal{P}$. We also suppose that $\int p(\theta)^{\alpha+1}d\theta$ exists, for any $\alpha>-1$. The function $\text{ENT}_{\alpha}(P)$ is concave in $\mathcal{P}$ for $\alpha>-1$.*

Thus, the α-entropy obeys the first requirement for an entropy measure in Definition 1. In particular, it can be shown that a scoring function producing the α-entropy is

$$S_{\alpha}(\theta,a(Q))=\frac{1}{\alpha}-\frac{1+\alpha}{\alpha}q(\theta)^{\alpha}+\int q(\theta)^{\alpha+1}d\theta.$$

The following are cases of interest:

- $\alpha \downarrow 0$: then $S_\alpha(\theta, a(Q)) \to S_0(\theta, Q) = -\log q(\theta)$: the log-score and $\text{ENT}_0(\Theta) = -\int \log p(\theta) p(\theta) d\theta$ is Shannon entropy.
- $\alpha = 1$: the scoring rule $S_1(\theta, a(Q))$ is $1 - 2q(\theta) + \int q(\theta)^2 d\theta$. This is known as quadratic score, and was first introduced by DeFinetti in [18]. The associated entropy measure is $H_1(P) = 1 - \int p(\theta)^2 d\theta$.
- $\alpha \downarrow -1$: then $\text{ENT}_{-1}(P) = -1 + \int d\theta$, which is unbounded except for the finite support case, when $\text{ENT}_\alpha(P)$ measures the support size (minus 1).
- $\alpha \uparrow \infty$: then $\text{ENT}_\alpha(\infty)$ is the modal value of the distribution.

Shannon entropy is invariant under Euclidean transformation, and monotone with respect to *contractive maps*, i.e., smooth transformation with Jacobian smaller that 1. This is shown, for example, in [19]. The next Theorem shows similar properties for the α-entropy.

Theorem 2 *Let Θ_1 be a random vector with density function $p(\theta_1)$, and suppose that $\text{ENT}_\alpha(\Theta_1)$ exists. Let $\Theta_2 = g(\Theta_1)$ be a smooth transformation, such that $g^{-1}(\cdot)$ exists, and let J be the Jacobian of the transformation $g^{-1}(\cdot)$. Then, if $\alpha > -1$,*

$$\begin{cases} \text{ENT}_\alpha(\Theta_2) > \text{ENT}_\alpha(\Theta_1) & \textit{if} \quad |J| < 1 \\ \text{ENT}_\alpha(\Theta_2) < \text{ENT}_\alpha(\Theta_1) & \textit{if} \quad |J| > 1 \end{cases}$$

Thus, Theorem 2 guarantees that the second requirement for an entropy measure is satisfied. We now consider the third requirement.

We first define the α-expectation of $h(\Theta)$ as

$$E_\alpha(h(\Theta)) = \int h(\theta) p(\theta)^{\alpha+1} d\theta = E(h(\theta) p(\Theta)^\alpha).$$

When $\alpha = 0$, then $E_\alpha(h(\Theta))$ is the ordinary expectation of $h(\Theta)$ and $E_0(\Theta) = 1$. The α-expectation of Θ and its α-entropy are in the following relationship:

$$\text{ENT}_\alpha(\Theta) = \frac{1 - E_\alpha(\Theta)}{\alpha}. \tag{7}$$

The next Theorem shows that the density function $p(\cdot)^\alpha$, in the α-expectation, is the function $g(\cdot)$ needed for the third requirement of an entropy measure.

Theorem 3 *Let Θ be a random variable with positive density $p(\theta)$. Then, for any variable Y with positive density $p(y)$ such that $p(y)^{\alpha+1}$ and $p(y,\theta)^{\alpha+1}$ are integrable, we have*

$$\text{ENT}_\alpha(Y, \Theta) = \text{ENT}_\alpha(Y) + E_\alpha(\text{ENT}_\alpha(\Theta|Y)) \tag{8}$$

for any $\alpha > -1$, where the α-expectation is with respect the marginal distribution of Y.

Decomposition (8) can be written alternatively as

$$\text{ENT}_\alpha(Y, \Theta) = \text{ENT}_\alpha(Y) + E[p(y)^\alpha \text{ENT}_\alpha(\Theta|Y)]$$

thus becoming (3) for $g(y) \equiv p(y)^\alpha$. The well known decomposition of the Shannon entropy, given in (1), is a special case of (8) which is obtained for $\alpha \downarrow 0$. A by-product of decomposition (1) is that the joint entropy of two independent variables is the sum of the marginal entropies:

$$\text{ENT}_0(Y, \Theta) = \text{ENT}_0(Y) + \text{ENT}_0(\Theta).$$

This additivity does not hold for the α-entropy, which, in the independent case, decomposes into

$$\text{ENT}_\alpha(Y, \Theta) = \text{ENT}_\alpha(Y) + \text{ENT}_\alpha(\Theta) - \alpha \text{ENT}_\alpha(Y) \text{ENT}_\alpha(\Theta), \tag{9}$$

where the Shannon case in (1), is obtained as $\alpha \to 0$ is clearly exposed. Theorems 1, 2 and 3 guarantee that the α-entropy is an entropy measure according to our Definition 1. In the next section, we shall use this class of entropy measures to define new design criteria.

α-ENTROPY AND OPTIMAL EXPERIMENTS

In this section, we use the gain of information of Definition 2 when the entropy function $H(\cdot)$ is the α-entropy, for $\alpha > -1$. Theorem 3 indicates that the function $g(\cdot)$ to be used, together with the α-entropy, is $g(y) \equiv p(y)^\alpha$. Thus, we consider the quantity

$$G_\alpha(\xi) = E\{p(Y|\xi)^\alpha [H(\Theta) - H(\Theta|Y, \xi)]\} \tag{10}$$

By using the definition of α-expectation, this can also be written as

$$G_\alpha(\xi) = E_\alpha \left[\text{ENT}_\alpha(\Theta) - \text{ENT}_\alpha(\Theta|\xi, Y)\right]$$

for which $G_0(\xi)$ is a limiting case for $\alpha \downarrow 0$. We define $G_\alpha(\xi)$ as the *α-expected change of α-entropy* about Θ, provided by the experiment ξ. Clearly, an interesting case is $G_\alpha(\xi) > 0$, so that the posterior distribution of Θ is less entropic than the prior distribution. The next Theorem provides a characterization of this.

Theorem 4 *For random variables Θ and Y, the expected change of uncertainty $G_\alpha(\xi)$ is non-negative if and only if*

$$\text{ENT}_\alpha(\Theta, Y|\xi) \leq \text{ENT}_\alpha(\Theta) + \text{ENT}_\alpha(Y|\xi) - \alpha \text{ENT}_\alpha(\Theta) \text{ENT}_\alpha(Y|\xi) \tag{11}$$

provided that all integrals exist. (This is (9) with equality replaced by inequality.)

Proof. After some algebra, (10) reduces to

$$\begin{aligned} \frac{1}{\alpha} \int p(\theta, y|\xi)^{\alpha+1} dy d\theta &\geq \frac{1}{\alpha} \int p(y|\xi)^{\alpha+1} q(\theta)^{\alpha+1} dy d\theta \\ &= \frac{1}{\alpha} \int p(y|\xi)^{\alpha+1} dy \int q(\theta)^{\alpha+1} d\theta. \end{aligned}$$

The result then follows by conversion of the integrals to α-entropies. $\square$

By the identity (9) there follows that the expected change of uncertainty is zero when Θ and Y are independent. Note that the condition (11) in Theorem 4 does not always hold, thus suggesting the existence of experiments which increase the α-entropy of Θ. One example is discussed next.

Example 0.2 (Simple Hypotheses Testing) Suppose the experiment ξ consists of tossing a coin, one time, and the result can be head or tail. Let Y be the binary variable, taking value 1 when the result is head and 0 otherwise. We introduce a variable Θ to model the probability that $Y = 1$ as follows. We suppose that Θ is itself a binary variable, taking value θ_1 with probability θ_1 and $\theta_2 = 1 - \theta_1$, with probability $1 - \theta_1$. These two values of Θ can be regarded as two simple hypotheses $H_0 : \Theta = \theta_1$ and $H_2 : \Theta = \theta_2$, with prior probabilities θ_1 and θ_2. Conditional on $\Theta = \theta_1$, we define

$$p(Y = 1|\theta_1, p, \delta) = \frac{p\theta_1 - \delta}{\theta_1}$$

while, conditional on $\Theta = \theta_2$, we define

$$p(Y = 1|\theta_2, p, \delta) = \frac{p\theta_2 + \delta}{\theta_2}$$

for known p and δ, such that $p(Y = 1|\theta_1, p, \delta)$ and $p(Y = 1|\theta_2, p, \delta)$ are in the interval $[0; 1]$. Thus, the joint probability distribution of Θ, Y is the table

	Y		
Θ	1	0	$p(\Theta)$
θ_1	$p\theta_1 - \delta$	$(1-p)\theta_1 + \delta$	θ_1
θ_2	$p\theta_2 + \delta$	$(1-p)\theta_2 - \delta$	θ_2
	p	$1-p$	1

Take $\alpha = 1$, so that the gain of information is

$$G_1(\xi) = \delta(2\delta - (1 - 2p)(1 - 2\theta_1))$$

Then, in addition to the constraint induced on δ by $0 \leq p(Y = 1|\theta_j) \leq 1$ for $i, j = 1, 2$, condition (11) gives the further constraint $\delta(2\delta - (1 - 2p)(1 - 2\theta_1)) \geq 0$. For example, let $p = \theta_1 = 1/3$ and $\delta > 0$, so that the conditions $0 \leq p(Y = 1|\theta_j) \leq 1$ for $i, j = 1, 2$ restricts the range of admissible values for δ to the interval $[0; 1/9]$. The expected gain of information is the function $4\delta^2 - 2/9\delta$, which is depicted in Figure 1, and the constraint $\delta(2\delta - (1 - 2p)(1 - 2\theta_1)) \geq 0$ reduces to $\delta > 1/18$. If $\delta > 1/18$, then, the expected gain of information is positive and, therefore, it is worthwhile conducting the experiment. The gain increases with δ and it is maximum when $\delta = 1/9$. On the other hand, when $\delta \leq 1/18$, the expected gain of information is negative, and the minimum is achieved when $\delta = 1/36$. Thus, with this information measure, the conclusion is that, when $0 < \delta \leq 1/18$ it is not worth conducting the experiment because, on average, the posterior

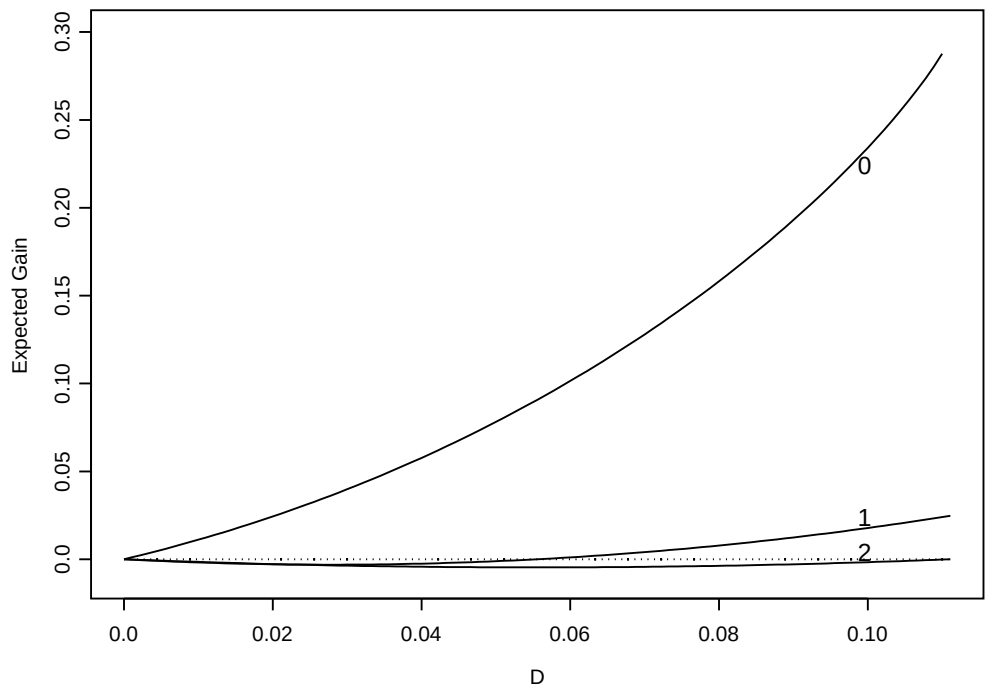

FIGURE 1. Gain of information in Example 0.2, for $\alpha = 0, 1$ and 2. The dotted line is $G = 0$, so that when $G_\alpha(\xi)$ is above the dotted line, the expected gain of information is positive, and negative otherwise.

distribution of Θ will be more entropic than the prior distribution, with the consequence of reducing our ability to discriminate between the two hypotheses H_0 and H_1. When $\alpha = 2$, then, using the same values of p and θ_1, and the constraint $\delta > 0$, it is easy to show that $G_2(\xi) = \delta(3/2\delta - 1/6)$, so that the expected gain of information is negative when $\delta < 1/9$ and, therefore, with this entropy measure, it is never worthwhile conducting the experiment. On the other hand, when $\alpha = 0$, then the expected gain of information is

$$\begin{aligned} G_0(\xi) &= \log(\frac{1}{4}2^{(2/3)}) + d\log(81) + \frac{1}{9}\log(1-9\,d)\times(1-9\,d) + \frac{1}{9}\log(2-9\,d) \\ &\times \quad (2-9\,d) + \frac{1}{9}\log(2+9\,d)\times(2+9\,d) + \frac{1}{9}\log(4-9\,d)\times(4-9\,d) \end{aligned}$$

which is non-negative for any $\delta > 0$, so that it is always worthwhile conducting the experiment, when $\delta > 0$. The values of $G_0(\xi)$ and $G_2(\xi)$ are in Figure 1, for different values of δ. □

In the next section, we consider more traditional regression type experiments.

REGRESSION TYPE EXPERIMENTS

Suppose the random p-vector Θ has a normal distribution with expectation θ_0 and variance-covariance matrix R_0^{-1} and the random n-vector $Y|\theta, \xi$ has a normal distri-

bution with expectation $X\theta$ and known variance-covariance matrix Σ. In this context, the experiment ξ is represented by the design matrix X, at which observations are taken. We suppose that both Σ and R_0 are known, and positive definite. Thus, the marginal distribution of Y is normal, with expectation $X\theta_0$ and variance-covariance matrix $XR_0^{-1}X^T + \Sigma$, and the posterior distribution of Θ is normal, with variance-covariance matrix R_1^{-1}, where $R_1 = R_0 + X^T\Sigma^{-1}X$ and expectation $\theta_1 = R_1^{-1}(R_0\theta_0 + X^Ty)$. The expected change of uncertainty is, by (10),

$$G_\alpha(X) = \frac{1}{\alpha}E_\alpha\left(\int p(\theta|y,X)^{\alpha+1}d\theta - \int q(\theta)^{\alpha+1}d\theta\right)$$

Now note that if the n-vector X has normal distribution with expectation μ and variance-covariance matrix V, then

$$\int p(x)^{\alpha+1}dx = [(2\pi^n)\det(V)]^{-\alpha/2}(\alpha+1)^{-n/2}.$$

By using repeatedly this result, it can be shown that $G_\alpha(X) > 0$ if and only if

$$\frac{(2\pi^{n+p})^{-\alpha/2}}{\alpha(\alpha+1)^{(n+p)/2}}\frac{\det(R_1)^{\alpha/2} - \det(R_0)^{\alpha/2}}{\det(\Sigma + XR_0^{-1}X^T)^{\alpha/2}} > 0$$

and, therefore, if and only if

$$\frac{1}{\alpha}\frac{\det(R_1)^{\alpha/2} - \det(R_0)^{\alpha/2}}{\det(\Sigma + XR_0^{-1}X^T)^{\alpha/2}} > 0. \tag{12}$$

By well known properties of the determinant, we can write $\det(R_1) = \det(R_0 + X^T\Sigma^{-1}X) = \det(R_0)\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})$ and $\det(\Sigma + XR_0^{-1}X^T) = \det(\Sigma)\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})$, where $R_0^{-1/2}$ is the inverse symmetric square root of R_0. Thus, (12) holds if and only if

$$\frac{1}{\alpha}\frac{\det(R_0)^{\alpha/2}[\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})^{\alpha/2} - 1]}{\det(\Sigma)^{\alpha/2}\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})^{\alpha/2}} > 0.$$

Since $\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2}) > \det(I_p) = 1$, the last inequality holds for any $\alpha > -1$. Thus, the expected uncertainty change in learning with conjugate normal distributions is always non-negative. Furthermore, if R_0 and Σ are both independent of X, maximizing $G_\alpha(X)$ reduces to maximizing

$$C_\alpha(X) = \frac{1}{\alpha}\frac{[\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})^{\alpha/2} - 1]}{\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})^{\alpha/2}}$$

The function $C(X)$ is increasing in $\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})$, so that maximizing $C_\alpha(X)$ is equivalent to maximizing $\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2})$. Given the identity $\det(I_p + R_0^{-1/2}X^T\Sigma^{-1}XR_0^{-1/2}) = \det(R_1)/\det(R_0)$, maximization of $C_\alpha(X)$ is also

obtained by maximizing $\det(R_1)$. Thus, for linear regression models with conjugate priors, the design criterion

$$\max_X \det(R_0 + X^T \Sigma^{-1} X)$$

corresponds to maximizing the α-expected change of α-entropy, for any $\alpha > -1$. This criterion is the well known Bayesian D-optimality, see [1], and it is known to be equivalent to maximizing the expected amount of Shannon information produced by an experiment. Our result shows that Bayesian D-optimality, in this context, has an information-based justification for more general measures of information.

ACKNOWLEDGEMENTS

The authors wish to thank Marco Ramoni and Paul Snow for the fruitful discussions about information.

REFERENCES

1. K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, **10**, pp. 273–304, 1995.
2. A. P. Dawid and P. Sebastiani, "Coherent dispersion criteria for optimal experimental design," *Annals of Statistics*, **27**, pp. 65–81, 1999.
3. D. V. Lindley, "On a measure of information provided by an experiment," *Annals of Mathematical Statistics*, **27**, pp. 986–1005, 1956.
4. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, **27**, pp. 379–423, 623–656, 1948.
5. S. Kullback, *Information Theory and Statistics*, Dover, New York, NY, 1968.
6. A. Renyi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statististics and Probability*, vol. 1, (Berkeley, CA), pp. 547–561, University of California Press, 1961.
7. R. M. Losee, "A discipline independent definition of information," *Journal of the American Society for Information Science*, **48**, (3), pp. 254–269, 1997.
8. P. Sebastiani and H. P. Wynn, "Renyi-type entropies and distances in Bayesian learning," Technical Report, Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2B, June 2000. Submitted for publication.
9. I. J. Good, "Rational decisions," **14**, pp. 107–114, 1952.
10. L. Savage, *The Foundations of Statistics*, Dover, New York, NY, 2nd revised ed., 1972.
11. A. D. Hendrickson and R. J. Buehler, "Proper scores for probability forecasters," *Annals of Statistics*, **42**, pp. 1916–1921, 1972.
12. W. F. Caselton and J. V. Zidek, "Optimal monitoring network designs," *Statistics and Probability Letters*, **2**, pp. 223–227, 1984.
13. M. C. Shewry and H. P. Wynn, "Maximum entropy sampling," *Journal of Applied Statistics*, **14**, pp. 165–170, 1987.
14. J. Sacks, W. Welch, T. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments (with Discussion)," *Statistical Science*, **4**, pp. 409–423, 1989.
15. R. A. Bates, R. J. Buck, E. Riccomagno, and H. P. Wynn, "Experimental design and and observation for large systems (with Discussion)," *Journal of the Royal Statistical Society, B*, **58**, pp. 77–111, 1996.
16. T. M. Cover and M. Thomas, *Elements of Information Theory*, Wiley, New York, NY, 1991.

17. L. Pronzato, H. P. Wynn, and A. A. Zhigljavsky, *Dynamical Search*, Chapman and Hall/CRC, Boca Raton, Fl, 1999.
18. B. DeFinetti, *Theory of Probability*, vol. 1, Wiley, New York, NY, 1974.
19. P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design," *Journal of the Royal Statistical Society, B*, **62**, pp. 145–157, 2000.