Bayesian Analysis (2007) **2**, Number 1, pp. 167–212

On the Measure of the Information in a Statistical Experiment

Josep Ginebra*

Abstract. Setting aside experimental costs, the choice of an experiment is usually formulated in terms of the maximization of a measure of information, often presented as an optimality design criterion. However, there does not seem to be a universal agreement on what objects can qualify as a valid measure of the information in an experiment. In this article we explicitly state a minimal set of requirements that must be satisfied by all such measures. Under that framework, the measure of the information in an experiment is equivalent to the measure of the variability of its likelihood ratio statistics or which is the same, it is equivalent to the measure of the variability of its posterior to prior ratio statistics and to the measure of the variability of the distribution of the posterior distributions yielded by it. The larger that variability, the more peaked the likelihood functions and posterior distributions that tend to be yielded by the experiment, and the more informative the experiment is. By going through various measures of variability, this paper uncovers the unifying link underlying well known information measures as well as information measures that are not yet recognized as such.

The measure of the information in an experiment is then related to the measure of the information in a given observation from it. In this framework, the choice of experiment based on statistical merit only, is posed as a decision problem where the reward is a likelihood ratio or posterior distribution, the utility function is convex, the utility of the reward is the information observed, and the expected utility is the information in an experiment. Finally, the information in an experiment is linked to the information and to the uncertainty in a probability distribution, and we find that the measure of the information in an experiment is not always interpretable as the uncertainty in the prior minus the expected uncertainty in the posterior.

Keywords: Convex ordering, design of experiments, divergence measure, Hellinger transform, likelihood ratio, measure of association, measure of diversity, measure of surprise, mutual information, optimal design, posterior to prior ratio, reference prior, stochastic ordering, sufficiency, uncertainty, utility, value of information.

1 Introduction

In the statistical science community there is a pervading feeling that the concept of "information carried by an experiment" is something intangible that can not be characterized. Review articles often list various measures, each addressing a particular aspect of what information means, but they do not identify any commonality among these measures. Papaioannou (2001) describes the current understanding by stating that "*While information is a basic and fundamental concept in statistics, there is no universal agree-*

*Departament d'Estadística, Universitat Politècnica de Catalunya. mailto:josep.ginebra@upc.edu

ment on how to define and measure it in a unique way". Clearly, there exists a need for an agreement on what qualifies as an information measure and of the features that make an experiment more informative than another.

Let $E = (X; P_\theta)$ denote a *statistical experiment* observing a random variable X with an unknown distribution P_θ, where the parameter $\theta \in \Omega$ is an index for the list of possible distributions of X. When experimenting, the goal is to learn about the unknown θ that explains X. Since many aspects of the association between X and θ help in identifying the P_θ responsible for producing an observation $X = x$, the information about θ in E is typically a highly multidimensional concept that can not be possibly captured completely by any single real valued quantity.

Nevertheless, to rank experiments in terms of the information "they carry", one has to do it through real valued measures that capture the one aspect of the information in E that one cares the most. It follows the need for a framework that encompasses all valid measures of the information in E that can be used as scales to induce a total information ordering on the space of available experiments, and maybe choose one of them. This paper makes that framework explicit by building on the sufficiency ordering of experiments considered in Blackwell (1951, 1953) and Le Cam (1964, 1986).

Section 2 introduces the background and notation on statistical experiments. Section 3 reviews the sufficiency ordering of experiments, which is also called the 'always at least as informative' ordering. That section also presents the Blackwell-Sherman-Stein and Le Cam theorem, establishing that the sufficiency ordering of experiments is equivalent to the convex ordering of their likelihood ratio statistics and to the convex ordering of the distribution of the posterior distributions attained under a given prior.

Definition 4.1 in Section 4 identifies a minimum set of requirements that must be satisfied by every *measure of the information in* E, making the sufficiency ordering into the only essential ingredient in the characterization of these measures. That characterization does neither assume that θ is a random variable nor that the experiment will be used in a statistical decision problem, even though it can be given decision theoretic and/or Bayesian interpretations.

Section 5 then explains how, as a consequence of this characterization, measuring the information about θ in E is essentially the same as measuring the variability of its likelihood ratio statistics, as in Definition 5.1-5.2. It follows that measuring the information in E is also the same as measuring the variability of its posterior to prior ratio statistics, and it is the same as measuring the variability of the distribution of the posterior distributions yielded by it. The larger that variability, the more peaked the likelihood functions and the posterior distributions that tend to be yielded by E, and the more informative E is. By considering various measures of the variability of these statistics, we present a broad spectrum of features associated to the informativity of E, and we uncover the framework underlying all the measures of the information being used by the design of experiments literature (DoE from now on), as well as measures of the information not yet recognized as such by them.

In this manuscript the comparison of experiments is always made based on statistical

merit only, irrespective of experimental costs. In practice, choosing among experiments requires compromising between the information they carry and their cost, but when comparing experiments just in terms of the information in them, no apologies are to be made for doing as if their cost was the same.

A source of confusion is that the term *information about* θ is used to denote differing concepts. A secondary contribution of the paper is to help distinguish and relate

- the measure of the *information about* θ *in experiment* $E = (X; P_\theta)$, also recognized as the *statistical information* or the *expected information in* E, which is relevant for comparing experiments in terms of statistical merit, it is our main object of interest, and which is dealt with in Sections 3 to 5 and 7,
- the measure of the *information about* θ *in an observation* $X = x$, also recognized as the *observed information in* $X = x$, which is relevant after the experiment is selected and carried out as a Bayesian model checking test statistic, and which is dealt with in Section 6.1, and
- the measure of the *information about* θ *in a given distribution* h *on* Ω, also recognized by Shannon as the *self-information about* θ *in its own distribution*, which is relevant when assessing the strength of knowledge about θ and as a measure of the homogeneity in a population h, and which is dealt with in Section 6.2.

Most of the statistical literature concentrates on inference for a given experiment and as a consequence, its main focus is the measure of the *information in* $X = x$. In the non-Bayesian DoE literature the information in E is then typically measured through real valued transformations of Fisher information matrices introduced in Kiefer (1959), and through divergence measures introduced in Csiszár (1963, 1967). On the other hand, the information theory literature stemming from Shannon (1948) starts by characterizing the *information about a random variable in its own distribution* through the negative of its entropy. In the Bayesian DoE literature the information in E is then measured through the cross entropy between X and θ as in Lindley (1956), in an approach generalized in Raiffa and Schlaiffer (1961) and in DeGroot (1962, 1984), where the information in E is measured through the negative of the Bayes risk and it is interpreted as the uncertainty in the prior minus the expected uncertainty in the posterior.

Different from that, this manuscript starts from and focuses on a characterization of the measure of the *information in experiment* E that encompasses as special cases the measures of information in Kiefer (1959), Csiszár (1963, 1967), Lindley (1956), Raiffa and Schlaiffer (1961) and DeGroot (1962).

In Section 6.1, the measure of the information about θ in an observation $X = x$ from E is defined to be a non-negative convex function of the corresponding likelihood ratio or posterior distribution, in Definition 6.1-6.2. All added, it turns that the choice of the most informative experiment can be posed as a decision problem where the reward from choosing experiment E is its likelihood ratio or posterior distribution statistic, the utility function is convex, the utility of the reward is the information in the observed outcome, and the expected utility from choosing E is the information in E.

In Section 6.2, the information about θ in a distribution h on Ω is defined to be the information in an observed outcome that updates a baseline prior into a posterior h. The uncertainty about θ in h is then measured as the information in a one-point distribution minus the information in h.

Section 7 explores the relationship between the information in E and the expected impact of E on the uncertainty about θ in its own distribution. By showing that the information in E is not always interpretable as the uncertainty in the prior *minus* the expected uncertainty in the posterior, this section clarifies the sense in which our definition of information generalizes the definition of information proposed in De Groot (1962) and adopted most often in Bayesian DoE.

Section 8 illustrates through an example how the framework described in this manuscript allows for a unified approach to the selection of an experiment, to the construction of a reference prior for a given experiment, to the assessment of the validity of the model and to the quantification of the impact of $X = x$ on the knowledge about θ.

It is important to emphasize that even though this manuscript might look like a review paper to some, Definitions 4.1, 5.1-5.2 and 6.1-6.2, their motivation, some examples and the interpretation of Proposition 3.1 are new. Readers mainly interested in statistical inference might want to read Sections 3.2.1, 6.1 and 5 first, because they provide a more intuitive starting point to the manuscript that does not rely on the axiomatic framework built in Sections 3 and 4. In that alternative presentation one first defines the measure of the information in a given observation $X = x$, and then presents the measure of the information in experiment E as the average of the information in all the observations that could have been yielded by it.

2 Background and notation

2.1 Statistical experiments

Definition 2.1. *A statistical experiment $E = \{(X, S_X); (P_\theta, \Omega)\}$ yields an observation on a random variable X defined on S_X, with an unknown probability distribution that is known to be in the family $(P_\theta, \theta \in \Omega)$.*

Following Wald (1950) and Blackwell (1951), here an experiment E is considered to be a family of probability measures, $(P_\theta, \theta \in \Omega)$, on a common sample space, S_X, one of which is assumed to be the distribution of X. One might think of each θ in Ω as representing a possible explaining "theory" and the parameter space Ω as representing the set of all conceivable "theories". When comparing experiments E and $F = \{(Y, S_Y); (Q_\theta, \Omega)\}$, the only necessary common thread between them is that the parameter space be the same, and that the same unknown θ governs the distribution of X and Y.

Note that here, a statistical experiment coincides with what the inference literature calls a parametric model. Thus, "a measure of the information about θ in experiment

E" is synonymous of "a measure of the information about θ in the statistical model $(P_\theta, \theta \in \Omega)$". To avoid measure theoretical details, throughout this paper we assume that the probability measures $(P_\theta, \theta \in \Omega)$ are dominated by a σ-finite measure μ (i.e., that there is a measure μ such that events of μ-measure zero also have P_θ-measure zero), and thus the corresponding density functions, p_θ, will always exist. We also assume that sample spaces are complete separable metric spaces. That covers all situations faced in the usual statistical inference and design of experiments practice.

As an example, *linear normal experiments* are the ones that yield $X \in R^n$ distributed as a $N_n(A\beta, \sigma^2 I)$, where A is a known $n \times p$ design matrix and β is a vector of regression parameters. Here θ denotes either β with $\Omega = R^p$ or (β, σ) with $\Omega = R^p \times [0, \infty)$, depending on whether σ is assumed known or unknown, and selecting an experiment consists of choosing a design matrix A.

An experiment is said to be *totally non-informative*, denoted by E_{tni}, if the distributions of X are the same for all θ. One can not learn about θ by observing from E_{tni}, and it is the baseline relative to which the information in every experiment is measured. At the other end, an experiment is said to be *totally informative*, denoted by E_{ti}, if for every pair $(\theta_i, \theta_j) \in \Omega \times \Omega$ the intersection of the support sets for P_{θ_i} and P_{θ_j} is an empty set, and thus if it is a family of mutually singular distributions. After performing E_{ti}, the P_θ that generated $X = x$ can be identified with certainty.

In Bayesian setups, the uncertainty about $\theta \in \Omega$ is modelled through a prior distribution π on Ω, which allows one to represent the experiment or statistical model $(P_\theta, \theta \in \Omega)$ through the marginal distribution of X, P_π, with density function $p_\pi(x) = E_\pi[p_\theta(x)]$, and to construct the joint distribution for (X, θ) with density function

$$f_\pi(x, \theta) = p_\theta(x)\pi(\theta) = p_\pi(x)\pi_E(\theta|x), \tag{1}$$

where $\pi_E(\theta|x)$ denotes the density of the posterior distribution of θ. Let the sampling, predictive and posterior densities of $F = (Y; Q_\theta)$ be denoted by $q_\theta(y)$, $q_\pi(y)$ and $\pi_F(\theta|y)$.

Note that when designing an experiment, data as well as parameters are unknown and the reasons against treating them symmetrically by considering both X and θ as random and by averaging over both parameter and sample spaces are a lot less compelling than for inference problems. In fact, it is our perception that the only difference between the Bayesian and the non-Bayesian way of planning for an experiment is in the way one interprets the optimality design criteria available.

2.2 Likelihood functions attainable and information in E

Given an observation $X = x$ yielded by experiment E, likelihood functions, $l_x(\theta)$, are functions of θ proportional to $p_\theta(x)$ with the constant of proportionality being arbitrary. Before performing the experiment, $l_X(\theta)$ can be regarded as a random function on Ω. For totally non-informative experiments, E_{tni}, likelihood functions are always constant, while for totally informative experiments, E_{ti}, likelihood functions are always zero everywhere except for one value θ in Ω. For any experiment, the relatively "flatter" the attainable likelihoods, the harder it is to identify the θ that explains the observations,

and the worse that experiment is for inferential purposes.

Adherents to likelihood based inference will recognize informative experiments to be the ones that provide highly concentrated likelihood functions. Given a choice, they will prefer experiments that tend to produce likelihoods with more pronounced peak(s) and the measures of the information in E should quantify this tendency. Comments in this direction can be found in Barnard (1959), in Barnard, Jenkins, and Winsten (1962, p. 323), and in Birnbaum (1962, pp. 293, 304).

When θ is real valued and Ω is open, the *relative peakedness* of the likelihood at θ can be measured through the squared relative rate of change of the likelihood function,

$$r_x(\theta) = (\dot{l}_x(\theta)/l_x(\theta))^2 = (\dot{p}_\theta(x)/p_\theta(x))^2, \tag{2}$$

where the dot indicates the derivative with respect to θ. Before the experiment is performed, x is unobserved and $r_X(\theta)$ is a random function on Ω with possibly a different distribution under each p_θ. Under regularity conditions (see, e.g., Lehmann 1983), one can assess the information about θ in E through the average of $r_X(\theta)$ under p_θ,

$$I^\theta_{Fi}(E) = E_{p_\theta}[(\frac{\dot{p}_\theta(x)}{p_\theta(x)})^2] = Var_{p_\theta}[\frac{\dot{p}_\theta(x)}{p_\theta(x)}], \tag{3}$$

which was introduced in Fisher (1922) and is called the Fisher information in E. The larger $I^\theta_{Fi}(E)$, the smaller the asymptotic variance for the maximum likelihood estimator of θ, and the more informative E is. When Ω is an open subset of R^p, the Fisher information is defined to be the covariance matrix of the vector of ratios between the partial derivatives of $l_x(\theta)$ and $l_x(\theta)$.

The DoE literature largely focuses on using real valued transformations of the Fisher information matrix, introduced in Kiefer (1959). However, Fisher information might not exist and when it does exist, it typically depends on the unknown value of θ, in which case different experiments could be optimal for different values of θ. Furthermore, the performance of E for tasks other than estimation under squared error loss should be assessed on the basis of how well those tasks can be performed, which may involve aspects of the association between X and θ not captured by Fisher information.

3 Sufficiency ordering of experiments and information

In statistical decision theoretic terms, the information about θ in an experiment E depends on the performance of E in relation to the terminal consequences of the statistical decisions made based on the data obtained from it. Sometimes one needs to select an experiment E based on its expected performance on a given decision problem with loss function $L(\theta, d)$ defined on $\Omega \times D$, where d is a decision and D is the space of decisions.

To make a decision based on data from $E = (X; P_\theta)$ one selects a decision rule, $\delta(x)$, that assigns to each $x \in S_X$ a possible decision d, and the performance of $\delta(X)$ under each $\theta \in \Omega$ is appraised through its risk function, $R_E(\theta, \delta) = E_{p_\theta}[L(\theta, \delta(x))]$.

In principle, one could assess the performance of E on that given terminal decision problem through the class of risk functions of its admissible rules (i.e., the rules that can not be improved upon for every θ), but that does not lead to any clear cut choice between two experiments. To narrow that choice down, one could compare experiments E and F on the basis of the risk function of a given pair of admissible rules, like their respective Bayes rules (minimizing the weighted average of $R_E(\theta, \delta)$ under a given distribution on Ω), or their minimax rules (minimizing the maximum of $R_E(\theta, \delta)$ over Ω). Most often though, one would still find E to be either better or worse than F depending on the value of θ. To attain a total ordering of the experiments available, one must compare them on the basis of a real number like the average risk for their Bayes rules (i.e., their Bayes risk), or the maximum risk for their minimax rule (i.e., their minimax risk).

By considering the choice between any two experiments, E and F, based on the Bayes or the minimax risk for a specific terminal decision problem, one might chose experiment E or F depending on the problem at hand. However, sometimes by observing X from E one can do at least as well as by observing Y from F for every terminal decision problem, and thus in particular for every statistical inference problem. These situations lead to the ordering of experiments considered next.

3.1 When is experiment E "sufficient for" or "always at least as informative as" experiment F?

Definition 3.1 (Blackwell 1951, 1953). *Experiment $E = (X; P_\theta)$ is said to be "sufficient for" $F = (Y; Q_\theta)$ if there exists a stochastic transformation of X to a random variable $W(X)$ such that $W(X)$ and Y have identical distribution under each $\theta \in \Omega$.*

Lehmann (1988, p. 521) re-phrases this by stating that "experiment E is sufficient for F if there exists a random variable Z with a known distribution and a function $g(\cdot, \cdot)$ such that for all $\theta \in \Omega$, X being distributed as P_θ implies that $g(X, Z)$ is distributed as Q_θ." In fact, there is no loss of generality in assuming that the distribution of Z is uniform on $(0, 1)$ and independent of X. Thus, E is "sufficient for" F whenever by using a realization $X = x$ of experiment E and an auxiliary randomization, Z, one can simulate data distributed as Y without knowing θ.

The sufficiency ordering of experiments is the central subject of study of the *comparison of experiments* literature stemming out of Blackwell's seminal papers. Brief expositions can be found in Blackwell and Girshik (1954), Savage (1954), Lehmann (1959, 1988), DeGroot (1970), LeCam (1975, 1996), Torgersen (1976), Vajda (1989), Shiryaev and Spokoiny (2000) and Gollier (2001). For a thorough presentation, see Heyer (1982), Strasser (1985), LeCam (1986) or Torgersen (1991a). For a review with an exhaustive list of references and examples, see Goel and Ginebra (2003).

In statistical inference, experiment $E = (X; P_\theta)$ is typically fixed and given and discussions are limited to the comparison between E and the sub-experiment E_T that yields a statistic based on X, $T(X)$. Clearly E is always "sufficient for" E_T in the sense of Definition 3.1. Furthermore, when $T(X)$ is a sufficient statistic for X, once

given $T(x) = t$ one can also generate data distributed as X without knowing θ, and hence in that case experiment E_T is also "sufficient for" E. Therefore, stating that a statistic $T(X)$ is sufficient for X is equivalent to stating that experiments E and E_T are "sufficient for" each other. But Definition 3.1 applies more generally, since it allows for the comparison of experiments on unrelated sample spaces. (In fact, as remarked in Le Cam (1975), 'to state that E is "sufficient for" F is the same as to state that there exists an experiment EF yielding (X, Y) for which X is a sufficient statistic for EF').

As an example of a sufficiency ordering of experiments, let E and F be a pair of linear normal experiments that observe X and Y from $N_{n_E}(A\beta, \sigma^2 I)$ and $N_{n_F}(B\beta, \sigma^2 I)$ respectively, where A and B are known $n_E \times p$ and $n_F \times p$ matrices. Hansen and Torgersen (1974) prove that when σ^2 is known E is "sufficient for" F if and only if $I^{\theta}_{Fi}(E) - I^{\theta}_{Fi}(F) = A^{'}A - B^{'}B$ is non-negative definite, and that for unknown σ^2 an additional condition is that $n_E \geq n_F + rank(A^{'}A - B^{'}B)$.

The Definition 3.1 making E sufficient for F if one can derive from X a random variable with the same distribution as Y using a known random mechanism that only depends on X, is grounded on a randomization argument seemingly devoid of statistical meaning. Nevertheless, that meaning follows from the well established fact that E is "sufficient for" F if, and only if E is "always at least as good as" F in the sense that for every decision problem and for every decision rule $\delta(Y)$ based on F, there exists a decision rule $\delta_*(X)$ based on E such that $R_E(\theta, \delta_*(x)) \leq R_F(\theta, \delta(y))$ for all θ. Consequently, when E is "sufficient for" F the Bayes and minimax risks under E are at most as large as the ones under F for every prior and loss.

Therefore, E is "sufficient for" F if and only if E is preferable to F for every statistical decision problem with a loss defined on $\Omega \times D$, which includes non-sequential estimation, testing, classification, the prediction of future observations and any other purely inferential problem, where learning about θ is the single goal of experimenting. Hence, the phrase *"E is always at least as informative as F"* is used as a synonym for *"E is sufficient for F"* (see, e.g., Blackwell and Girshik 1954; Lehmann 1988). This also explains why any ordering of experiments that respects the sufficiency ordering is called an information ordering in Torgersen (1991a), and why the sufficiency ordering will be essential in the characterization of the measure of the information in an experiment.

Comparisons in the sense of this sufficiency ordering are always made on the basis of statistical merit only, ignoring experimental costs. By stating that E is "sufficient for" or "always at least as informative as" F if and only if E is preferable to F for every decision problem with a loss on $\Omega \times D$, one excludes from consideration the comparison of experiments under one-period bandit and stochastic control type problems, which have loss functions defined on $\Omega \times S_X$ (see, e.g., Gonzalez and Ginebra 2001), and the comparison under mixed problems with the goal of maximizing both information about θ and outcome, which have loss functions defined on $\Omega \times D \times S_X$ (see, e.g., Verdinelli and Kadane 1992). One also excludes the comparison of experiments in terms of their performance under problems with loss functions that depend on the experiment itself, like the ones that include in the loss experimental costs that depend on sample size.

3.2 Variability of likelihood ratio statistics and information in E

The Blackwell-Sherman-Stein and Le Cam theorems presented below establish the equivalence between the sufficiency ordering of experiments, the convex ordering of their likelihood ratio statistics, and the convex ordering of the distribution of the posterior distributions attained under a given prior. That will enable us to propose measuring the information in E through measures of the variability of its likelihood ratio and posterior distribution statistics, as described in Section 5.

Here we focus on the case where Ω has a finite number of elements, $\Omega = \{\theta_1, \ldots, \theta_k\}$, because it is an important case in its own (under it, Fisher information is not even defined), and because it provides the basic tool for the cases where Ω is infinite.

3.2.1 The vector of likelihood ratios as a likelihood function

Let $E = (X; P_\theta)$ be an experiment on Ω, and let $P_\pi = \sum_{i=1}^k \pi_i P_{\theta_i}$ be a convex combination of the elements in $(P_{\theta_i}, \theta_i \in \Omega)$ that dominates all the elements in that family (i.e., P_π is such that any measurable set of P_π-measure zero has P_{θ_i}-measure zero for all i). In that case, the 'likelihood to averaged likelihood ratio statistic'

$$T_\pi(X) = \frac{1}{p_\pi(X)}(p_{\theta_1}(X), \ldots, p_{\theta_k}(X)) = \frac{1}{E_\pi[l_X(\theta)]}\ (l_X(\theta_1), \ldots, l_X(\theta_k)), \tag{4}$$

is minimal sufficient for E (see, e.g., Basu 1975; Lehmann 1983), and its distribution characterizes the statistical properties of E. In particular, P_π can be any convex combination with strictly positive weights (i.e. with $\pi_i > 0$ for $i = 1, \ldots, k$), and we focus on this case, but when all measures in E are dominated by one of them, say P_{θ_1}, then

$$T_{\theta_1}(X) = \frac{1}{p_{\theta_1}(X)}(p_{\theta_2}(X), \ldots, p_{\theta_k}(X)) \tag{5}$$

is also minimal sufficient for E and all the discussion that follows in terms of $T_\pi(X)$ can be rephrased in terms of $T_{\theta_1}(X)$ without any loss of generality. The range of values taken by $T_\pi(X)$, $\{T_\pi(x), x \in S_X\}$, is a subset of the set

$$K_\pi = \{u = (u_1, \ldots, u_k) \in R^k : u_i \geq 0 \ \text{ and } \ \sum_{i=1}^k \pi_i u_i = 1\}, \tag{6}$$

which is the convex hull of $\{(1/\pi_1, \ldots, 0), \ldots, (0, \ldots, 1/\pi_k)\}$. The range of values taken by $T_{\theta_1}(X)$ when X ranges in S_X, $\{T_{\theta_1}(x), x \in S_X\}$, is a subset of R_+^{k-1}.

As a function of θ, with x fixed, the vector of observed likelihood ratios $T_\pi(x)$ is proportional to $(p_{\theta_1}(x), \ldots, p_{\theta_k}(x))$ with constant of proportionality $1/p_\pi(x)$ and therefore, $T_\pi(x)$ is a standardized version of the likelihood function. *As a function of* x, the j-th coordinate of $T_\pi(x)$, $p_{\theta_j}(x)/p_\pi(x)$, is the density function for the distribution of X under $\theta = \theta_j$ when the dominating measure is P_π instead of μ and $T_\pi(x)$ becomes the list of conceivable density functions for X. (Note that the assumption that inferences should not depend on the dominating measure is tantamount to the constant of proportionality of the likelihood function being irrelevant).

Remark: A choice of a particular set of weights, $(\pi_1, \ldots, \pi_k)$, is just a matter of choice of a dominating measure P_π, and of a version of standardized likelihood function $T_\pi(x)$, and all that can be devoid of any Bayesian connotation. In fact, in the sufficiency ordering literature one typically restricts attention to uniform weights, $\pi_i = 1/k$. On the other hand, in Bayesian terms one is entitled to interpret π as a prior distribution and $T_\pi(X)$ as the 'posterior to prior ratio statistic',

$$T_\pi(X) = (\frac{\pi_E(\theta_1|X)}{\pi_1}, \ldots, \frac{\pi_E(\theta_k|X)}{\pi_k}). \tag{7}$$

3.2.2 The Blackwell-Sherman-Stein and Le Cam theorem

Here, we compare the experiments on Ω $E = (X; P_\theta)$ and $F = (Y; Q_\theta)$, through the distribution of their corresponding likelihood ratio statistics, $T_\pi(X)$ and $S_\pi(Y)$. For totally non-informative experiments, E_{tni}, the likelihood function is constant, and $T_\pi(X) = (1, \ldots, 1)$ with probability one. For totally informative experiments, E_{ti}, the likelihood is zero everywhere except at one $\theta_j \in \Omega$, and $T_\pi(X) = (0, \ldots, 1/\pi_j, \ldots, 0)$, which is an extreme point of K_π. In general, the further $T_\pi(X)$ *tends* to fall away from $(1, \ldots, 1)$ towards an extreme point of K_π, the easier it is to guess θ and the more informative E is. Given that for every experiment

$$E_{p_\pi}[T_\pi(x)] = (1, \ldots, 1), \tag{8}$$

the more informative E is, the more spread out is the distribution of $T_\pi(X)$ (under $X \sim P_\pi$) away from $(1, \ldots, 1)$ towards extreme points of K_π. Therefore it should not come as a surprise that the Blackwell-Sherman-Stein theorem, enunciated next, relates "E being always at least as informative as F" to the distribution of $T_\pi(X)$ when X is P_π-distributed being more variable than the distribution of $S_\pi(Y)$ when Y is Q_π-distributed, where $Q_\pi = \sum_{i=1}^k \pi_i Q_{\theta_i}$.

Proposition 3.1. *Experiment $E = (X; P_\theta)$ is "sufficient for" experiment $F = (Y; Q_\theta)$ if and only if for some strictly positive set of weights π,*

$$E_{p_\pi}[\phi(T_\pi(x))] \geq E_{q_\pi}[\phi(S_\pi(y))] \tag{9}$$

for every convex function $\phi(\cdot)$ on K_π.

Equivalently, this proposition can be re-stated in terms of the convex ordering of likelihood ratio statistics as "experiment E is sufficient for F if and only if the distribution of $T_\pi(X)$ when $X \sim P_\pi$ is larger in the convex order than the distribution of $S_\pi(Y)$ when $Y \sim Q_\pi$, i.e., if and only if

$$T_\pi(X)|_{p_\pi} \geq_{cx} S_\pi(Y)|_{q_\pi}." \tag{10}$$

Since convex functions take on their larger values over "extreme regions", any measure of the form $E[\phi(U)]$ with a convex $\phi(\cdot)$ can be interpreted as a measure of the dispersion of the random variable U. Consequently, "E is sufficient for F" is equivalent to "the

distribution of $T_\pi(X)$ under P_π is always more variable than the distribution of $S_\pi(Y)$ under Q_π, no matter how variability is measured". In particular, when E is sufficient for F, the distribution of any set of coordinates of $T_\pi(X)$ is always more variable than the distribution of that same set of coordinates of $S_\pi(Y)$. For details on the convex order of distributions and its interpretation as a variability order, see Shaked and Shantikumar (1994, chap. 2 and 5) and the Appendix.

Blackwell (1951, 1953), Sherman (1951) and Stein (1951) prove Proposition 3.1 for the case where π is the uniform distribution on Ω. Nevertheless, when Proposition 3.1 holds for any one strictly positive π, it holds for any other set of weights, π, under which P_π and Q_π dominate E and F. In particular, when E and F are dominated by P_{θ_1} and Q_{θ_1}, one can replace P_π and Q_π by P_{θ_1} and Q_{θ_1}, and $T_\pi(X)$ and $S_\pi(Y)$ by $T_{\theta_1}(X)$ and $S_{\theta_1}(Y)$ in Proposition 3.1, and one can re-phrase that result by stating that E is "sufficient for" F if and only if $E_{p_{\theta_1}}[\chi(T_{\theta_1}(x))] \geq E_{q_{\theta_1}}[\chi(S_{\theta_1}(y))]$ for every convex function $\chi(\cdot)$ on R_+^{k-1}.

Since $T_\pi(X)$ and $S_\pi(Y)$ are likelihood functions, Proposition 3.1 compares the informativity of E and F through the variability of the distribution of the likelihood functions they yield. On the other hand, given that any convex function $\phi(u)$ on K_π can be posed as $\phi(u) = \varphi(\pi_1 u_1, \ldots, \pi_k u_k)$ where $\varphi(h_1, \ldots, h_k)$ is a convex function on the simplex of R^k, Proposition 3.1 can also be posed in Bayesian terms as follows.

Proposition 3.2. *Experiment E is "sufficient for" F if and only if for a given strictly positive prior distribution π on Ω,*

$$E_{p_\pi}[\varphi(\pi_E(\theta|x))] \geq E_{q_\pi}[\varphi(\pi_F(\theta|y))] \tag{11}$$

for every convex function $\varphi(\cdot)$ on the simplex of R^k, where $\pi_E(\theta|x)$ and $\pi_F(\theta|y)$ are the posterior distributions under the same prior π.

This proposition can be re-phrased as "E is sufficient for F if and only if for some strictly positive prior π,

$$(\pi_E(\theta_1|X), \ldots, \pi_E(\theta_k|X))|_{p_\pi} \geq_{cx} (\pi_F(\theta_1|Y), \ldots, \pi_F(\theta_k|Y))|_{q_\pi}." \tag{12}$$

Given that $E[\varphi(\cdot)]$ measures the variability of the distribution of distributions on the simplex of R^k, stating that "E is sufficient for F" is equivalent to stating that "the distribution of $(\pi_E(\theta_1|X), \ldots, \pi_E(\theta_k|X))$ under P_π is always more variable than the distribution of $(\pi_F(\theta_1|Y), \ldots, \pi_F(\theta_k|Y))$ under Q_π, no matter how variability is measured." Also, "E is sufficient for F" implies that $\pi_E(\theta_i|X)$ is always more variable around $E_{p_\pi}[\pi_E(\theta_i|x)] = \pi_i$ than $\pi_F(\theta_i|Y)$ is around $E_{q_\pi}[\pi_F(\theta_i|y)] = \pi_i$, coordinate-wise.

I formulate the version of Blackwell-Sherman-Stein theorem in Proposition 3.1 before the one in Proposition 3.2 for historical reasons, and because it frames the sufficiency ordering of experiments in terms of the convex ordering of their likelihood functions, which might make that result more appealing than by framing it in terms of the convex ordering of their posterior distributions.

Remark: When E is an experiment on a parameter space with an infinite number of elements, the infinite-dimensional vector of likelihood ratios, $T_\pi(X)$, and $\pi_E(\theta|X)$, are stochastic processes indexed by $\theta \in \Omega$ and with the corresponding distributions induced by $X \sim P_\pi(x)$. For the statement and proof of a result analogous to Proposition 3.1 that applies for experiments on countable or uncountable parameter spaces see Le Cam (1986, pp. 43-44). In the infinite Ω case, it is established that an experiment E is "sufficient for" or "almost at least as informative as" F if and only if Proposition 3.1 holds for the restrictions of experiments E and F to every finite subset of Ω. Given that the convex ordering of a stochastic process is implied by the corresponding finite dimensional convex orderings (see, e.g., Bassan and Scarsini 1991), this allows one to generalize Proposition 3.1-3.2 both to countable and continuous parameter spaces.

4 What is a valid measure of information in E?

Given two experiments, often neither of them is sufficient for the other and therefore, the sufficiency ordering is partial and does not serve the purpose of ranking experiments and determining the most informative one. The following definition identifies the minimal set of requirements for a function on a set of experiments on Ω to qualify as a valid measure of the information in them, inducing a total ordering in that set.

Definition 4.1. *A measure of the information about θ in an experiment E assigns a value $I(E)$ such that*

1. *$I(E)$ is a real number,*
2. *$I(E_{tni}) = 0$, and*
3. *whenever E_1 and E_2 are such that E_1 is "sufficient for" E_2, then $I(E_1) \geq I(E_2)$.*

Under information measure $I(\cdot)$, and with experimental costs being the same, one would prefer E to F whenever $I(E) > I(F)$. The requirement that $I(\cdot)$ be a real number guarantees that the ordering induced by $I(\cdot)$ is total and thus it singles out which experiment to choose. That excludes matrices as well as real valued functions of θ or π, but it allows for the evaluation of real functions of θ or π at $\theta = \theta_0$ or $\pi = \pi_0$.

The second requirement setting $I(E_{tni})$ equal to 0, captures the fact that one should never pay not to see the outcome of an experiment. Nevertheless, when comparing any two experiments the value taken by $I(E_{tni})$ is irrelevant and therefore optimality design criteria can differ from an information measure by a constant term. As a consequence, the definition of *optimality design criteria* would consist only of the first and third requirements in Definition 4.1.

The third requirement in the above definition is needed because when E is at least as good as F for every terminal decision problem, then E has to be preferred to F under every information measure. This requirement can also be justified through the randomization argument in Definition 3.1, because when one can reproduce data like the one from experiment F starting from the data from E plus an auxiliary randomization,

without knowing θ, then E has to be preferred to F under every information measure. Given that this randomization argument uses only the statistical model, $(P_\theta; \theta \in \Omega)$, one does not need to assume that θ is a random variable nor that the experiments will be used in a decision problem to justify Definition 4.1.

Lindley (1956), Kiefer (1959) and De Groot (1962) argue that the measures of the statistical information they propose do respect the sufficiency ordering, thus implicitly recognizing the role played by that ordering in the characterization of the measure of the information in an experiment. Definition 4.1 goes one step further by making the sufficiency ordering into the only essential requirement in that characterization. For more specific references to the necessity of the third requirement in Definition 4.1, see Vajda (1989, chap. 6) and Torgersen (1991a, sec.7.2, 1994, p. 314-318).

The following properties are straightforward consequences of Definition 4.1.

Property 1: The information in E is never smaller than the information in totally non-informative experiments, and is never larger than the information in totally informative ones, $0 < I(E) < I(E_{ti})$.

Property 2: The information in experiment E observing X, and the information in sub-experiment E_T observing a statistic $T(X)$, satisfy $I(E) \geq I(E_T)$ with equality if the statistic $T(X)$ is sufficient for X.

Property 3: If $I(\cdot)$ is an information measure and $r(\cdot)$ is a real increasing function on the range of $I(\cdot)$ with $r(0) = 0$, then $r(I(\cdot))$ is an information measure.

Property 4: Given any collection of measures of the information in an experiment, $\{I^\eta(\cdot), \eta \in \Gamma\}$, their linear combinations with non-negative coefficients, their supremum and their infimum values, are all valid measures of information. By letting $\eta = \theta$ and $\Gamma = \Omega$, or $\eta = \pi$ and Γ be the set of probability measures on Ω, this allows one to construct measures of information out of real valued functions of θ or π the way it is illustrated in Examples 1, 4 and 8 below.

Property 5: Given any sequence of measures of the information in an experiment, $I_1(\cdot), I_2(\cdot), \ldots$, its limit, if it exists, is a valid information measure.

Example 1: In Stone (1961) and Goel and DeGroot (1979) it is proven that when E is "sufficient for" F, the difference of the Fisher information matrices of E and F, $I^\theta_{Fi}(E) - I^\theta_{Fi}(F)$, is non-negative definite. Consequently, if $\gamma(\cdot)$ is a real valued function of non-negative definite symmetric matrices such that $\gamma(\{0\}) = 0$ and that $\gamma(M_1) \geq \gamma(M_2)$ whenever $M_1 - M_2$ is non-negative definite, then

$$I^\theta_{Fi,\gamma}(E) = \gamma(I^\theta_{Fi}(E)) \tag{13}$$

satisfies the second and third requirements in Definition 4.1. Examples of (13) include the determinant of $I^\theta_{Fi}(E)$, $I^\theta_D(E) = |I^\theta_{Fi}(E)|$, and the A, E and all the other optimality design criteria introduced by Kiefer (1959) and extensively used by the DoE literature. Other than for location experiments though, (13) is still a function of θ and thus fails the

first requirement in Definition 4.1. Nevertheless, one can convert (13) into information measures by resorting to Property 4 and either

1. using an average of (13) with respect to a given distribution π_0 on Ω, $I^{\pi_0}_{Fi,\gamma}(E) = E_{\pi_0}[\gamma(I^{\theta}_{Fi}(E))]$, which in particular includes guessing θ to be equal to θ_0, or
2. using $I^{s}_{Fi,\gamma}(E) = \max_{\theta\in\Omega}\gamma(I^{\theta}_{Fi}(E))$, or $I^{i}_{Fi,\gamma}(E) = \min_{\theta\in\Omega}\gamma(I^{\theta}_{Fi}(E))$.

Clearly, all information measures must abide by Definition 4.1, because any measure of information that preferred experiment F to E even though one could obtain data like the one from F from the data from E plus an auxiliary randomization, would make for a silly information measure. Some may want to argue though that the list of requirements in Definition 4.1 is incomplete.

- Some might expect the measure of the information in an experiment to be additive in the sense that if $E_i = (X_i; P^i_\theta)$ for $i = 1, \ldots, N$ is a set of experiments on the same Ω, and if $\Pi^N_{i=1}E_i$ is an experiment observing all X_i independently given θ, then $I(\Pi^N_{i=1}E_i) = \sum^N_{i=1} I(E_i)$. But by requiring that, one would discard basically all measures of the information in E except $I^{\theta_0}_{Fi}(E)$ with $\theta_0 \in R$, (19) and (50).

  For example, under a linear normal experiment E_i yielding $X_i \in R^n$ distributed $N_n(A_i\theta, \sigma^2 I)$ with $\theta \in R^p$ and $n < p$, θ is not estimable and $I_D(E_i) = |A_i^{'}A_i| = 0$ and yet, performing several such experiments independently leading to a total sample size larger than p could make θ estimable and $I_D(\Pi_i E_i) = |I_{Fi}(\Pi_i E_i)| > 0$, in which case $I_D(\Pi_i E_i) > \sum_i I_D(E_i)$. Additivity would also exclude the measures in Sections 5.3 and 5.4, under which $I(\Pi_i E_i) \leq \sum_i I(E_i)$.

  Also, given that an infinite independent replication of any experiment E with $P_{\theta_i} \neq P_{\theta_j}$ if $\theta_i \neq \theta_j$, is equivalent to a totally informative experiment, E_{ti}, the information in such an infinite replication would be equal to $I(E_{ti})$. Under the additivity requirement, the information in E_{ti} would then have to either be infinite or 0 if $I(E) = 0$, and one would have to rule out all the measures of the information with $0 < I(E_{ti}) < \infty$. In fact, note that carrying out a totally informative experiment twice independently, provides the same information as carrying it out once, which also violates this type of additivity requirement.

- Given experiments $E_0 = (X_0; P^0_\theta)$ and $E_1 = (X_1; P^1_\theta)$ on Ω and given $p \in (0, 1)$, the mixture experiment $(1-p)E_0 + pE_1$ is the one obtained by first observing a Bernoulli random variable I on $\{0, 1\}$ with probability of success p, and then performing experiment E_I. Some might expect the measure of the information in an experiment to be linear under mixture experiments, $I((1-p)E_0 + pE_1) = (1-p)I(E_0) + pI(E_1)$. Indeed, all the measures covered by Definition 5.1 satisfy this condition, but imposing it on all measures of the information in E by listing it in Definition 4.1, would rule out measures based on the negative of the minimax risk for a given loss, measures covered by Example 1, (38) and (51).

- Fisher information matrices and Bayes risk typically depend on which parametrization one chooses. Adding the requirement that the information about θ in an experiment E always be the same as the information about any one-to-one function

$\psi = \psi(\theta)$ would also exclude most of the design optimality criteria considered in Example 1 and in Section 5.3. This lack of invariance has long been known and accepted by the DoE literature, and even though in specific contexts being invariant under a re-parametrization is a plus, that loses its appeal when parameters have a definite physical meaning (see, e.g., Pukelsheim, 1993, p. 137).

It is important to emphasize that Definition 4.1 lists the minimal set of requirements for functions on a set of experiments to qualify as measures of the information in them (much like the Kolmogorov axiom set lists the minimal set of requirements for functions on a field of events to qualify as measures of their probability). That does not mean that in specific settings like for example the ones considered in information theory or in Bernardo (2005a, 2005b), one is not entitled to impose additivity, linearity, invariance or any other additional requirements and in that way reduce the set of measures of the information in an experiment under consideration.

5 Generalized divergence measures of the information in an experiment E

As a consequence of Proposition 3.1, all measures of the variability of $T_\pi(X)$, expressed as $E_{p_\pi}[\phi(T_\pi(x))]$ for a convex function $\phi(\cdot)$, satisfy the first and third requirement of Definition 4.1 and therefore they qualify as design optimality criteria. As a consequence of Proposition 3.2, the same can be said of all measures of the variability of $\pi_E(\theta|X)$, expressed as $E_{p_\pi}[\varphi(\pi_E(\theta|x))]$ for a convex function $\varphi(\cdot)$.

Definition 5.1-5.2 below, encompassing the definitions of information in E of Lindley (1956), Raiffa and Schlaifer (1961), DeGroot (1962) and Csiszár (1963, 1967) as special cases, trivially further restrict $\phi(\cdot)$ and $\varphi(\cdot)$ so that $E_{p_\pi}[\phi(T_\pi(x))]$ and $E_{p_\pi}[\varphi(\pi_E(\theta|x))]$ also satisfy the second requirement in Definition 4.1 and qualify as measures of the information in E. In Section 5.6, $T_\pi(x)$ and $\pi_E(\theta|x)$ are interpreted as the reward from choosing and performing experiment E, $\phi(\cdot)$ and $\varphi(\cdot)$ become utility functions, and $E_{p_\pi}[\phi(T_\pi(x))]$ and $E_{p_\pi}[\varphi(\pi_E(\theta|x))]$ become the expected utility from choosing E.

5.1 Definition of generalized divergence measures

Definition 5.1. *The generalized ϕ-divergence measure of the information about θ in an experiment $E = (X; P_\theta)$ is*

$$I_\phi(E) = E_{p_\pi}[\phi(T_\pi(x))] = \int_{S_X} \phi(\frac{p_{\theta_1}(x)}{p_\pi(x)}, \ldots, \frac{p_{\theta_k}(x)}{p_\pi(x)}) p_\pi(x) dx, \tag{14}$$

where $\phi(u)$ is a real valued convex function on K_π such that $\phi(1, \ldots, 1) = 0$.

Condition $\phi(1, \ldots, 1) = 0$ sets $I_\phi(E_{tni})$ equal to 0 (every convex function can be made to satisfy this condition by subtracting from it, its value at $(1, \ldots, 1)$). It is easy to check that any two such convex functions with a difference that is linear on K_π

lead to the same $I_\phi(\cdot)$, but this is all the arbitrariness there is in the choice of $\phi(\cdot)$ because Torgersen (1991a, p. 354) proves that if $E_{p_\pi}[\phi_1(T_\pi(x))] = E_{p_\pi}[\phi_2(T_\pi(x))]$ for all experiments on Ω, then $\phi_1(\cdot) - \phi_2(\cdot)$ has to be linear on K_π. (Note though that in the context of one subset of experiments on Ω, different $I_\phi(\cdot)$ might induce the same ordering).

Now, one can re-phrase Proposition 3.1 by stating that E is "sufficient for" or "always at least as informative as" F, if and only if for one non vanishing π it holds that $I_\phi(E) \geq I_\phi(F)$ for every convex $\phi(\cdot)$ on K_π. Hence, the claim that measuring the information in E is essentially the same as *measuring the variability of its likelihood ratio statistics*; the more variable $T_\pi(X)$ is around $E_{p_\pi}[T_\pi(x)] = (1, \ldots, 1)$ in the sense of $E_{p_\pi}[\phi(\cdot)]$, the more peaked the likelihood functions that tend to be yielded by E, and the more informative E is in the sense of $I_\phi(\cdot)$.

The Bayesian interpretation of Definition 5.1 follows from the fact that

$$I_\phi(E) = E_{p_\pi}[\phi(\frac{\pi_E(\theta_1|x)}{\pi_1}, \ldots, \frac{\pi_E(\theta_k|x)}{\pi_k})], \tag{15}$$

and therefore that measuring the information in E is essentially the same as *measuring the variability of its posterior to prior ratio statistics*. Given that any convex function $\phi(u)$ on K_π can be posed as $\phi(u) = \varphi(\pi_1 u_1, \ldots, \pi_k u_k)$ where $\varphi(h_1, \ldots, h_k)$ is a convex function on the simplex of R^k, one can re-phrase Definition 5.1 as follows.

Definition 5.2. *The generalized ϕ-divergence measure of the information about θ in an experiment $E = (X; P_\theta)$ is*

$$I_\phi(E) = E_{p_\pi}[\varphi_\pi(\pi_E(\theta|x))] = \int_{S_X} \varphi_\pi(\pi_E(\theta_1|x), \ldots, \pi_E(\theta_k|x)) p_\pi(x) dx, \tag{16}$$

where $\pi_E(\theta|x)$ is the posterior under prior π and where $\varphi_\pi(h) = \phi(h_1/\pi_1, \ldots, h_k/\pi_k)$ is a convex function on the simplex of R^k such that $\varphi_\pi(\pi_1, \ldots, \pi_k) = 0$.

Condition $\varphi_\pi(\pi) = 0$ sets $I_\phi(E_{tni})$ equal to 0, and any two such convex functions with a difference that is linear will lead to the same measure $I_\phi(\cdot)$. For every E,

$$0 \leq I_\phi(E) \leq I_\phi(E_{ti}) = \pi_1 \varphi_\pi(1, \ldots, 0) + \ldots + \pi_k \varphi_\pi(0, \ldots, 1). \tag{17}$$

One can now re-phrase Proposition 3.2 by stating that E is "sufficient for" F if and only if for one strictly positive π, it holds that $I_\phi(E) \geq I_\phi(F)$ for every convex $\phi(u) = \varphi(\pi_1 u_1, \ldots, \pi_k u_k)$. Hence, the claim that measuring the information in E is also the same as *measuring the variability of the distribution of the posterior distributions* that tend to be yielded by it; The more variable $\pi_E(\theta|X)$ is around $E_{p_\pi}[\pi_E(\theta|X)] = (\pi_1, \ldots, \pi_k)$ in the sense of $E_{p_\pi}[\varphi_\pi(\cdot)]$, the more peaked the posterior distributions that tend to be yielded by E, and the more informative E is in the sense of $I_\phi(\cdot)$.

Section 6.1 defines the measure of the information in an observation $X = x$ yielded by experiment E to be equal to $\phi(T_\pi(x))$ or to $\varphi_\pi(\pi_E(\theta|x))$ for non-negative convex functions $\phi(\cdot)$ and $\varphi_\pi(\cdot)$. That will allow one to interpret $I_\phi(E)$ as the average of the

information in all the observations that could have been yielded by E, and it will provide one motivation for Definition 5.1-5.2 that does not rely on Proposition 3.1-3.2 and on the axiom set in Definition 4.1.

Remark: For experiments on infinite Ω, likelihood ratio statistics become stochastic processes indexed by θ. From the remark in Section 3.2.2 it follows that one can extend generalized divergence measures to experiments on infinite Ω, either be it countable or uncountable, by basing them on any arbitrary finite subset of coordinates of their infinite-dimensional likelihood ratio process or posterior distribution. Alternatively, given that the convex ordering of a stochastic process is implied by the corresponding finite dimensional convex orderings (see, e.g., Bassan and Scarsini 1991), one is also entitled to use as a measure of the information in E either (14) with $\phi(\cdot)$ being a convex functional of the complete infinite-dimensional likelihood ratio process, or (16) with $\varphi(\cdot)$ being a convex functional of the posterior distribution, as it is exemplified in Section 8.

Definition 5.1-5.2 reduces the choice of a measure of the information in an experiment E on Ω, to the choice of a convex function $\phi(\cdot)$ on K_π, or to the choice of a convex function $\varphi_\pi(\cdot)$ on the simplex of R^k. Next, a wide array of generalized divergence measures is presented, each bringing a different perspective on what statistical information means. They include the most well known information measures as well as measures not yet recognized as such by the DoE literature.

5.2 Csiszár divergences as generalized divergence measures

An experiment on $\Omega = \{\theta_1, \theta_2\}$ consisting of an ordered pair $(P_{\theta_1}, P_{\theta_2})$ is called a dichotomy. For them, the more informative E is, the easier it is to distinguish θ_1 from θ_2 based on outcomes of E, the further $p_{\theta_1}(X)$ tends to be away from $p_{\theta_2}(X)$ when $X \sim P_{\theta_1}$, the further $p_{\theta_2}(X)/p_{\theta_1}(X)$ tends to be away from 1, and given that $E_{p_{\theta_1}}[p_{\theta_2}(x)/p_{\theta_1}(x)] = 1$, the more variable $p_{\theta_2}(X)/p_{\theta_1}(X)$ is when $X \sim P_{\theta_1}$. For a closely related argument see Good (1979, 1985).

For non-dichotomous experiments E dominated by P_{θ_i}, this argument applies to each pair (θ_i, θ_j), and it provides the intuitive appeal behind the generalized divergences measures obtained from $\phi(u_1, \ldots, u_k) = u_i g(u_j/u_i)$ with $g(\cdot)$ being a convex function on R_+ with $g(1) = 0$,

$$I_g^{\theta_i,\theta_j}(E) = E_{p_{\theta_i}}[g(\frac{p_{\theta_j}(x)}{p_{\theta_i}(x)})], \tag{18}$$

that can also be interpreted as measures of the variability of $p_{\theta_j}(X)/p_{\theta_i}(X)$ when $X \sim P_{\theta_i}$. These numbers, known as Csiszár divergences, were independently related to the information in E by Csiszár (1963, 1967) and by Ali and Silvey (1966), and their statistical meaning stems from their relation to the Bayes risk when testing θ_j against θ_i (see, e.g., Torgersen 1991b, 1994; Vajda 1989).

Other than for dichotomies, Csiszár divergences are functions of (θ_i, θ_j) and not real numbers as required of information measures, but they can be made into information measures by using properties 4 and 5 in Section 4. These properties can also be used

to symmetrize (18) through $I_g^{\theta_i,\theta_j}(E) + I_g^{\theta_j,\theta_i}(E)$ or through $\min\{I_g^{\theta_i,\theta_j}(E), I_g^{\theta_j,\theta_i}(E)\}$.

When $g(u) = u \log u$, (18) becomes

$$I_{KL}^{\theta_i,\theta_j}(E) = E_{p_{\theta_j}}[\log \frac{p_{\theta_j}(x)}{p_{\theta_i}(x)}], \tag{19}$$

which is the Kullback-Leibler divergence between P_{θ_j} and P_{θ_i} introduced by Good (1950, 1960) and Kullback (1959). The two symmetrized versions of (19) described above, are respectively recognized in the literature as the Jeffreys divergence and as the intrinsic discrepancy, advocated for in Bernardo (2005a) on the grounds of its additivity, of its parametrization invariance, and of its finiteness.

When $g(u) = |u^{1/r} - 1|^r$ with $r \geq 1$, (18) becomes

$$I_S^{\theta_i,\theta_j,r}(E) = \int_{S_X} |p_{\theta_j}(x)^{1/r} - p_{\theta_i}(x)^{1/r}|^r dx, \tag{20}$$

which relates the information in E with distances between P_{θ_i} and P_{θ_j}; for $r = 1$ it is their statistical or variational distance, and for $r = 2$ it is their Hellinger distance.

When $g(u) = \text{sign}(1-t)(1-u^t)$ with $t > 0$ and $t \neq 1$, (18) becomes

$$I_R^{\theta_i,\theta_j,t}(E) = \text{sign}(1-t)(1 - E_{p_{\theta_i}}[(\frac{p_{\theta_j}(x)}{p_{\theta_i}(x)})^t]), \tag{21}$$

that are measures studied by Rényi (1961); for $t = 2$ or $g(u) = (u-1)^2$, one obtains

$$I_R^{\theta_i,\theta_j,t=2}(E) = \int_{S_X} \frac{(p_{\theta_j}(x) - p_{\theta_i}(x))^2}{p_{\theta_i}(x)} dx = Var_{p_{\theta_i}}[\frac{p_{\theta_j}(x)}{p_{\theta_i}(x)}]. \tag{22}$$

When $g(u) = \max_i\{u_i, 1\} - 1$, one obtains

$$I_M^{\theta_i,\theta_j}(E) = \int_{S_X} \max\{p_{\theta_j}(x), p_{\theta_i}(x)\} dx - 1. \tag{23}$$

Finally note that Kullback (1959, pp. 26-28), Torgersen (1991a, pp. 52-56) and others, derive Fisher information matrices through limiting arguments involving (19), and therefore all the information measures in Example 1 can be included within this generalized divergence framework in that way.

5.3 Expected value of sample information as generalized divergences

When it comes to comparing experiments in terms of their performance under one given non-sequential statistical decision problem, the negative of the Bayes risk for the given prior and loss and the negative of the minimax risk for the given loss, both satisfy the first and third requirements in Definition 4.1 and thus qualify as optimality design criteria, but their value for totally non-informative experiments is not 0 and thus they fail to make it as measures of the information in an experiment.

In Bayesian decision theory, Raiffa and Schlaiffer (1961) and DeGroot (1962) appraise the statistical worth of experiment $E = (X; P_\theta)$ through the expected value of the sample information (EVSI) in E, which is defined as

$$I_V^{\pi,L}(E) = E_{p_\pi}[E_{\pi_E(\theta|x)}[L(\theta, d^\pi) - L(\theta, d^{\pi_E(\theta|x)})]], \tag{24}$$

where d^π denotes the Bayes decision with respect to the prior distribution (i.e., d^π minimizes $E_\pi[L(\theta, d)]$), and where $d^{\pi_E(\theta|x)}$ denotes the Bayes decision with respect to the posterior distribution. That is, $I_V^{\pi,L}(E)$ is the expected savings in loss when, *a posteriori*, the best decision based on the posterior distribution obtained through E is used, instead of the best decision based on the prior. Given that

$$I_V^{\pi,L}(E) = E_\pi[L(\theta, d^\pi)] - E_{p_\pi}[E_{\pi_E(\theta|x)}[L(\theta, d^{\pi_E(\theta|x)})]], \tag{25}$$

where the first term does not depend on E and the second term coincides with the Bayes risk for the problem, and given that $I_V^{\pi,L}(E_{tni}) = 0$, it follows that (24) satisfies all the requirements of Definition 4.1 and thus it qualifies as a measure of the information in E. The experiment maximizing $I_V^{\pi,L}(E)$ is the one minimizing the corresponding Bayes risk, and it holds that for every E,

$$0 < I_V^{\pi,L}(E) < I_V^{\pi,L}(E_{ti}) = E_\pi[L(\theta, d^\pi)]. \tag{26}$$

Example 2: In estimation problems, the decision space is Ω. For $\theta \in R$ and $L_1(\theta, d) = (d-\theta)^2$, the Bayes decision with respect to π is $d^\pi = E_\pi[\theta]$, and $E_\pi[L_1(\theta, d^\pi)] = Var_\pi[\theta]$. From (24) and the conditional variance identity it follows that

$$I_V^{\pi,L_1}(E) = Var_\pi[\theta] - E_{p_\pi}[Var_{\pi_E(\theta|x)}[\theta]] = Var_{p_\pi}[E_{\pi_E(\theta|x)}[\theta]], \tag{27}$$

or which is the same,

$$I_V^{\pi,L_1}(E) = E_{p_\pi}[E_{\pi_E(\theta|x)}[\theta]^2] - E_\pi[\theta]^2. \tag{28}$$

Thus, the more the posterior Bayes decision, $d^{\pi_E(\theta|x)} = E_{\pi_E(\theta|x)}[\theta]$, varies with x, the more informative is experiment E. In estimation problems with $\theta \in R^p$ and $L_1(\theta, d) = (d-\theta)'H(d-\theta)$, where H is a known $p \times p$ non-negative definite symmetric matrix, the Bayes decision under π is $d^\pi = E_\pi[\theta]$, $E_\pi[L_1(\theta, d^\pi)] = \text{trace}\{HVar_\pi[\theta]\}$ and

$$I_V^{\pi,L_1}(E) = \text{tr}\{H(Var_\pi[\theta] - E_{p_\pi}[Var_{\pi_E(\theta|x)}[\theta]])\} = \text{tr}\{HVar_{p_\pi}[E_{\pi_E(\theta|x)}[\theta]]\}. \tag{29}$$

Example 3: In classification problems, Ω has k elements, $\Omega = \{\theta_1, \ldots, \theta_k\}$, and X has to be classified as coming from one of the k distributions. When the loss is zero if the classification is correct (i.e., $L_2(\theta_i, d = \theta_i) = 0$), and it is one if it is incorrect (i.e., $L_2(\theta_i, d = \theta_j) = 1$ when $\theta_i \neq \theta_j$), the Bayes decision under π is the θ_i that maximizes $\pi(\theta)$ and $E_\pi[L_2(\theta, d^\pi)] = 1 - \max_i \pi_i$. It follows that the EVSI in E for this problem is the expected increase in the modal probability of the distribution of θ due to E,

$$I_V^{\pi,L_2}(E) = E_{p_\pi}[\frac{\max_i\{\pi_i l_x(\theta_i)\}}{E_\pi[l_x(\theta)]}] - \max_i \pi_i = E_{p_\pi}[\max_i \pi_E(\theta_i|x)] - \max_i \pi_i. \tag{30}$$

In general, when Ω has k elements and the decision space has J elements, every loss function can be represented through a $k \times J$ matrix $\{L_{ij}\}$ where L_{ij} is the non-negative loss incurred when θ is θ_i and one picks decision d_j. In that case, the EVSI in E is

$$I_V^{\pi,L}(E) = E_{p_\pi}[\max_{j=1,\ldots,J}\{-\sum_{i=1}^{k} L_{ij}\pi_i \frac{p_{\theta_i}(x)}{p_\pi(x)}\}] - \max_{j=1,\ldots,J}\{-\sum_{i=1}^{k} L_{ij}\pi_i\}, \tag{31}$$

which is a generalized divergence measure both under the non-negative convex function

$$\phi_{\pi,L}^{(1)}(u) = \max_{j=1,\ldots,J}\{-\sum_{i=1}^{k} L_{ij}\pi_i u_i\} + \sum_{i=1}^{k} L_{id^\pi}\pi_i u_i, \tag{32}$$

as well as under the possibly negative convex function

$$\phi_{\pi,L}^{(2)}(u) = \max_{j=1,\ldots,J}\{-\sum_{i=1}^{k} L_{ij}\pi_i u_i\} - \max_{j=1,\ldots,J}\{-\sum_{i=1}^{k} L_{ij}\pi_i\}. \tag{33}$$

For details on Bayes risk based measures of the information in an experiment, see Lindley (1961), Chaloner and Verdinelli (1995), Dawid (1998), or Dawid and Sebastiani (1999).

5.4 Mutual information and its extensions as generalized divergences

If one treats θ as a random variable, one can measure the degree of association or dependency between X and θ implicit in their joint distribution $f_\pi(x,\theta)$ through the real valued random variable,

$$T_\pi(X,\theta) = \frac{f_\pi(X,\theta)}{p_\pi(X)\pi(\theta)} = \frac{\pi_E(\theta|X)}{\pi(\theta)} = \frac{p_\theta(X)}{p_\pi(X)} = \frac{l_X(\theta)}{E_\pi(l_X(\theta))}. \tag{34}$$

The stronger the association between X and θ, the better X explains θ, and the more informative E should be about θ. In particular, when X and θ are independent, $T_\pi(X,\theta) = 1$ with probability one and $E = E_{tni}$. When X and θ are functionally dependent, $T_\pi(X,\theta)$ is 0 for all (X,θ) except for (X,θ_j) where it is $1/\pi_j$, and $E = E_{ti}$. For an arbitrary experiment, the stronger the association between X and θ, the further the value of $f_\pi(X,\theta)$ tends to be away from the value of $p_\pi(X)\pi(\theta)$, and the further $T_\pi(X,\theta)$ tends to be away from one. Since $E_{p_\pi\pi}[T_\pi(x,\theta)] = 1$ for all E, the stronger that association, the more variable $T_\pi(X,\theta)$ when (X,θ) is distributed as $p_\pi\pi$.

The connection between variability of $T_\pi(X,\theta)$ and association of X and θ is made explicit in Ali and Silvey (1965). Its connection to the information in E follows as a consequence of using Proposition 3.1 with $\phi_\pi(u) = \sum_{i=1}^{k}\pi_i g(u_i)$, where $g(\cdot)$ is convex on R_+; that leads to E being "sufficient for" F implying that $T_\pi(X,\theta)|_{p_\pi\pi} \geq_{cx} S_\pi(Y,\theta)|_{q_\pi\pi}$, which in turn means that E being "sufficient for" F implies that the distribution of $T_\pi(X,\theta)$ under $p_\pi\pi$ is always more variable than the distribution of $S_\pi(Y,\theta) = q_\theta(Y)/q_\pi(Y)$ under $q_\pi\pi$, no matter how variability is measured.

When $\phi_\pi(u) = \sum_{i=1}^k \pi_i g(u_i)$ with the function $g(\cdot)$ being convex on R_+ and with $g(1) = 0$, (14) becomes

$$I_g^\pi(E) = E_{p_\pi \pi}[g(T_\pi(x, \theta))], \tag{35}$$

which serves as a measure of the variability of $T_\pi(X, \theta)$ when $(X, \theta) \sim p_\pi \pi$, as a measure of the association between X and θ, and as a measure of the information in E.

Example 4: A special case of (35) obtained with $\phi_\pi(u) = \sum_{i=1}^k \pi_i u_i \log u_i$ is

$$I_{MI}^\pi(E) = E_{f_\pi}[\log \frac{f_\pi(x, \theta)}{p_\pi(x)\pi(\theta)}] = E_{p_\pi}[-\log p_\pi(x)] - E_\pi[E_{p_\theta}[-\log p_\theta(x)]], \tag{36}$$

which in information theory is recognized as the mutual information or cross entropy between X and θ (see, Cover and Thomas 1991; Barron 1999). In purely Bayesian terms

$$I_{MI}^\pi(E) = E_\pi[-log\pi] - E_{p_\pi}[E_{\pi_E(\theta|x)}[-\log \pi_E(\theta|x)]], \tag{37}$$

which is the expected reduction of the entropy of the distribution of θ. Following the lead by Lindley (1956, 1972), Good (1960) and Rényi (1967a,b), (36) has become the default design optimality criteria in Bayesian DoE. Taking advantage of the analogy between (25) and (37), DeGroot (1962, 1979) and Bernardo (1979a) pose $I_{MI}^\pi(E)$ as an example of EVSI. It is also the Bayes risk when estimating the true density, p_θ, under the KL-divergence loss (see, e.g., Haussler and Opper 1997).

Considering the set $\{I_{MI}^\pi(\cdot), \pi \in \Gamma\}$, where Γ is the set of probability measures on Ω, and using Property 4 in Section 4 leads to

$$I_C(E) = \sup_{\pi \in \Gamma} I_{MI}^\pi(E), \tag{38}$$

which is a measure of the information in E that corresponds to the concept of capacity (Cover and Thomas 1991), and which is related to the minimax risk for estimating the true density, p_θ, under the KL-divergence loss (see, e.g., Haussler and Opper 1997). The π that maximizes (38) is closely related to the reference prior for E constructed in Bernardo (1979b) and thus $I_C(E)$ can be interpreted as the mutual information in E under that reference prior, which itself is a distribution that depends on E.

As an alternative to (36) one has proposed measuring information in E through the negative of an average of the entropy of P_θ, $I = -E_\pi[E_{p_\theta}[-\log p_\theta(x)]]$, and through $G = E_\pi[-log\pi] + I$ (see, e.g., Soofi 2000, p. 1351). Unfortunately, I and G can both take different values on E and on experiment E_T observing a sufficient statistic for E and therefore, according to I and G one could prefer E to E_T or E_T to E. The third requirement in Definition 4.1 rules that there should be a tie between E and E_T, and therefore I and G are not valid measures of the information in E.

Example 5: When $\phi_{\pi,r}(u) = \sum_{i=1}^k \pi_i |u_i^{1/r} - 1|^r$ with $r \geq 1$, (35) becomes

$$I_S^{\pi,r}(E) = \int_{S_X} \int_\Omega |(f_\pi(x, \theta))^{1/r} - (p_\pi(x)\pi(\theta))^{1/r}|^r d\theta dx \tag{39}$$

$$= E_{p_\pi}[\frac{E_\pi[|l_x(\theta)^{1/r} - E_\pi[l_x(\theta)]^{1/r}|^r]}{E_\pi[l_x(\theta)]}] \tag{40}$$

which relates the information in E to the distance between the joint distribution of (X, θ) and the product of their marginals. For $r = 1$, it coincides with twice the measure of the association between X and θ proposed in Silvey (1964). An alternative way to pose (40) in terms of average distance between posterior and prior is

$$I_S^{\pi,r}(E) = E_{p_\pi}[\sum_{i=1}^{k} |\pi_E(\theta_i|x)^{1/r} - \pi_i^{1/r}|^r], \tag{41}$$

which is as in (16) with $\varphi_{\pi,r}(h) = \sum_{i=1}^{k} |h_i^{1/r} - \pi_i^{1/r}|^r$.

Example 6: When $\phi_{\pi,t}(u) = \text{sign}(1-t)\sum_{i=1}^{k} \pi_i(1-u_i^t)$ with $t > 0$ and $t \neq 1$,

$$I_R^{\pi,t}(E) = \text{sign}(1-t)(1 - E_{p_\pi \pi}[(\frac{f_\pi(x,\theta)}{p_\pi(x)\pi(\theta)})^t]), \tag{42}$$

that are information measures considered in Goel and DeGroot (1981). For $t = 2$,

$$I_R^{\pi,t=2}(E) = Var_{p_\pi \pi}[T_\pi(X,\theta)] = E_{p_\pi}[\frac{Var_\pi[l_x(\theta)]}{E_\pi[l_x(\theta)]^2}], \tag{43}$$

which links the information in E with the variance of $T_\pi(X, \theta)$. The larger the coefficient of variation (or reciprocal of the signal to noise ratio) of the values of the likelihood function, $(l_x(\theta_1), \ldots, l_x(\theta_k))$ that tend to be yielded by E, the larger the information in E as measured through $I_R^{\pi,t=2}(E)$. Also,

$$I_R^{\pi,t=2}(E) = E_{p_\pi}[\sum_{i=1}^{k} \frac{(\pi_E(\theta_i|x) - \pi_i)^2}{\pi_i}], \tag{44}$$

which is as in Definition 5.2 with $\varphi_\pi(h) = \sum_i (h_i - \pi_i)^2/\pi_i$.

Example 7: When $\phi_\pi(u) = \sum_{i=1}^{k} \pi_i \max\{u_i, 1\} - 1$, (35) becomes

$$I_M^\pi(E) = \int_{S_X} \int_\Omega \max\{p_\pi(x)\pi(\theta), f_\pi(x,\theta)\} d\theta dx - 1 \tag{45}$$

$$= E_{p_\pi}[\frac{E_\pi[\max_i\{l_x(\theta_i), E_\pi[l_x(\theta)]\}]}{E_\pi[l_x(\theta)]}] - 1, \tag{46}$$

that can also be posed as:

$$I_M^\pi(E) = E_{p_\pi}[\sum_{i=1}^{k} \max\{\pi_E(\theta_i|x), \pi_i\}] - 1, \tag{47}$$

which is as in (16) with $\varphi_\pi(h) = \sum_{i=1}^{k} \max\{h_i, \pi_i\} - 1$.

5.5 Other examples of generalized divergence measures

Example 8: Consider $\phi_t(u) = 1 - u_1^{t_1} u_2^{t_2} \dots u_k^{t_k}$, where $t = (t_1, \dots, t_k)$ is a probability measure on Ω and thus such that $t_i \geq 0$ and that $\sum_{i=1}^k t_i = 1$, which is a non-negative convex function on the set K_t. The corresponding generalized ϕ-divergence measure is

$$I_{H_1}^t(E) = 1 - \int_{S_X} \Pi_{i=1}^k p_{\theta_i}(x)^{t_i} dx = 1 - E_{p_\pi}[\frac{\Pi_{i=1}^k l_x(\theta_i)^{t_i}}{E_\pi[l_x(\theta)]}], \tag{48}$$

that is a natural generalization of (21). It can be checked that $0 \leq I_{H_1}^t(E) \leq 1$, that $I_{H_1}^t(E) = 0$ if, and only if $E = E_{tni}$, and that $I_{H_1}^t(E) = 1$ if, and only if $E = E_{ti}$. The smaller the ratio between a weighted geometric and a weighted arithmetic mean of the likelihood values, $(l_x(\theta_1), \dots, l_x(\theta_k))$, that tend to be yielded by E, the larger $I_{H_1}^t(E)$. Its Bayesian interpretation follows from the fact that (48) can also be written as:

$$I_{H_1}^t(E) = 1 - E_{p_\pi}[\frac{\Pi_{i=1}^k \pi_E(\theta_i|x)^{t_i}}{\Pi_{i=1}^k \pi_i^{t_i}}], \tag{49}$$

which is as in Definition 5.2 with $\varphi_t(h) = 1 - \Pi_{i=1}^k (h_i/\pi_i)^{t_i}$. As it will be argued in detail in Section 7, this important family of information measures is not encompassed by the definition of measure of the information proposed in DeGroot (1962).

What makes these measures of information special is the fact that $I_{H_1}^t(E) = 1 - H_E(t)$, where $H_E(t)$ is a function on the distributions on Ω with finite support known as the Hellinger transform of E. This transform is the Laplace transform of the distribution of the vector of logarithms of the coordinates of $T_\pi(X)$ under P_π, and as such it characterizes the distribution of $T_\pi(X)$ and given that $T_\pi(X)$ is a sufficient statistic for E, it also characterizes the experiment E (see, e.g., Goel 1988; Torgersen 1991a; Le Cam and Yang 2000; Shiryaev and Spokoiny 2000). In pattern recognition, $H_E(t)$ are used as upper bounds of the Bayes risk for classification (see, Fazekas and Liese 1996).

Since $r(u) = -\log(1-u)$ is strictly increasing in $[0,1]$ and $r(0) = 0$,

$$I_{H_2}^t(E) = -\log(1 - I_{H_1}^t(E)) = -\log H_E(t) \tag{50}$$

is also an information measure, which is additive under independent experiments. By considering the set of information measures $\{I_{H_2}^t(\cdot), t \in \Gamma\}$, where Γ is the set of probability measures on Ω, and using Property 4 in Section 4, one can construct an information measure that does not depend on t through

$$I_{Ch}(E) = \sup_{t \in \Gamma} I_{H_2}^t(E) = -\inf_{t \in \Gamma} \log H_E(t). \tag{51}$$

When $k = 2$, $I_{Ch}(E)$ is the information number defined in Chernoff (1952) and used to obtain an upper bound for the logarithm of the Bayes risk for the problem of testing simple versus simple hypotheses (see, e.g., Cover and Thomas 1991, p. 312).

Example 9: Consider the convex function $\phi_\pi(u) = \sqrt[r]{\pi_1 u_1^r + \dots + \pi_k u_k^r} - 1$ where $r \geq 2$. The corresponding generalized ϕ-divergence measure is

$$I_{RM}^{\pi,r}(E) = \int_{S_X} \sqrt[r]{\pi_1 p_{\theta_1}(x)^r + \dots + \pi_k p_{\theta_k}(x)^r}\, dx - 1 = E_{p_\pi}[\frac{\sqrt[r]{E_\pi[l_x(\theta)^r]}}{E_\pi[l_x(\theta)]}] - 1. \tag{52}$$

When $r = 2$, the larger the values taken by the ratio between the weighted root mean square and arithmetic means of the $(l_x(\theta_1), \ldots, l_x(\theta_k))$ that tend to be yielded by E, the larger $I_{RM}^{\pi,r=2}(E)$. Furthermore, it is easy to check that

$$I_{RM}^{\pi,r}(E) = E_{p_\pi}[\sqrt[r]{E_\pi[(\frac{\pi_E(\theta|x)}{\pi(\theta)})^r]}] - 1 = E_{p_\pi}[\sqrt[r]{E_{\pi_E(\theta|x)}[(\frac{\pi_E(\theta|x)}{\pi(\theta)})^{r-1}]}] - 1, \quad (53)$$

which is as in (16) with $\varphi_\pi(h) = \sqrt[r]{\pi_1^{1-r}h_1^r + \ldots + \pi_k^{1-r}h_k^r} - 1$.

Example 10: The generalized ϕ-divergence obtained with $\phi(u) = \max_i u_i - 1$ is

$$I_{MR}(E) = \int_{S_X} \max_i p_{\theta_i}(x)\, dx - 1 = E_{p_\pi}[\frac{\max_i l_x(\theta_i)}{E_\pi[l_x(\theta)]}] - 1, \quad (54)$$

and it links information in E with the ratio between the maximum and the average values of the likelihood functions that tend to be attained through experiment E, and with the maximum of the corresponding posterior to prior ratio,

$$I_{MR}(E) = E_{p_\pi}[\max_i \frac{\pi_E(\theta_i|x)}{\pi_i}] - 1, \quad (55)$$

which is as in Definition 5.2 with $\varphi_\pi(h) = \max_i(h_i/\pi_i) - 1$.

Example 11: Given the link between information in E and variability of its likelihood ratio statistics $T_\pi(X)$, it is natural to expect that the variance-covariance matrix of this statistic, $Var_{p_\pi}[T_\pi(x)]$, will be related to information. If $A = \{a_{ij}\}$ is a known $k \times k$ symmetric non-negative definite matrix and if $\mathbf{1}' = (1, \ldots, 1)$, the generalized ϕ-divergence measure obtained with $\phi_A(u) = (u - \mathbf{1})'A(u - \mathbf{1})$ is

$$I_Q^A(E) = E_{p_\pi}[(T_\pi(X) - \mathbf{1})'A(T_\pi(X) - \mathbf{1})] = trace(AVar_{p_\pi}[T_\pi(X)]). \quad (56)$$

In particular, when $A = ll'$ for a known $l \in R^k$ one obtains

$$I_Q^l(E) = l'Var_{p_\pi}[T_\pi(x)]l. \quad (57)$$

Given that E "sufficient for" F implies that $I_Q^l(E) \geq I_Q^l(F)$ for every $l \in R^k$ and therefore it implies that $Var_{p_\pi}[T_\pi(X)] - Var_{q_\pi}[S_\pi(Y)]$ is a non-negative definite matrix, any real valued transformation, $\gamma(Var_{p_\pi}[T_\pi(x)])$, with $\gamma(\cdot)$ such that $\gamma(\{0\}) = 0$ and that $\gamma(A) \geq \gamma(B)$ whenever $A - B$ is non-negative definite, is an information measure. For example, this applies to the trace of $Var_{p_\pi}[T_\pi(x)]$.

5.6 The choice of an experiment as a decision theory problem

The generalized divergence measures in Definition 5.1-5.2 are linear under mixture experiments, i.e., they are all such that $I_\phi((1-p)E_0 + pE_1) = (1-p)I_\phi(E_0) + pI_\phi(E_1)$. In fact, from results in Torgersen (1991a, p. 353-355) it follows that generalized divergence

measures are the only measures of information abiding by Definition 4.1 and having this property. Hence, adding the condition that $I(E)$ be linear under mixtures to Definition 4.1 characterizes the information measures covered by Definition 5.1-5.2.

Furthermore, when comparing experiments the requirement setting $I(E_{tni}) = 0$ is irrelevant, and in that case, the conditions $\phi(1, \ldots, 1) = 0$ and $\varphi_\pi(\pi) = 0$ in Definition 5.1-5.2 can be disposed of. As a consequence, setting aside experimental cost, the choice of experiment based on statistical merit can be posed as a decision problem as follows.

Given a strictly positive set of weights, $\pi = (\pi_1, \ldots, \pi_k)$, one considers the reward from choosing E to be its likelihood ratio statistic $T_\pi(X)$, with a distribution on $\{T_\pi(x), x \in S_X\} \subset K_\pi$ induced by $X \sim P_\pi$ and denoted by $T_\pi(X)|_{p_\pi}$. By defining the utility function on these rewards to be any given convex function on K_π, $\phi(\cdot)$ (which here is neither assumed to be non-negative nor such that $\phi(1, \ldots, 1) = 0$), the choice of the most informative experiment is equivalent to the choice of the $T_\pi(X)|_{p_\pi}$ that maximizes the expected utility $E_{p_\pi}[\phi(T_\pi(x))]$.

Analogously, given a strictly positive prior distribution π, one can consider the reward from choosing E to be its posterior distribution statistic, $\pi_E(\theta|X)$, with a distribution on $\{\pi_E(\theta|x), x \in S_X\}$ induced by $X \sim P_\pi$ and denoted by $\pi_E(\theta|X)|_{p_\pi}$. By defining the utility function on these rewards to be any given convex function on the simplex of R^k, $\varphi(\cdot)$ (which here is neither assumed to be non-negative nor such that $\varphi(\pi) = 0$), the choice of the most informative experiment is equivalent to the choice of the $\pi_E(\theta|X)|_{p_\pi}$ that maximizes the expected utility $E_{p_\pi}[\varphi(\pi_E(\theta|x)]$.

To take experimental costs into consideration, one would have to include in the utility function an extra term that would typically depend on sample size and/or the specific outcome observed.

Note that the choice of an experiment is a decision problem, but it is not a *statistical* decision problem, because one is not entitled to carry out an experiment to help decide which experiment to chose. In this context, the measure of the information in an experiment E, $I_\phi(E)$, can be interpreted as the expected utility from carrying out that experiment. In the next section we interpret the utility of the reward actually obtained after having performed experiment E to be the information observed in $X = x$.

6 Measure of the observed information

So far we have considered the measure of the information about θ to be "expected" from an experiment, $E = (X; P_\theta)$, in advance of observing the data, or what is the same, the utility expected from choosing experiment E and carrying it out.

Here, we turn our attention to the measure of the information about θ "observed" in a given realization $X = x$ from E, or what is the same, to the utility of the reward that one actually obtains from E. That in turn will allow one to measure the information and the uncertainty about θ in any distribution h on Ω.

6.1 Measure of the information about θ in an observation $X = x$

Once the experiment E has been chosen and carried out, the issue arises as to how informative did the observed $X = x$ actually turn out to be. In statistical inference, experiment E is fixed and given, which explains why in that context the term information most often refers to the information in $X = x$ and not to the one in E; that is the case for example in Good (1950, 1966), Barnard (1951, 1959), Birnbaum (1962, 1969), Basu (1975), or Barndorff-Nielsen (1978). Even though the information in $X = x$ and the information in E are closely related, that connection is rarely made explicit.

The sufficiency principle rules that the information about θ in any observation $X = x$ from experiment E should depend only on the likelihood function determined by x, $l_x(\theta)$. In fact, given that $T_\pi(x)$ is a version of $l_x(\theta)$, that principle rules that the information in $X = x$ should depend only on the position of $T_\pi(x)$ in the set K_π, defined in Section 3.2.1 to be the convex hull of $\{(1/\pi_1, \ldots, 0), \ldots, (0, \ldots, 1/\pi_k)\}$, or what is the same, it should depend only on the position of $\pi_E(\theta|x)$ on the simplex of R^k.

The flatter $l_x(\theta)$, the closer $T_\pi(x)$ is to $(1, \ldots, 1)$, the closer $\pi_E(\theta|x)$ is to π, and even though neither the likelihood principle nor the sufficiency principle have anything to say about it, common wisdom dictates that the less informative $X = x$ should be considered to be. The more peaked $l_x(\theta)$ is at some $\theta_j \in \Omega$, the closer $T_\pi(x)$ is to the corresponding extreme point of K_π, the further the posterior distribution $\pi_E(\theta|x)$ is from the prior π towards the one-point distribution with $\pi_E(\theta_j|x) = 1$, and common wisdom dictates that the more informative $X = x$ should be considered to be.

In the limit, observations are said to be totally non-informative and are denoted by x_{tni}, if they lead to a constant likelihood, $l_x(\theta) = C$, and therefore if $T_\pi(x_{tni}) = (1, \ldots, 1)$, and if $\pi_E(\theta|x) = \pi$. Observations are said to be totally informative and are denoted by x_{ti}, if they lead to a likelihood that is zero everywhere except at one $\theta_j \in \Omega$, and therefore if $T_\pi(x_{ti})$ is an extreme point of K_π, and if one obtains a degenerate one-point posterior distribution. The experiments E_{tni} and E_{ti} defined in Section 2.1, yield observations in x_{tni} and in x_{ti} with probability one. For most experiments though, the sample space does not have any totally informative or totally non-informative points, and the class of points denoted by x_{tni} and x_{ti} are empty.

The sufficiency principle dictates that the information in $X = x$ has to be measured through functions of $T_\pi(x)$ and common wisdom dictates that these functions have to be such that the further $T_\pi(x)$ is away from $(1, \ldots, 1)$ towards an extreme point of K_π, the larger the values they take. The following definition, which encompasses the definition of measure of the information in $X = x$ given in DeGroot (1984) as a special case, naturally restricts attention to non-negative convex functions of $T_\pi(x)$.

Definition 6.1. *The generalized ϕ-divergence measure of the information about θ in a realization $X = x$ from experiment $E = (X; P_\theta)$ is*

$$I_\phi(x) = \phi\left(\frac{p_{\theta_1}(x)}{p_\pi(x)}, \ldots, \frac{p_{\theta_k}(x)}{p_\pi(x)}\right) = \phi\left(\frac{\pi_E(\theta_1|x)}{\pi_1}, \ldots, \frac{\pi_E(\theta_k|x)}{\pi_k}\right), \tag{58}$$

where $\phi(u)$ is a non-negative convex function on K_π with $\phi(1, \ldots, 1) = 0$.

Relative to Definition 5.1, the only new feature required of $\phi(u)$ in Definition 6.1 is that it be non-negative, which leads to $I_\phi(x)$ being minimized at x_{tni}, with $I_\phi(x_{tni}) = 0$. Because of the convexity of $\phi(u)$ and of K_π, $I_\phi(x) = \phi(T_\pi(x))$ is maximized at one of the extreme points of K_π, when $x = x_{ti}$.

In Bayesian terms, the larger the impact of $X = x$ on the beliefs about θ, the further the 'posterior to prior ratio statistic' $T_\pi(x)$ is away from $(1, \ldots, 1)$ towards an extreme point of K_π, and by the convexity of $\phi(\cdot)$, the larger $I_\phi(x) = \phi(T_\pi(x))$. Given that any convex $\phi(\cdot)$ on K_π can be posed as $\phi(u) = \varphi(\pi_1 u_1, \ldots, \pi_k u_k)$ where $\varphi(\cdot)$ is convex on the simplex of R^k (i.e., the space of probability measures over Ω), one can re-phrase Definition 6.1 in a way that captures that the further $\pi_E(\theta|x)$ is away from π towards an extreme point of the simplex, the more informative $X = x$ is.

Definition 6.2. *The generalized ϕ-divergence measure of the information about θ in a realization $X = x$ from experiment E is*

$$I_\phi(x) = \varphi_\pi(\pi_E(\theta_1|x), \ldots, \pi_E(\theta_k|x)), \tag{59}$$

where $\pi_E(\theta|x)$ is the posterior under the prior π, and where $\varphi_\pi(h) = \phi(h_1/\pi_1, \ldots, h_k/\pi_k)$ is a non-negative convex function on the simplex of R^k with $\varphi_\pi(\pi_1, \ldots, \pi_k) = 0$.

In Section 5.6, $I_\phi(x)$ plays the role of the utility of the reward obtained from E. By definition, it follows that the average of the information in all the observations that could have been yielded by E, $E_{p_\pi}[I_\phi(x)]$, is the information in E, $I_\phi(E)$, and that average can also be interpreted as the expected utility of E. Given $\phi(\cdot)$, the more informative the observations that tend to be yielded by E in the sense of $I_\phi(x)$, the more informative E is in the sense of $I_\phi(E)$, and the larger the expected utility of E.

This argument allows one to motivate Definition 5.1-5.2 and Section 5.6 without relying on Proposition 3.1-3.2 and on the axiomatic framework of Definition 4.1. Definition 6.1-6.2 would in fact have made for a very good alternative starting point for the manuscript. In that alternative presentation, instead of building the case for Definition 5.1-5.2 starting from first principles, one could have defined $I_\phi(E)$ to be just the average of the information in all the observations that could have been yielded by E.

Finding $I_\phi(x) = \phi(T_\pi(x))$ to be unpredictably larger than $I_\phi(E) = E_{p_\pi}[\phi(T_\pi(x))]$ by identifying it to be an outlier of the distribution of $I_\phi(X)$ under P_π, indicate that

1. $X = x$ was unusually informative about θ, or maybe that
2. the data were not coming from any distribution in (P_θ, Ω) and therefore the assumed statistical model (experiment) was wrong, or most likely that
3. in the words of DeGroot (1984, p. 290), "the prior distribution might have been misleading in that it was probably concentrated around an incorrect value of θ."

In fact, $I_\phi(x)$ can be interpreted as a *measure of the surprise about θ in $X = x$* which makes it useful as a Bayesian model checking test statistic to assess the compatibility between $X = x$ and $\pi(\theta)$, in the spirit of Box (1980) and of Bayarri and Berger (1999).

Requiring that $I_\phi(x) = \phi(T_\pi(x))$ and not just its expectation, $I_\phi(E) = E_{p_\pi}[\phi(T_\pi(x))]$, be non-negative amounts to treating the three types of feedback listed above as valuable. Disposing of the non-negativity requirement in Definition 6.1-6.2 would give the impression that after the experiment one can be worse off because his prior assumptions were misguided or by mere bad luck.

It is important to remark that even though one can define $I_\phi(E)$ and the expected utility of E through possibly negative convex functions on K_π, when Ω is finite one can always obtain one non-negative convex function on K_π that leads to that same $I_\phi(E)$ by adding to $\phi(\cdot)$ an appropriate linear function. In that way, one can always associate one (non-negative) ϕ-divergence measure of the information in $X = x$, $I_\phi(x)$, to every ϕ-divergence measure of the information in E, $I_\phi(E)$.

In the context of the EVSI measures considered in subsection 5.3,

$$I_V^{\pi,L}(x) = \phi_{\pi,L}^{(1)}(T_\pi(x)) = E_{\pi_{E}(\theta|x)}[L(\theta, d^\pi) - L(\theta, d^{\pi_E(\theta|x)})] \tag{60}$$

is the ϕ-divergence measure under the non-negative convex function in (32) with $I_V^{\pi,L}(E) = E_{p_\pi}[I_V^{\pi,L}(x)]$. It measures how much one saves after the experiment by *a posteriori* using the Bayes decision a posteriori, $d^{\pi_E(\theta|x)}$, instead of the Bayes decision a priori, d^π, and it is called the conditional value of sample information in Raiffa and Schlaiffer (1961). Under the loss function in Example 2 $I_V^{\pi,L_1}(x)$ is the square of the difference between posterior and prior expected values of θ. Under the loss in Example 3 $I_V^{\pi,L_2}(x)$ is the modal posterior probability minus the posterior probability of the prior mode.

Under the possibly negative convex function in (33)

$$\phi_{\pi,L}^{(2)}(T_\pi(x)) = E_\pi[L(\theta, d^\pi)] - E_{\pi_E(\theta|x)}[L(\theta, d^{\pi_E(\theta|x)})], \tag{61}$$

which is also such that $I_V^{\pi,L}(E) = E_{p_\pi}[\phi_{\pi,L}^{(2)}(T_\pi(x))]$ but which can be negative and therefore it does not qualify as a measure of the information in $X = x$. Instances where one observes a sample that leads to a posterior expected loss that is larger than the prior expected loss can be a highly informative warning.

Table 1 lists the measures $I_\phi^\pi(x)$ associated to Examples 2 to 10. It includes as special cases the KL-divergence between posterior and prior, and the coefficient of variation, one minus the ratio between weighted geometric and arithmetic averages, and the ratio between the maximum and the average of the values taken by the likelihood function.

Remark: The likelihood principle rules that, if the likelihood function $l_x(\theta)$ obtained from E is proportional to the likelihood function $l_y(\theta)$ obtained from F, the conclusions drawn from $X = x$ should be identical to the conclusions drawn from $Y = y$, and in that case, one should require that the information in $X = x$ has to be equal to the information in $Y = y$. Given that when $l_x(\theta)$ is proportional to $l_y(\theta)$ then $T_\pi(x) = S_\pi(y)$, as a consequence of the likelihood principle one should always use the same convex function $\phi(\cdot)$ to compare the information in observations from different experiments. Other than that though, nothing else in our interpretation of informativity in terms of peakedness

Ex.	$\phi_\pi(u)$	$I^\pi_\phi(x)$	$I^\pi_\phi(x)$
2	$(E_\pi[\theta u] - E_\pi[\theta])^2$	$(E_{\pi_E(\theta\|x)}[\theta] - E_\pi[\theta])^2$	$(\frac{E_\pi[\theta l_x(\theta)]}{E_\pi[l_x(\theta)]} - E_\pi[\theta])^2$
3	$\max_i\{\pi_i u_i\} - \pi(\theta^\pi_m)u(\theta^\pi_m)$	$\max_\theta\{\pi_E(\theta\|x)\} - \pi_E(\theta^\pi_m\|x)$	$\frac{\max_\theta\{\pi(\theta)l_x(\theta)\}}{E_\pi[l_x(\theta)]} - \frac{\pi(\theta^\pi_m)l_x(\theta^\pi_m)}{E_\pi[l_x(\theta)]}$
4	$\sum_{i=1}^k \pi_i u_i \log u_i$	$E_{\pi_E(\theta\|x)}[\log \frac{\pi_E(\theta\|x)}{\pi(\theta)}]$	$E_\pi[\frac{l_x(\theta)}{E_\pi[l_x(\theta)]} \log \frac{l_x(\theta)}{E_\pi[l_x(\theta)]}]$
5	$\sum_{i=1}^k \pi_i \|u_i^{1/r} - 1\|^r$, $1 \le r$	$E_\pi[\|(\frac{\pi_E(\theta\|x)}{\pi(\theta)})^{1/r} - 1\|^r]$	$E_\pi[\|(\frac{l_x(\theta)}{E_\pi[l_x(\theta)]})^{1/r} - 1\|^r]$
6	$\sum_{i=1}^k \pi_i(1-u_i^t)$, $0<t<1$	$1 - E_\pi[(\frac{\pi_E(\theta\|x)}{\pi(\theta)})^t]$	$1 - \frac{E_\pi[l_x(\theta)^t]}{E_\pi[l_x(\theta)]^t}$
6	$\sum_{i=1}^k \pi_i(u_i^t - 1)$, $1<t$	$E_\pi[(\frac{\pi_E(\theta\|x)}{\pi(\theta)})^t] - 1$	$\frac{E_\pi[l_x(\theta)^t]}{E_\pi[l_x(\theta)]^t} - 1$
6	$\sum_{i=1}^k \pi_i(u_i^2 - 1)$	$E_\pi[(\frac{\pi_E(\theta\|x)}{\pi(\theta)})^2] - 1$	$\frac{Var_\pi[l_x(\theta)]}{E_\pi[l_x(\theta)]^2}$
7	$\sum_{i=1}^k \pi_i \max\{u_i, 1\} - 1$	$E_\pi[\max\{\frac{\pi_E(\theta\|x)}{\pi(\theta)}, 1\}] - 1$	$E_\pi[\max\{\frac{l_x(\theta)}{E_\pi[l_x(\theta)]}, 1\}] - 1$
8	$1 - u_1^{\pi_1} \ldots u_k^{\pi_k}$	$1 - \Pi_{i=1}^k (\frac{\pi_E(\theta_i\|x)}{\pi_i})^{\pi_i}$	$1 - \frac{\Pi_{i=1}^k l_x(\theta_i)^{\pi_i}}{E_\pi[l_x(\theta)]}$
9	$\sqrt[r]{\pi_1 u_1^r + \ldots + \pi_k u_k^r} - 1$	$\sqrt[r]{E_\pi[(\frac{\pi_E(\theta\|x)}{\pi(\theta)})^r]} - 1$	$\frac{\sqrt[r]{E_\pi[l_x(\theta)^r]}}{E_\pi[l_x(\theta)]} - 1$
10	$\max_i u_i - 1$	$\max_\theta\{\frac{\pi_E(\theta\|x)}{\pi(\theta)}\} - 1$	$\frac{\max_\theta\{l_x(\theta)\}}{E_\pi[l_x(\theta)]} - 1$

Table 1: Measures of the *information in* $X = x$ associated to the measures of the *information in* E of Examples 2 to 10. Each measure is presented both in terms of distance between the posterior and prior densities as well as in terms of a convex function of the likelihood. In Example 3, θ^π_m denotes the mode of the prior distribution.

of the likelihood and of variability of likelihood ratios follows from the likelihood or the weaker sufficiency principle. The rationale for Definitions 5.1-5.2 and 6.1-6.2 is grounded on Proposition 3.1-3.2 and on the axiom set in Definition 4.1.

6.2 Measure of the information about θ in a distribution h on Ω

Here, the information about a random variable θ in a distribution h on Ω is measured through the information in an observation that updates a baseline reference distribution treated as a prior, into a posterior h. The uncertainty about θ in h is then measured as the information in a one-point distribution, h_{ct}, minus the information in h.

Definition 6.3. *The information about θ in the distribution $h = (h_1, \ldots, h_k)$ on Ω is the information in any observation $X = x$ that updates the uniform prior distribution, $h_{1/k} = (1/k, \ldots, 1/k)$, into a posterior distribution h, $\pi_E(\theta|x) = h$. Within the context of generalized divergence measures this leads to measuring information in h through*

$$I(h) = \phi_{1/k}(kh_1, \ldots, kh_k), \tag{62}$$

where $\phi_{1/k}(u)$ are non-negative convex functions on $K_{h_{1/k}}$ with $I(h_{1/k}) = \phi_{1/k}(1, \ldots, 1) = 0$ and with $I(h_{ct}) = \phi_{1/k}(k, \ldots, 0) = \ldots = \phi_{1/k}(0, \ldots, k)$. This is the same as

$$I(h) = \varphi_{1/k}(h_1, \ldots, h_k), \tag{63}$$

where $\varphi_{1/k}(h_1, \ldots, h_k)$ are non-negative convex functions on the simplex of R^k with $I(h_{1/k}) = \varphi_{1/k}(h_{1/k}) = 0$ and with $I(h_{ct}) = \varphi_{1/k}(1, \ldots, 0) = \ldots = \varphi_{1/k}(0, \ldots, 1)$.

By requiring that $\varphi_{1/k}(h_{1/k}) = 0$, this definition arbitrarily considers $I(h)$ to be smallest for the uniform distribution $h_{1/k}$, with $I(h_{1/k}) = 0$, but it could be made smallest for any other baseline distribution h_{ref} on Ω by replacing $h_{1/k}$ by h_{ref} in Definition 6.3.

The replacement of $h_{1/k}$ by other baseline distributions is very much indicated in the infinite Ω case, when uniform distributions lose their unique role as baseline distributions and one might want to use an alternative reference prior instead, in a way analogous to the one used in Clarke (1996) to measure the information in a prior distribution in terms of equivalent sample size. Note though that the lack of a universal agreement on what counts as a minimally informative reference distribution for uncountable Ω, coupled with the fact that for them one-point distributions are not absolutely continuous, here complicates considerably the jump from finite to uncountable Ω.

Because of the convexity of $I(\cdot)$, $I(h)$ is largest at the extreme points of the simplex, that correspond to degenerate one-point distributions, h_{ct}, with $I(h_{ct}) = \varphi_{1/k}(1, \ldots, 0)$. Requiring that $\varphi_{1/k}(1, \ldots, 0) = \ldots = \varphi_{1/k}(0, \ldots, 1)$ forces the information about θ in all one-point distributions, $I(h_{ct})$, to be the same irrespective of the $\theta_j \in \Omega$ 'held true' by h_{ct}. (By assuming equal information for all one-point distributions, one is not assuming that the information in a totally informative observation has to be the same irrespective of the one-point posterior distribution that that observation is leading to).

Note that even though the $\varphi_\pi(h)$ given in examples 5 to 7 and 8 to 10 do not take the same value on all the extreme points of the simplex, when Ω is finite one can always find one non-negative convex function that satisfies $\varphi_\pi(1, \ldots, 0) = \ldots = \varphi_\pi(0, \ldots, 1)$ and generates any generalized ϕ-divergence measure $I_\phi(E)$, by adding to $\varphi_\pi(h)$ an appropriate function that is linear on the simplex and vanishes at π. For example, the convex function $\varphi_\pi(h) = \sum_{i=1}^k \max\{h_i, \pi_i\} - 1$ that yields the measure in (47) fails this condition, but the non-negative convex function $\varphi_\pi(h) = \sum_{i=1}^k \pi_i(h_i + \max\{h_i/\pi_i, 1\}) - \sum_{i=1}^k \pi_i(\pi_i + 1)$ satisfies this condition and yields the same measure in (47).

On the other hand, note that once the baseline h_{ref} is agreed upon to be $h_{1/k}$ or any other minimally informative reference distribution, only the subset of non-negative convex functions on the simplex of R^k that are maximized on all h_{ct} and are 0 on h_{ref} qualify as measures of the information in h. In Examples 3 to 10 with $h_{ref} = h_{1/k}$, that restricts consideration to the measures obtained as $I(h) = \phi_{h_{1/k}}(kh) = \varphi_{h_{1/k}}(h)$.

In the context of Example 4, the information in E is interpreted as the expected reduction of the uncertainty about θ when the uncertainty is measured through the entropy of its distribution, and Rényi (1967a, 1967b) relates the uncertainty about θ in h to "the amount of missing information on θ when nothing else is known about θ except that its distribution is h." We next define the uncertainty about θ in h associated to generalized divergence measures by analogy.

Definition 6.4. *The uncertainty about θ in a distribution h on Ω is the information in a one-point distribution minus the information in h, $U(h) = I(h_{ct}) - I(h)$. Within*

the context of generalized divergences the uncertainty in h is thus measured through

$$U(h) = \phi_{1/k}(kh_{ct}) - \phi_{1/k}(kh), \tag{64}$$

or what is the same, through

$$U(h) = \varphi_{1/k}(h_{ct}) - \varphi_{1/k}(h), \tag{65}$$

where $\phi_{1/k}(u)$ and $\varphi_{1/k}(h)$ satisfy the conditions set in Definition 6.3.

Note that $U(h)$ is a non-negative concave function on the simplex of R^k (i.e., on the space of probability measures on Ω), which is minimized by all one-point distributions, h_{ct}, with $U(h_{ct}) = 0$, and which is maximized by the uniform distribution, $h_{1/k}$, with $U(h_{1/k}) = I(h_{ct}) = \varphi_{1/k}(1, \ldots, 0)$. This coincides with the definition of uncertainty function given in DeGroot (1962) and with the definition of measure of the diversity or heterogeneity of a population with probability measure h given in Rao (1982), which makes $I(h)$ a measure of the homogeneity of that population.

One could in fact, define uncertainty measures being maximized by any other reference distribution, h_{ref}, by replacing $h_{1/k}$ by h_{ref} in Definitions 6.3 and 6.4. It is important to emphasize though, that h_{ref} should be a distribution commonly agreed upon to represent maximum uncertainty (minimum information) about θ, irrespective of the prior π that one is going to eventually use to compute the posterior $\pi_E(\theta|x)$ in the context of an experiment. Once the baseline distribution, h_{ref}, is chosen, only the subset of non-negative concave functions on the simplex of R^k that are 0 on all h_{ct} and are maximized on h_{ref}, qualify as measures of the uncertainty in h. In the context of Examples 3 to 10 with $h_{ref} = h_{1/k}$, that restricts consideration to the uncertainty measures obtained as $U(h) = \phi_{h_{1/k}}(kh_{ct}) - \phi_{h_{1/k}}(kh) = \varphi_{h_{1/k}}(h_{ct}) - \varphi_{h_{1/k}}(h)$.

Table 2 lists the measures of the uncertainty about θ in h associated to Examples 2 to 10. This list includes the variance of θ, and the entropy, the Gini-Simpson index, the geometric average, one minus the maximum, and one minus the root mean square average of $h = (h_1, \ldots, h_k)$. Even though Table 2 presents the version of these measures for finite Ω, some of them can be extended by analogy to distributions on infinite Ω.

7 Information in E and uncertainty about θ

Here, the goal is to relate the information in E with the expected impact of E on the uncertainty in the distribution of θ, and to clarify the sense in which Definition 5.1-5.2 generalizes the definition of information in an experiment given in DeGroot (1962), which is the one typically adopted in Bayesian DoE and which encompasses the measures of information in Lindley (1956) and in Raiffa and Schlaiffer (1961), but which leaves out measures like the ones in Examples 5 to 10.

In DeGroot (1962), the information about θ in E is defined to be the uncertainty in the prior *minus* the expected uncertainty in the posterior, where as in Definition 6.4 uncertainty is measured through non-negative concave functions on the simplex of R^k,

Example	$\phi_\pi(u)$	$U(h)$
2	$(E_\pi[\theta u] - E_\pi[\theta])^2$	$\mathrm{Var}_h[\theta]$
3	$\max_i \{\pi_i u_i\} - \pi(\theta_m^\pi) u(\theta_m^\pi)$	$1 - \max_i h_i$
4	$\sum_{i=1}^k \pi_i u_i \log u_i$	$E_h[-\log h]$
5	$\sum_{i=1}^k \pi_i \lvert u_i - 1\rvert$	$2(1 - 1/k - \sum_{i=1}^k \lvert h_i - 1/k\rvert)$
6	$\sum_{i=1}^k \pi_i (1 - u_i^t)$, $0 < t < 1$	$k^{t-1}(E_h[h^{t-1}] - 1)$
6	$\sum_{i=1}^k \pi_i (u_i^t - 1)$, $1 < t$	$k^{t-1}(1 - E_h[h^{t-1}])$
6	$\sum_{i=1}^k \pi_i (u_i^2 - 1)$	$k(1 - E_h[h])$
7	$\sum_{i=1}^k \pi_i \max\{u_i, 1\} - 1$	$2 - 1/k - \sum_{i=1}^k \max\{h_i, 1/k\}$
8	$1 - u_1^{\pi_1} \dots u_k^{\pi_k}$	$k(h_1 \dots h_k)^{1/k}$
9	$\sqrt[r]{\pi_1 u_1^r + \dots + \pi_k u_k^r} - 1$	$\sqrt[r]{k}(1 - \sqrt[r]{E_h[h^{r-1}]})$
10	$\max_i u_i - 1$	$k(1 - \max_i h_i)$

Table 2: Measures of the *uncertainty about* θ *in its own distribution* $h(\theta)$ associated to the measures of the *information in* E of Examples 2 to 10. Except for Example 2, the baseline maximum uncertainty distribution is the uniform, $h_{1/k}$, and in these cases the measures of the *information about* θ *in* $h(\theta)$ can be obtained from $I(h) = U(h_{1/k}) - U(h)$.

$U(\cdot)$, taking the value 0 at all the degenerate one-point distributions, h_{ct}, and "typically attaining their maximum at or near the distribution $(1/k, \dots, 1/k)$,"

$$I_{DG}(E) = U(\pi) - E_{p_\pi}[U(\pi_E(\theta|x))] = E_{p_\pi}[I(\pi_E(\theta|x))] - I(\pi). \tag{66}$$

Therefore, the setting in De Groot (1962) covers all the measures of the information in E that can be interpreted as the *expected additive increase (decrease) of the information (uncertainty) about* θ *in its distribution*, when one updates the prior π based on the outcome of E. Even though De Groot does not make explicit how one would go about selecting the baseline maximum uncertainty distribution, it is implied that the uncertainty function, $U(\cdot)$, and thus the baseline distribution that maximizes $U(\cdot)$, is to be selected in a way that does not depend on the prior π.

Clearly, all the measures posed as $I_{DG}(E)$ are generalized divergence measures,

$$I_\phi(E) = E_{p_\pi}[\phi(\frac{\pi_E(\theta|x)}{\pi})] = E_{p_\pi}[\varphi_\pi(\pi_E(\theta|x))], \tag{67}$$

with $\varphi_\pi(h) = U(\pi) - U(h)$, because $\varphi_\pi(\cdot)$ is convex and $\varphi_\pi(\pi) = 0$. It is thus natural to ask wether or not all generalized divergence measures can be posed as in (66) and if not, to determine which $I_\phi(E)$ can be posed as in (66), and which can not.

By comparing (66) with (25), it is clear that the setting in De Groot (1962) includes all the measures considered in Section 5.3, with uncertainty being measured through $U_L(h) = E_h[L(\theta, d^h)]$, which is already recognized in Raiffa and Schlaiffer (1961) as the value of perfect information when $\pi = h$. By comparing (66) with (37), it is also clear that (66) includes the mutual information in Example 4, with $U_{MI}(h) = E_h[-\log h]$.

In general, to pose any generalized divergence measure, $I_\phi(E)$, as in (66), one would 'just' have to define $U(h)$ to be equal to $\varphi_\pi(h_{ct}) - \varphi_\pi(h)$, where $\varphi_\pi(h)$ is one of the convex functions defining that $I_\phi(E)$. The problem is that unless one can find one such $U(h)$ that is as in Definition 6.4 (i.e., it is non-negative, it is 0 for all h_{ct}, and it is maximized at a commonly agreed upon maximum uncertainty h_{ref}), $U(h)$ will not make it into a meaningful measure of the uncertainty in h, and one will not be able to interpret $I_\phi(E)$ as prior uncertainty minus expected posterior uncertainty.

In particular, it can be checked that the generalized divergence measures in (41), (44), (47), (49), (53), and in (55), can not be posed as in (66) with $U(\cdot)$ satisfying Definition 6.4, and therefore the measures of the information in an experiment covered by Examples 5 to 10 can not be interpreted as prior uncertainty minus expected posterior uncertainty as in De Groot (1962). For example, in order to pose the generalized divergence measure in Example 7,

$$I^\pi_M(E) = E_{p_\pi}[\sum_{i=1}^{k} \max\{\pi_E(\theta_i|x), \pi_i\}] - 1, \tag{68}$$

as in (66), the only non-negative concave function $U(h)$ with $U(h_{ct}) = 0$ that would allow it is

$$2 - \sum_{i=1}^{k} (\pi_i h_i + \max\{h_i, \pi_i\}), \tag{69}$$

and in order to pose the generalized divergence measure in Example 8,

$$I^\pi_{H_1}(E) = 1 - E_{p_\pi}[\frac{\Pi_{i=1}^k \pi_E(\theta_i|x)^{\pi_i}}{\Pi_{i=1}^k \pi_i^{\pi_i}}], \tag{70}$$

as in (66), the only non-negative concave function $U(h)$ with $U(h_{ct}) = 0$ that would allow it is

$$(\frac{h_1}{\pi_1})^{\pi_1} \dots (\frac{h_k}{\pi_k})^{\pi_k}, \tag{71}$$

but (69) and (71) are maximized by $h = \pi$, which is the prior distribution used to compute $\pi_E(\theta|x)$, instead of being maximized at one commonly agreed maximum uncertainty distribution, h_{ref}. Therefore (69) and (71) fail to make it into meaningful measures of the uncertainty in h. Calling measure of uncertainty to an object that assumes that one's subjective prior π represents maximum uncertainty about θ, is not something that De Groot is likely to have settled for.

Among all the measures covered by (41), (44), (47), (49), (53) and by (55), in Examples 5 to 10, only the ones computed under $\pi = h_{ref}$, $I_\phi^{h_{ref}}(E)$, can be posed as in (66) with a $U(\cdot)$ satisfying the requirements of Definition 6.4, but these special cases only allow one to assess the value of E through its impact on the posterior distributions obtained from that reference prior and not through its impact on the posterior distributions obtained from other priors.

Definition 4.1 leads to measuring the information in E through $I_\phi(E)$, in Definition 5.1-5.2, which encompasses $I_{DG}(E)$ as a special case but which also includes measures

like the ones in Examples 5 to 10 that are not interpretable as prior uncertainty minus expected posterior uncertainty about θ. In the extended setting of Definition 5.1-5.2, the information about θ in E still relates to the expected impact of E on the uncertainty about θ, but that impact does not have to be measured on an additive scale, as in (66). Even though some might wish that the measure of the *statistical information* in E and the measure of the *self-information* about θ in its own distribution always add up in the sense of De Groot (1962), the two concepts being measured are different enough to allow for other types of relations between their measures.

Of course, if one was willing to consider an uncertainty measure to be any non-negative concave function on the simplex of R^k taking the value 0 on all extreme points of that set, without any reference to a standard baseline maximum uncertainty distribution, then all measures in Examples 5 to 10 could be posed as in (66), and in this extended sense Definition 5.1 and the definition of information in De Groot (1962) would coincide. But we believe that assuming that one's subjective prior *de facto* represents maximum uncertainty about θ, the way allowed for if one adopts this extended definition of uncertainty measure, goes against the spirit of what De Groot was trying to capture through his definition of information in an experiment.

On the side, note that carrying out experiment E could lead to $\pi_E(\theta|x)$ being less concentrated than π, and therefore one could be left with less information (more uncertainty) about θ in $\pi_E(\theta|x)$ than there was in π (i.e., the entropy or the variance of the observed posterior could be larger, or its modal probability smaller than the ones of the prior). Therefore, even in the context of Examples 2 to 4 that fit in the framework of De Groot (1962), $I(\pi_E(\theta|x)) - I(\pi)$ can be negative and it should not be used as a measure of the information about θ in $X = x$. Finding that a priori one was too certain about θ is quit informative and worthwhile.

It is unfortunate that the term "information" has come to mean a wide array of different concepts, such as *statistical information* in experiment E, *observed information* in $X = x$ and *self-information* about θ in a distribution h on Ω. But the word "information" already refers to all these concepts and it is therefore important that one precisely distinguishes and relates its meanings and its corresponding measures, the way we have tried to do in this manuscript.

8 An illustrative example

The measures of the information in $E = (X; P_\theta)$ can be used for comparing experiments based on statistical merit in design of experiments, and they are the base for the construction of reference priors for a given experiment (Bernardo 1979b, 2005a; Dawid 1979), and for the construction of minimally informative models for a given prior (Yuan and Clarke 1999). The measures of the information in $X = x$ serve as measures of the surprise about θ in $X = x$ which makes them useful as Bayesian model checking test statistics. The measures of the information in a distribution h can be used to assess the strength of knowledge about the corresponding random variable, and as measures of concentration or of the homogeneity of a population with probability measure h.

By linking the measure of the information in E, in $X = x$, and in h, under the generalized divergence framework set up in Definitions 5.1-5.2, 6.1-6.2, and 6.3, one is allowed a unified approach to all these problems. Here we briefly sketch where those links lead to in the context of linear normal experiments.

Consider $E = (X; P_\theta)$ to be a linear normal regression experiment that yields $X \in R^n$ distributed $N_n(A\theta, \sigma^2 I)$ with known σ and with $\theta \in R^p$. In this context, selecting an experiment requires choosing an $n \times p$ design matrix A. Let the prior distribution, π, be normal, $N_p(m_0, \sigma^2 V_0)$, with known $E_\pi[\theta] = m_0$ and $Var_\pi[\theta] = \sigma^2 V_0$, and thus let the prior predictive, $p_\pi(x)$, be $N_n(Am_0, \sigma^2(I + AV_0A^{'}))$ and the posterior distribution, $\pi_E(\theta|x)$, be $N_p(m_F, \sigma^2 V_F)$ with $E_{\pi_E(\theta|x)}[\theta]$ equal to $m_F = (V_0^{-1} + A^{'}A)^{-1}(V_0^{-1}m_0 + A^{'}X)$ and with $Var_{\pi_E(\theta|x)}[\theta] = \sigma^2 V_F$, where $V_F = (V_0^{-1} + A^{'}A)^{-1}$.

When E and F are any two such linear normal experiments, E is "sufficient for" F if and only if $I^\theta_{Fi}(E) - I^\theta_{Fi}(F)$ is non-negative definite (Hansen and Torgersen, 1974). Given that here $I^\theta_{Fi}(E) = A^{'}A$ which does not depend on θ, for this type of experiments one can restrict attention to the measures of the information covered by Example 1, that in this context can all be posed as $\gamma(A^{'}A)$, with $\gamma(\cdot)$ being a real function such that $\gamma(M_1) \geq \gamma(M_2)$ whenever $M_1 - M_2$ is non-negative definite. That is precisely what the DoE literature requires of an optimality design criteria ever since Kiefer (1959), even though they do not link this requirement to the sufficiency argument given above.

Next, we compute the measures of the information in E, in $X = x$, and in $h(\theta)$, associated to Examples 2, 3, 4 and 8, for linear normal experiments, and find that all the measures of the information in E are indeed functions of $A^{'}A$ as described above.

Example 2 (cont): When $H = I$, and therefore when comparing linear normal experiments in terms of their performance under estimation with loss function $L(\theta, d) = (d - \theta)^{'}(d - \theta)$, one measures the information in E through

$$I_V^{\pi, L_1}(E) = \text{trace}\{\sigma^2(V_0 - (V_0^{-1} + A^{'}A)^{-1})\}, \tag{72}$$

which leads one to choose an experiment with a design matrix A minimizing the trace of the posterior variance-covariance, $V_F = (V_0^{-1} + A^{'}A)^{-1}$, among the set of available experiments (see, e.g., Chaloner 1984). The measure of the information in $X = x$ associated to (72) is

$$I_V^{\pi, L_1}(x) = (m_F - m_0)^{'}(m_F - m_0). \tag{73}$$

Finding $I_V^{\pi, L_1}(x)$ to be surprisingly larger than $I_V^{\pi, L_1}(E)$ by finding that when $X \sim P_\pi$ the probability that $I_V^{\pi, L_1}(X) > I_V^{\pi, L_1}(x)$ is very small might indicate that the prior assumptions might have not been reasonable.

Moreover, in this context the uncertainty about θ in a distribution on R^p is measured through the trace of its variance-covariance matrix and therefore, the impact of $X = x$ on the uncertainty about θ, $U_V^{L_1}(\pi) - U_V^{L_1}(\pi_E(\theta|x))$, is the trace of $\sigma^2(V_0 - (V_0^{-1} + A^{'}A)^{-1})$, that in this case is always non-negative and it coincides with $I_V^{\pi, L_1}(E)$, which only holds for linear normal experiments.

Example 3 (cont): Here, one measures the information in E through

$$I_V^{\pi,L_2}(E) = (2\pi\sigma^2)^{-p/2}(|(V_0^{-1} + A^{'}A)^{-1}|^{-1/2} - |V_0|^{-1/2}), \tag{74}$$

which leads one to choose an experiment with an A maximizing the determinant of $(V_0^{-1} + A^{'}A)$ and thus the determinant of $V_{p_\pi}[X]$. The measure of the information in $X = x$ associated to (74) is the modal posterior probability density minus the posterior probability density at the prior mode,

$$I_V^{\pi,L_2}(x) = \frac{1}{(2\pi\sigma^2)^{p/2}|V_F|^{1/2}}(1 - \exp\{-\frac{1}{2\sigma^2}(m_F - m_0)^{'}V_F^{-1}(m_F - m_0)\}). \tag{75}$$

In the context of this example, the information about θ in a distribution on R^p, is measured through the value of the probability density at its mode, and the impact of $X = x$ on the information about θ, $I_V^{L_2}(\pi_E(\theta|x)) - I_V^{L_2}(\pi)$, is the difference between the modal posterior density and the modal prior density, which in this case coincides with $I_V^{\pi,L_2}(E)$ but not for more general type of experiments.

Example 4 (cont): Here, one measures the information in E through

$$I_{MI}^{\pi}(E) = (\log|V_0| - \log|(V_0^{-1} + A^{'}A)^{-1}|)/2, \tag{76}$$

which like in the previous example leads one to choose an experiment with an A that maximizes the determinant of $V_{p_\pi}[X]$ (see, e.g., Sebastiani and Wynn 2000). The measure of the information in $X = x$ associated to (76) is the Kullback-Leibler divergence between the posterior and the prior distributions,

$$2I_{MI}^{\pi}(x) = \log\frac{|V_0|}{|V_F|} + \frac{1}{\sigma^2}(m_F - m_0)^{'}V_0^{-1}(m_F - m_0) + \text{trace}\{V_0^{-1}V_F\} - p. \tag{77}$$

Here the uncertainty about θ is measured through the entropy of its distribution, and the reduction of the uncertainty due to $X = x$, $U_{MI}(\pi) - U_{MI}(\pi_E(\theta|x))$, is the entropy of the posterior minus the entropy of the prior, that for these linear normal experiments coincides with $I_{MI}^{\pi}(E)$, but not in general.

Example 8 (cont): Goel and Padilla (1994), following the lead in Goel (1988), extends (48) for infinite Ω through $I_{H_1}^{\pi}(E) = 1 - E_{p_\pi}[\exp\{E_\pi[\ln\{p_\theta(x)/p_\pi(x)\}]\}]$, and computes it for various exponential family experiments. Using their results, it follows that under our linear normal regression model,

$$I_{H_1}^{\pi}(E) = 1 - \exp\{-\text{trace}\{AV_0A^{'}\}/2\}, \tag{78}$$

and

$$I_{H_2}^{\pi}(E) = \text{trace}\{AV_0A^{'}\}/2, \tag{79}$$

which is the only measure of the information in E considered in this section that is additive under independent experiments. In the context of this measure, one chooses the experiment with an A maximizing the trace of $AV_0A^{'}$ and thus maximizing the trace

of $V_{p_\pi}[X]$; when $V_0 = I$, that reduces to maximizing the trace of $I_{Fi}(E)$. The measure of the information in $X = x$ associated to $I^{\pi}_{H_1}(E)$ is

$$I^{\pi}_{H_1}(x) = 1 - \exp\{E_\pi[\ln p_\theta(x)] - \ln E_\pi[p_\theta(x)]\}, \tag{80}$$

that for these linear normal experiments becomes

$$I^{\pi}_{H_1}(x) = 1 - \frac{|V_0|^{1/2}}{|V_F|^{1/2}} \exp\{-\frac{1}{2}(\frac{1}{\sigma^2}(m_0 - m_F)' V_F^{-1}(m_0 - m_F) + \text{trace}\{V_F^{-1}V_0\} - p)\}. \tag{81}$$

Given that $I^{\pi}_{H_1}(E)$ can not be written as (66), here the impact of E and of $X = x$ on the information in the distribution of θ can not be measured on an additive scale.

9 Concluding remarks

The role of likelihood ratios in statistical inference is widely recognized but their role in DoE is not. In this paper our main goal was to draw attention to the role played by the convex ordering of likelihood ratio statistics in the characterization of the measure of the information in an experiment, and therefore in the foundations of DoE. As a consequence of ignoring that link, the DoE literature lacks a clear understanding on why do optimality design criteria qualify as such and its scope is too narrow in that it focuses on a small subset of valid optimality design criteria.

Some researchers in DoE might be disappointed by the fact that Definition 4.1 is making the choice of an information measure as wide as possible instead of narrowing it down, but that is an unavoidable consequence of information being a highly multidimensional concept. As a consequence of Blackwell-Sherman-Stein theorem, in Proposition 3.1-3.2, if one imposed any extra requirement in Definition 4.1 other than linearity under mixture experiments, some of the generalized divergence measures covered by Definition 5.1-5.2 would be excluded from consideration as information measures and Definition 4.1 would stop characterizing all valid measures of the information in an experiment.

The secondary goal of the paper was to present Definition 6.1-6.2 on how observed information should be measured. That definition is very simple and its motivation is hard to argue against even outside the sufficiency ordering framework and yet, we have not found it anywhere in the literature (even though it owes a lot to DeGroot 1984). An alternative way of presenting our ideas would begin by defining $I_\phi(x)$ as in Definition 6.1-6.2, and then defining $I_\phi(E)$ to be equal to $E_{p_\pi}[I_\phi(x)]$. Instead, we focused on a characterization of the measure of the information in E that naturally leads to Definition 5.1-5.2 and makes Definition 6.1-6.2 into its off-shot.

The choice of an experiment based on statistical merit only, is a *decision problem* in which the reward from experiment $E = (X; P_\theta)$ is its likelihood ratio or posterior distribution, the utility function is convex, the utility of the reward is the observed information in $X = x$, and the expected utility of E is the statistical information in E.

As remarked earlier, the axiom set in Definition 4.1 does neither assume that θ is random nor that the results from E will be used in a *statistical decision problem*.

Nevertheless, one might expect the information in E to depend on the strength of knowledge about θ because the less uncertain one is about θ, the less value one finds in the outcome to come from E (in the limit, anyone with a one-point prior should not learn anything from any experiment). When choosing the design optimality criteria to use, it is very convenient to think in terms of loss functions and/or prior distributions the way it is illustrated in Section 5.3, (with the understanding that if one needs to take experimental costs into consideration one also has to include in the loss terms that typically depend on sample size and/or the specific outcome observed, as described in Lindley, 1972, 2000 and in Bernardo and Smith, 1994). Once an optimality criteria is chosen, there is no difference between the Bayesian and the non-Bayesian way of using it to plan for an experiment and thus, we find it unfortunate that some insist in exporting to DoE the same divide that separates Bayesian from Non-Bayesian inference.

By letting Ω be the source from which an input message, θ, is picked, E and F be discrete memoryless channels and $X = x$ and $Y = y$ be the corresponding output messages, comparing the fidelity of channels E and F as in information theory is analogous to comparing the information in experiments E and F. For the ones asking for a comparison of Definition 4.1 to the axiomatic approach to information theory (see, e.g., Shore and Johnson 1980, Ebanks et al. 1998), note that their axiom sets apply to the measure of the *information observed* and not to the measure of the *information in E* that is the object in Definition 4.1. Furthermore, the typical axiom sets in information theory implicitly assume that the goal is to summarize/update the information about θ in a distribution on Ω as in Bernardo (1979a), which explains that if one abides by their main axiom sets, one ends up normatively requiring that the information in E be measured through its mutual information in Example 4. Definition 4.1 encompasses all valid measures of the information in E and therefore it includes measures that might not be useful in specific settings like the one of information theory.

The framework presented in this manuscript covers as special cases the comparison of the information in experiment $E = (X; P_\theta)$ that observes from X with the information in the sub-experiment E_T that observes from a statistic $T(X)$, and the comparison of E with experiments that observe censored or truncated versions of X.

On the other hand, in this manuscript it is assumed that all the information gathered is to be used after the experiment is completed and therefore it excludes from consideration sequential experiments, where the information in different parts of the experiment are used for different purposes. An extension of Blackwell-Sherman-Stein theorem is needed in order to relate sequential sufficiency as defined in Greenshtein (1996) to variability orderings of sequential likelihood ratio statistics, which would help extend generalized divergence measures to sequential experiments.

Appendix: convex functions and convex ordering

Real valued functions $\phi(\cdot)$ defined on a convex set C in R^k are convex if $\phi(\alpha u + (1-\alpha)v) \leq \alpha\phi(u) + (1-\alpha)\phi(v)$ for every $\alpha \in (0,1)$ and every u and v in C. If $\phi(\cdot)$ is convex and $h(\cdot)$ is convex and non-decreasing on the range of $\phi(\cdot)$, then $h(\phi(\cdot))$ is convex.

Any convex function $g(u)$ on R (R_+) induces convex functions $\phi_i(u_1,\ldots,u_k) = g(u_i)$ on R^k (R_+^k) by acting coordinate wise. Given any arbitrary collection of convex functions on the same subset, any linear combination of these functions with non-negative coefficients and the pointwise supremum of this collection of functions are convex. In particular, if $g(\cdot)$ is convex on R (R_+), $a_0 \in R$ and $t_i \geq 0$, then $\phi(u_1,\ldots,u_k) = a_0 + \sum_{i=1}^k t_i g(u_i)$ and $\phi(u_1,\ldots,u_k) = \max\{a_0, \max_i\{g(u_i)\}\}$ are convex functions on R^k (R_+^k). For a detailed coverage of convex analysis, see Rockafellar (1970).

Convex functions take on their larger values over "extreme regions." Any measure of the form $E[\phi(U)]$ with a convex function $\phi(\cdot)$, serves as a measure of the variability of the random vector U. The convex ordering defined next, allows one to compare random variables in terms of their variabilities.

Definition 9.1. *If $U = (U_1,\ldots,U_k)$ and $V = (V_1,\ldots,V_k)$ are random vectors such that*

$$E[\phi(U)] \geq E[\phi(V)], \tag{82}$$

for every real valued function $\phi(\cdot)$ that is convex on the union of the supports of U and V, then U is said to be larger than V in the convex order, denoted by $U \geq_{cx} V$.

When $U \geq_{cx} V$, U is more likely to take extreme values than V and thus U is more spread out than V. In fact, one can re-phrase Definition 9.1 by stating that U is larger than V in the convex order, if U is more variable than V irrespective of the way one measures variability. In particular, when U and V are real valued random variables, $U \geq_{cx} V$ implies that for every $a \in R$, $E[(U-a)^2] \geq E[(V-a)^2]$, and thus $Var[U] \geq Var[V]$. In general, $U \geq_{cx} V$ implies that $Var[U] - Var[V]$ is a non-negative definite matrix. Also, $U \geq_{cx} V$ implies $U_i \geq_{cx} V_i$ for $i = 1,\ldots,k$, as well as the convex ordering between any given set of components of U and V. For a very nice exposition on convex orderings, see Shaked and Santikumar (1994, chap.2 and 5).

References

Ali, S. M., and Silvey, S. D. (1965). "Association between random variables and the dispersion of a Radom Nykodim derivative." *Journal of the Royal Statistical Society, Series B*, 27: 100-107.

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

Barnard, G. A. (1951). "The theory of information (with discussion)." *Journal of the Royal Statistical Society, Series B*, 13: 131-142.

—— (1959). Discussion of "Optimum experimental designs," by J. Kiefer. *Journal of the Royal Statistical Society, Series B*, 21: 311-312.

Barnard, G. A., Jenkins, G. M., and Winsten, C. B. (1962). "Likelihood inference and time series (with discussion)." *Journal of the Royal Statistical Society, Series B*, 24: 321-372.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. New York: Wiley.

Barron, A. R. (1999). "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems." In J. M. Bernardo, A. P. Dawid, and A. F. M. Smith, (eds.), *Bayesian Statistics 6*, 27-52. Oxford: Oxford University Press.

Bassan, B., and Scarsini, M. (1991). "Convex orderings for stochastic processes." *Commentationes Mathematicae Universitatis Carolinae*, 32: 115-118.

Basu, D. (1975). "Statistical information and likelihood (with discussion)." *Sankhya A*, 37: 1-71.

Bayarri, M. J., and Berger, J. O. (1999). "Quantifying surprise in the data and model verification." In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 6*, 53-82. Oxford: Oxford University Press.

Bernardo, J. M. (1979a). "Expected information as expected utility." *The Annals of Statistics*, 7: 686-690.

—— (1979b). "Reference posterior distributions for Bayesian inference (with discussion)." *Journal of the Royal Statistical Society, Series B*, 41: 113-147.

—— (2005a). "Reference analysis." In D. Dipak, and C. R. Rao (eds.), *Bayesian Thinking: Modeling and Computation, Handbook of Statistics 25*, 17-90. Amsterdam: North Holland.

—— (2005b). "Intrinsic credible regions: An objective Bayesian approach to interval estimation (with discussion)." *Test*, 14: 317-384.

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

Birnbaum, A. (1962). "On the foundations of statistical inference (with discussion)." *Journal of the American Statistical Association*, 67: 269-326.

—— (1969). "Concepts of statistical evidence." In S. Morgenbesser, P. Suppes, and M. White (eds.), *Science and Methodology*, 112-143. Saint Martin's Press.

Blackwell, D. (1951). "Comparison of experiments." In *Proceengs 2nd Berkeley Symposium on Mathematical Statistics and Probability*, 93-102. Berkeley: University of California Press.

—— (1953). "Equivalent comparison of experiments." *Annals of Mathematical Statistics*, 24: 265-272.

Blackwell, D., and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.

Box, G. E. P. (1980). "Sampling and Bayes inference in scientific modelling and robustness." *Journal of the Royal Statistical Society, Series A*, 143: 383-420.

Chaloner, K. (1984). "Optimal Bayesian experimental designs for linear models." *The Annals of Statistics*, 12: 283-300.

Chaloner, K., and Verdinelli, I. (1995). "Bayesian experimental design: A review." *Statistical Science*, 3: 273-304.

Chernoff, H. (1952). "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations." *Annals of Mathematical Statistics*, 23: 493-507.

Clarke, B. (1996). "Implications of reference priors for prior information and for sample size." *Journal of the American Statistical Association*, 91: 173-184.

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Csiszár, I. (1963). "Eine informationstheoretische ungleichung und ihre Anwendung auf den beweis der ergodizitat von markoffschen ketten." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 8: 85-108.

—— (1967). "Information-type measures of difference of probability distributions, and indirect observations." *Studia Scientarum Mathematicarum Hungarica*, 2: 191-213.

Dawid, A. P. (1979). Discussion of "Reference posterior distributions for Bayesian inference," by J. Bernardo. *Journal of the Royal Statistical Society, Series B*, 41: 132-133.

—— (1998). "Coherent measures of discrepancy, uncertainty and dependence, with applications to Bayesian predictive experimental design." Technical Report 139, Dept. of Statistical Science, University College, London.

Dawid, A. P., and Sebastiani, P. (1999). "Coherent dispersion criteria for optimal experimental design." *The Annals of Statistics*, 27: 65-81.

DeGroot, M. H. (1962). "Uncertainty, information and sequential experiments." *Annals of Mathematical Statistics*, 33: 404-419.

—— (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

—— (1979). Discussion of "Reference posterior distributions for Bayesian inference," by J. Bernardo. *Journal of the Royal Statistical Society, Series B*, 41: 135-136.

—— (1984). "Changes in utility as information." *Theory and Decision*, 17: 283-303.

Ebanks, B., Sahoo, P., and Sander, W. (1998). *Characterization of Information Measures*. Singapore: World Scientific Press.

Fazekas, I., and Liese, F. (1996). "Some properties of the Hellinger transform and its application in classification problems." *Computers and Mathematical Applications,* 31: 107-116.

Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics." *Phyilosophical Transactions of the Royal Society*, A 222: 309-368.

Goel, P. K. (1983). "Information measures and Bayesian hierarchical models." *Journal of the American Statistical Association*, 78: 408-410.

—— (1988). "Comparison of experiments and information in censored data." In S. Gupta, and J. Berger (eds.), *Statistical Decision Theory and Related Topics IV* (Vol. 2), 335-349. New York: Springer Verlag.

Goel, P. K., and DeGroot, M. H. (1979). "Comparison of experiments and information measures." *The Annals of Statistics*, 5: 1066-1077.

—— (1981). "Information about hyperparameters in hierarchical models." *Journal of the American Statistical Association*, 76: 140-147.

Goel, P. K., and Ginebra, J. (2003). "When is one experiment 'always better than' another?" *Journal of the Royal Statistical Society, Series D*, 52: 515-537.

Goel, P. K., and Padilla, M. L. R. (1994). "Generalized Hellinger transforms as information measures." In *ASA Proceedings of the Bayesian Statistical Science Section*, 78-83.

Gollier, C. (2001). *The Economics of Risk and Time*. Cambridge: The MIT Press.

González, E., and Ginebra, J. (2001). "Bayesian heuristic for multiperiod control." *Journal of the American Statistical Association*, 96: 1113-1121.

Good, I. J. (1950). *Probability and the Weighting of Evidence*. New York: Hafner Publisher.

—— (1960). "Weight of evidence, corroboration, explanatory power, information and the utility of experiments." *Journal of the Royal Statistical Society, Series B*, 22: 319-331.

—— (1966). "A derivation of the probabilistic explicata of information." *Journal of the Royal Statistical Society, Series B*, 28: 578-581.

—— (1979). "Studies in the history of probability and statistics. A. M. Turing's statistical work in world war II." *Biometrika*, 66: 393-396.

—— (1985). "Weight of evidence: A brief survey." In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds.), *Bayesian Statistics 2*, 249-270. Amsterdam: Elsevier Science Publisher.

Greenshtein, E. (1996). "Comparison of sequential experiments." *The Annals of Statistics*, 24: 436-448.

Hansen, O. H., and Torgersen, E. N. (1974). "Comparison of linear normal experiments." *The Annals of Statistics*, 2: 367-373.

Haussler, D., and Opper, M. (1997). "Mutual information, metric entropy and cumulative relative entropy risk." *The Annals of Statistics*, 25: 2451-2492.

Heyer, H. (1982). *Theory of Statistical Experiments*. New York: Springer Verlag.

Kiefer, J. (1959). "Optimum experimental designs (with discussion)." *Journal of the Royal Statistical Society, Series B*, 21: 272-319.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Le Cam, L. (1964). "Sufficiency and approximate sufficiency." *Annals of Mathematical Statistics*, 35: 1419-1455.

—— (1975). "Distances between experiments." In *A Survey of Statistical Design and Linear Models*, 383-396. Amsterdam: North Holland.

—— (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer Verlag.

—— (1996). "Comparison of experiments. A short review." In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen (eds.), *Statistics, Probability and Game Theory*, 127-138. IMS Lecture Notes-Monograph Series, Vol. 30.

Le Cam, L., and Yang, G. L. (2000). *Asymptotics in Statistics; Some Basic Concepts* (2nd ed.). New York: Spriner Verlag.

Lehmann, E. L. (1959, 1986). *Testing Statistical Hypothesis* (1rst and 2nd ed.). New York: Wiley.

—— (1983). *Theory of Point Estimation*. New York: Wiley.

—— (1988). "Comparing location experiments." *The Annals of Statistics*, 16: 521-533.

Lindley, D. V. (1956). "On a measure of the information provided by an experiment." *Annals of Mathematical Statistics*, 27: 986-1005.

—— (1961). "Dynamic programming and decision theory." *Applied Statistics*, 10: 39-51.

—— (1972). *Bayesian Statistics, a Review*. Philadelphia: SIAM.

—— (2000). "The philosophy of statistics (with discussion)." *The Statistician*, 49: 293-337.

Papaioannou, T. (2001). "On distances and measures of information: A case of diversity." In Ch. A. Charalambides, M. V. Koutras, and N. Balakrishnan (eds.), *Probability and Statistical Models with Applications*, 503-515. Boca Raton: Chapman and Hall.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley.

Raiffa, H., and Schlaifer, R. O. (1961). *Applied Statistical Decision Theory.* Cambridge: M.I.T. Press.

Rao, C. R. (1982). "Analysis of diversity: A unified approach." In S. Gupta, and J. Berger (eds.), *Statistical Decision Theory and Related Topics III* (Vol. 2), 233-250. New York: Academic Press.

Rényi, A. (1961). "On measures of entropy and information." In *Proceedings 4th Berkeley Symposium on Mathematical Statistics and Probability*, 547-561. Berkeley: University of California Press.

—— (1967a). "Statistics and information theory." *Studia Scientarum Mathematicarum Hungarica*, 2: 249-256.

—— (1967b). "On some basic problems of statistics from the point of view of information theory." In *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*, 531-543. Berkeley: University of California Press.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton New Jersey: University Press.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Sebastiani, P., and Wynn, H. P. (2000). "Maximum entropy sampling and optimal Bayesian experimental design." *Journal of the Royal Statistical Society, Series B*, 62: 145-157.

Shaked, M., and Shantikumar, J. G. (1994). *Stochastic Orders and their Applications.* New York: Academic Press.

Shannon, C. E. (1948). "A mathematical theory of communications." *Bell System Tech. Journal*, 27: 379-423, 623-656.

Sherman, S. (1951). "On a theorem of Hardy, Littlewood, Polya and Blackwell." *Proceedings of the National Academy of Sciences*, 37: 826-831.

Shiryaev, A. N., and Spokoiny, V. G. (2000). *Statistical Experiments and Decisions. Asymptotic Theory.* Singapore: World Scientific Press.

Shore, J. E., and Johnson, R. W. (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy." *IEEE Transactions on Information Theory*, 26: 26-37.

Silvey, S. D. (1964). "On a measure of association." *Annals of Mathematical Statistics*, 35: 1157-1166.

Soofi, E. (2000). "Principal information theoretic approaches." *Journal of the American Statistical Association*, 95: 1349-1353.

Stein, C. (1951). "Notes on a seminar on theoretical statistics; Comparison of experiments." Unpublished report.

Stone, M. (1961). "Non-equivalent comparison of experiments and their use for experiments involving location parameters." *Annals of Mathematical Statistics*, 32: 326-332.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. New York: Walter Gruyter.

Torgersen, E. N. (1976). "Comparison of statistical experiments (with discussion)." *Scandinavian Journal of Statistics*, 3: 186-208.

—— (1991a). *Comparison of Experiments*. Cambridge: Cambridge University Press.

—— (1991b). "Stochastic orders and comparison of experiments." In K. Mosley, and M. Scarsini (eds.), *Stochastic Orders and Decision under Risk*, 334-371. IMS Lecture Notes-Monograph Series, Vol. 19.

—— (1994). "Information orderings and stochastic orderings." In *Stochastic Orders and their Applications*, 275-319. New York: Academic Press.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Dordretch: Kluwer.

Verdinelli, I., and Kadane, J. B. (1992). "Bayesian designs for maximizing information and output." *Journal of the American Statistical Association*, 87: 510-515.

Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

Yuan, A., and Clarke, B. (1999). "A minimally informative likelihood for decision analysis: Robustness and illustration." *Canadian Journal of Statistics*, 27: 649-665.

Acknowledgments

The author is extremely grateful to Prem K. Goel for all the insights that he has generously shared with him; Without his help and encouragement this manuscript would have never been written. He is also grateful to Enrique Gonzalez-Davila, to the associate editor, and to the referee for their constructive questions and comments.

This paper is intended as a tribute to D. Blackwell, L. Le Cam, E. N. Torgersen, and M. De Groot, on the work of which this manuscript relies so much upon.