Maximum entropy derived and generalized under idempotent probability to address Bayes-frequentist uncertainty and model revision uncertainty

David R. Bickel

Maximum entropy derived and generalized under idempotent probability to address Bayes-frequentist uncertainty and model revision uncertainty

David R. Bickel

April 18, 2019

Ottawa Institute of Systems Biology
Department of Biochemistry, Microbiology, and Immunology
Department of Mathematics and Statistics
University of Ottawa
451 Smyth Road
Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670
dbickel@uottawa.ca

Abstract

Typical statistical methods of data analysis only handle determinate uncertainty, the type of uncertainty that can be modeled under the Bayesian or confidence theories of inference. An example of indeterminate uncertainty is uncertainty about whether the Bayesian theory or the frequentist theory is better suited to the problem at hand. Another example is uncertainty about how to modify a Bayesian model upon learning that its prior is inadequate.

Both problems of indeterminate uncertainty have solutions under the proposed framework. The framework is based on an information-theoretic definition of an incoherence function to be minimized. It generalizes the principle of choosing an estimate that minimizes the reverse relative entropy between it and a target posterior distribution such as a confidence distribution. The simplest form of the incoherence function, called the incoherence distribution, is a min-plus probability distribution, which is equivalent to a possibility distribution rather than a measure-theoretic probability distribution.

An analog of Bayes's theorem for min-plus probability leads to a generalization of minimizing relative entropy and thus of maximizing entropy. The framework of minimum incoherence is applied to problems of Bayesian-confidence uncertainty and to parallel problems of indeterminate uncertainty about model revision.

Keywords: Bayes-frequentist continuum; Bayesian model checking; blended inference; coding theory; fiducial inference; information theory; Kullback-Leibler divergence; possibility theory

1 Introduction

Uncertainty in statistical inference has multiple layers. The lowest layer is uncertainty about an unknown quantity or about the truth of a hypothesis according to a mathematical model.

The next layer is uncertainty about which model to rely on under some higher level model. A simple remedy is Bayesian model averaging, integrating multiple posterior distributions with respect to a posterior distribution over the models. The resulting average posterior distribution may then generate estimates that minimize posterior expected loss. That approach is insufficient in itself when there is uncertainty about the prior over the models or about the highest level hyperprior over candidates for that prior. The more frequentist approach of fiducial model averaging (Bickel, 2018a) is also insufficient when there is uncertainty about the highest level confidence distribution.

The first two layers involve what Walley (1991, §5.1.2), citing the 1933 edition of Knight (2012), calls *determinate uncertainty*, that which can be represented in terms of a probability distribution. The third layer of uncertainty in statistical inference involves *indeterminate uncertainty*, model uncertainty that cannot be represented by a single posterior distribution given the data (Walley, 1991, §5.1.2), whether that posterior is a confidence distribution or the result of Bayes's theorem.

Some forms of indeterminate uncertainty succumb to minimizing risk over a set of candidate models. Unlike Bayesian model averaging, that approach can discriminate between Bayesian models that have no hyperprior over them (e.g., Grünwald and Dawid, 2004) and can even adjudicate between frequentist and Bayesian point estimates (Samaniego, 2010). Other forms of indeterminate uncertainty can be reduced a problem of minimizing expected loss with respect to a distribution that maximizes entropy in the broad sense of minimizing relative entropy. Extreme forms call for a generalization of maximum entropy such as the framework proposed in this paper.

In that framework, indeterminate uncertainty is made determinate by finding the *most coherent distribution*, the posterior distribution that minimizes a measure of the incoherence between the uncertain distributions of that quantity. If parameter estimates or other actions are needed, they may then be found by minimizing the expected loss with respect to the most coherent distribution. Even when there is no unique most coherent distribution, basing inference only on the most coherent distributions can benefit from a substantial reduction in indeterminacy.

Incoherence between Bayesian and frequentist inference occurs in many applications when there is indeterminate uncertainty about which theory of statistics to use.

Example 1. Let $p(0)$ denote a Bayesian posterior probability that ψ, the parameter of interest, is 0 or some other null hypothesis value. Simple lower bounds on the Bayes factor that quantifies the relevancy of the evidence for the null hypothesis H_0 are available under various sets of assumptions reviewed in Held and Ott (2018). From such a lower bound on the Bayes factor, Bayes's theorem generates $\underline{p}(0)$, a lower bound on $p(0)$, given any prior probability of H_0 or given any interval of such prior probabilities.

That approach to hypothesis testing faces two major obstacles. First, a lower bound on the posterior probability of H_0, no matter how low, cannot in itself imply that the posterior probability of H_0 is low (Sellke et al., 2001), for that could be anywhere in the interval $\left[\underline{p}(0), 1\right]$. For example, if $\underline{p}(0) = 10^{-7}$, then $p(0) \in \left[10^{-7}, 1\right]$, which is not much more useful than the trivial $p(0) \in [0, 1]$.

Second, many argue that since H_0 cannot be exactly true, it should always be assigned 0 prior probability, necessarily leading to $p(0) = 0$ regardless of what data are observed (e.g., Bernardo, 2011). The view that the posterior probability of H_0 must be 0 agrees with the standard practice of reporting a confidence interval for ψ in the sense that the confidence density for a continuous parameter (Efron, 1993) puts 0 fiducial probability at each possible value of ψ, including 0. Further, confidence distributions tend to be close to objective Bayes posteriors, posterior distributions derived from improper priors. While there are important exceptions (Fraser, 2011; Bahamyirou and Marchand, 2015), confidence distributions and objective Bayes posteriors are similar for large enough samples under broad conditions (Veronese and Melilli, 2018a,b).

Giving due weight to the arguments that raise the second obstacle provides a straightforward path through the first obstacle. That will be seen to result from minimizing incoherence. ▲

Another example of Bayes-frequentist indeterminacy calls for blending fully Bayesian inference on one hand with frequentist or objective Bayes inference on the other.

Example 2. A scientist observes $x = 2$ as the value of $X \sim \mathrm{N}(\theta, 1)$ for an unknown θ. In the domain of application, the prior distribution of θ is $\mathrm{N}\left(0, \sigma_0^2\right)$, with $\sigma_0 = 1/8$ according to the working hypothesis. Since the working hypothesis might be incorrect, the scientist considers not only the resulting fully Bayesian posterior but also the usual confidence intervals for θ, all of which may be encoded in as the single confidence distribution $\mathrm{N}(x, 1) = \mathrm{N}(2, 1)$ (e.g., Schweder and Hjort, 2016). (In this case, $\mathrm{N}(2, 1)$ is also the objective

Bayes posterior from the uniform prior.) Judging both the fully Bayesian approach and the frequentist approach relevant to inference about θ, the scientist uses $\mathrm{N}\left(0, \widetilde{\sigma}_0^2\right)$ as the prior distribution, where the value of $\widetilde{\sigma}_0 > 1/8$ is determined by blending the confidence distribution with the fully Bayesian posterior such that incoherence is minimized. ▲

In addition to Bayes-frequentist incoherence, a manifestation of indeterminate uncertainty that challenges statistical data analysis is that of model selection in light of inference after model selection but without a unique hyperprior over the models. Understood broadly, indeterminate model selection includes inference after updating a prior distribution in light of prior-data conflict (Walter and Augustin, 2009; Evans and Jang, 2011; Bickel, 2018b) in spite of its violation of Bayes's theorem (cf. Mayo, 2018, §6.4).

Example 3. A scientist who had been heavily relying on $\mathrm{N}\left(0, \sigma_0^2\right)$ with $\sigma_0 = 1/8$ as the prior distribution of θ observes that $X = 2$ and becomes aware of independent arguments against $\sigma_0 = 1/8$. Since those arguments do not counter all the evidence that seemed to support $\sigma_0 = 1/8$, they do not warrant jumping to a default prior, proceeding as if it were never reasonable to use $\sigma_0 = 1/8$. Nonetheless, the initial prior must be weakened in the direction of the uniform prior (cf. Evans, 2015, §5.7), which is practically equivalent to $\mathrm{N}\left(0, \sigma_0^2\right)$ for sufficiently high σ_0. That may be accomplished by increasing σ_0 to σ_0^*, the value that minimizes the incoherence between the posteriors from the initial prior and the uniform prior. ▲

In spite of the similarities between Examples 2-3, their most coherent distributions differ since $\sigma_0^* < \widetilde{\sigma}_0$. The extent of the discrepancy is recorded in Figure 1.

The formalism of minimum incoherence is presented in Section 2 with an eye to statistical theory and applications. That framework justifies a generalization of maximum entropy, which solves the problems of Example 1 and a problem of model revision conditional on a new insight (Section 3). Examples 2 and 3 instead require hierarchical incoherence, the topic of Section 4. Finally, Appendix A compares merits of the recommended framework to those of a generalized principle of maximum expected utility.

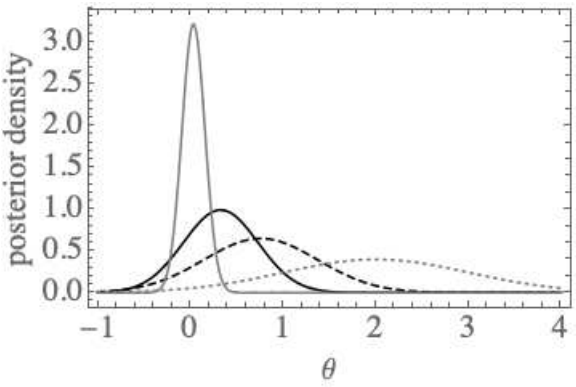

Figure 1: Posterior density functions of θ, the mean of a unit-variance observable random variable X, given the observation that $X = 2$. The two lighter curves correspond to the fully Bayesian posterior based on the normal prior of mean 0 and standard deviation 1/8 (solid) and to the confidence distribution that is equal to the objective Bayes posterior based on the uniform prior (dotted). Between the two lighter curves, each of the two darker curves represents the most coherent density function for statistical inference according to whether it is a blend of Bayesian and frequentist inference (dashed), as per Example 2, or whether it results from revising the first prior (solid), as per Example 3.

2 Incoherence as generalized redundancy under indeterminate uncertainty

2.1 Source distributions and coding distributions for statistics

The incoherence framework specified in Section 2.2 follows the distinction in information theory between "source distributions" and "coding distributions," which is analogous to the distinction in statistical theory between true distributions and distributions that estimate them, respectively. Applications of minimum incoherence require identifying source distributions and coding distributions in the problem at hand. For the examples in Section 1, the source distributions and coding distributions are posterior distributions understood broadly enough to include not only Bayesian posterior distributions but also prior-free posteriors such as confidence distributions and other fiducial distributions (e.g., Hannig et al., 2016) as well as posteriors from inferential models (Martin and Liu, 2013).

In Examples 1 and 2, the fully Bayesian posteriors would be source distributions, whereas the objective Bayes posteriors and fiducial distributions would be coding distributions according to arguments that the latter approximate or estimate the former. Singh, Xie, and Strawderman (2007) and Xie and Singh (2013)

considered confidence distributions as estimates of ψ, the parameter of interest. In the case that ψ is an indicator of the truth of a hypothesis, as in Examples 1-4, a fiducial probability in the form of an observed confidence level (Polansky, 2007) serves as an estimate of ψ (Bickel, 2012a). Wilkinson (1977, §6.2) instead presented fiducial probability, a generalization of an observed confidence level, not as a level of belief but rather as an *estimate* of a level of belief since it cannot be updated via Bayes's theorem. If *true* levels of belief are equated with Bayesian posterior probabilities (Wilkinson, 1977, §6.2), that would suggest that confidence distributions and other fiducial distributions are estimates of Bayesian posterior distributions.

That seemingly academic viewpoint has two concrete implications for statistical inference. First, it indicates that fiducial distributions may well be useful as estimates in spite of criticisms that they deviate from some certain laws of probability (e.g., Grundy, 1956; Lindley, 1958; Evans, 2015, §3.6), for estimates need not have all the properties of what they estimate. Second, it guides the application of Section 2.2 by letting fiducial distributions as estimates serve as coding distributions and the fully Bayesian targets of estimation serve as source distributions.

In contrast, in Example 3, the fully Bayesian posterior, now considered inadequate, would no longer qualify as a source distribution. However, because it is not abandoned entirely, it could serve as a coding distribution that in some sense estimates a source distribution for purposes of model revision. Since a candidate for the revised model in that example is based on the uniform prior, the corresponding objective Bayes posterior is appropriate as a candidate source distribution.

It will be seen in Section 4 that the differences in what are considered source distributions and coding distributions account for the discrepancy noted in Figure 1.

2.2 Derivation of the incoherence function

In the problem of source coding, a *source distribution* p generates an outcome to be encoded by an algorithm called a *code* that is idealized as a *coding distribution* q (Picard, 2004, §1.2). If p and q are probability mass functions on some finite set Θ of possible outcomes, then the number of characters a code requires to describe an outcome is abstracted as the *ideal codelength* (cf. Grünwald, 2007), the amount of information $S(q(\theta))$ that would be gained by observing an outcome θ in addition to the information in its probability

$q(\theta)$ according to q as the coding distribution. Assuming that $q(\theta) > 0 \implies S(q(\theta)) \geq 0$ and that

$$q(\theta), q(\theta') > 0 \implies S(q(\theta)\, q(\theta')) = S(q(\theta)) + S(q(\theta'))$$

for outcomes θ and θ', it follows that $S(q(\theta)) \propto -\log q(\theta)$, with the result that $S(q(\theta)) = -\log q(\theta)$ may be adopted as a convention (Aczél and Daróczy, 1975, §0.2). That defines $E_p(S(q)) = \sum_\theta p(\theta) S(q(\theta))$ as the expected ideal codelength of q with respect to p.

The *redundancy* of a coding distribution q with respect to a source distribution p is this difference in expected ideal codelengths:

$$\operatorname{rdndc}_p(q) \coloneqq E_p(S(q)) - E_p(S(p)) = E_p\left(\log \frac{p}{q}\right) = \sum_\theta p(\theta) \log \frac{p(\theta)}{q(\theta)}, \tag{1}$$

which is the Kullback-Leibler divergence between p and q (Cover and Thomas, 2006, §13.1). Relaxing the assumption that Θ is finite, the redundancy a coding probability distribution q with respect to source probability distribution p is the Kullback-Leibler divergence

$$\operatorname{rdndc}_p(q) = E_p\left(\log \frac{d\,p}{d\,q}\right)$$

if q dominates p or $\operatorname{rdndc}_p(q) = \infty$ otherwise. For example, $\operatorname{rdndc}_p(q) = \int p(\theta) \log(p(\theta) / q(\theta))\, d\theta$ if p and q are source and coding probability density functions on Θ with respect to the Lebesgue measure. Given a source distribution p, the goal is to determine which members of a set $\mathcal{Q}$ of possible coding distributions minimize the redundancy:

$$\operatorname{best}_p(\mathcal{Q}) \coloneqq \{q' \in \mathcal{Q} : \operatorname{rdndc}_p(q') = \inf_{q \in \mathcal{Q}} \operatorname{rdndc}_p(q)\}. \tag{2}$$

The function minimized, $q \mapsto \operatorname{rdndc}_p(q)$, is called *reverse relative entropy* to distinguish it from $p \mapsto \operatorname{rdndc}_p(q)$, called *relative entropy*.

In typical statistical applications, Θ is an uncountable set of parameter values that index the distributions of some family $\{x \mapsto f_\theta(x) : \theta \in \Theta\}$ of probability density functions that could have generated the observed sample x; that is, x is a realization of the random sample $X \sim f_\theta$ for some unknown $\theta \in \Theta$. Accordingly,

the source distribution p is a posterior distribution such as a confidence distribution (Bickel, 2012a) or the *Bayesian posterior* that results from applying Bayes's theorem to a prior distribution. As a distribution intended to estimate p, the coding distribution q may also be a posterior distribution.

Under the indeterminate uncertainty introduced in Section 1, p is not known but is considered as a member of a set $\mathcal{P}$ of posterior distributions. For example, $\mathcal{P}$ may be the set of Bayesian posterior distributions corresponding to different priors, or $\mathcal{P}$ could be the set of confidence distributions corresponding to different methods of constructing confidence intervals (Bickel, 2012b). To generalize minimum redundancy to that setting, we need an "incoherence function" to minimize over a set of "decisions" that is analogous to the function $q \mapsto \text{rdndc}_p(q)$ that is minimized over the set $\mathcal{Q}$ of coding distributions in equation (2). Such a function is defined by its desirable properties.

Definition 1. Let $\mathfrak{Q}$ and $\mathfrak{P}$ denote the power sets (collections of all subsets) of $\mathcal{Q}$ and $\mathcal{P}$, respectively. $\mathfrak{Q}$ and $\mathfrak{P}$ induce $\mathfrak{Q} \otimes \mathfrak{P}$, the power set of the *decision set* $\mathcal{Q} \times \mathcal{P}$. The *set of most coherent decisions* in $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ restricted to a non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$ is

$$\text{cohrnt}(\mathcal{D}_0 \,|\, \mathcal{D}_1) \coloneqq \left\{ d' \in \mathcal{D}_0 : I\left(d' \,|\, \mathcal{D}_1\right) = \min\left(\inf_{d \in \mathcal{D}_0} I(d \,|\, \mathcal{D}_1), I(\emptyset | \, \mathcal{D}_1)\right) \right\}, \tag{3}$$

where $I(d \,|\, \mathcal{D}_1) = I(\{d\} \,|\, \mathcal{D}_1)$ for all $d \in \mathcal{Q} \times \mathcal{P}$, assuming the I written in equation (3) is a conditional incoherence function according to the following definition. The function $I(\bullet)$ on $\mathfrak{Q} \otimes \mathfrak{P}$ is an *incoherence function*, and the function $I(\bullet|\bullet)$ on $(\mathfrak{Q} \otimes \mathfrak{P}) \times (\mathfrak{Q} \otimes \mathfrak{P})$ is a *conditional incoherence function* if all of these conditions hold:

1. $I(\mathcal{D}_0 \,|\, \mathcal{D}_1) = I(d \,|\, \mathcal{D}_1)$ for every $d \in \text{cohrnt}(\mathcal{D}_0 \,|\, \mathcal{D}_1)$ and every non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$. (The incoherence of making a decision in $\mathcal{D}_0$ given that it must be in $\mathcal{D}_1$ is the incoherence of making one of the most coherent such decisions. That is because the decision maker minimizes incoherence according to equation (3).)

2. $I(\mathcal{D}_0) = I(\mathcal{D}_0 \,|\, \mathcal{Q} \times \mathcal{P})$ for all $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$. (The incoherence of making a decision in $\mathcal{D}_0$ is the incoherence of making a decision in $\mathcal{D}_0$ given that it is in the decision set.)

3. There is a real number $\underline{I}$ such that $I(\mathcal{D}_0 \,|\, \mathcal{D}_1) = \underline{I}$ for every non-empty $\mathcal{D}_0, \mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$ such that $\mathcal{D}_1 \subset \mathcal{D}_0$. (Since every decision must be in the set to which it is restricted, the incoherence of making

a decision in a set that includes the whole restriction set does not depend on those particular sets.)

4. There are no non-empty $\mathcal{D}_0, \mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$ such that $\text{cohrnt}\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = \emptyset$. (At least one of the decisions in $\mathcal{D}_0$ restricted to $\mathcal{D}_1$ has to be among the most coherent.)

5. All incoherence functions share a function ϕ such that $I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = \phi\left(I\left(\mathcal{D}_0 \cap \mathcal{D}_1\right), I\left(\mathcal{D}_1\right)\right)$ for all $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ and all non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$. (Following the analogous condition that Cox (1961) used to characterize standard conditional probability, this constrains the incoherence of making a decision in $\mathcal{D}_0$ given that the decision must be taken in $\mathcal{D}_1$; see also Lapointe and Bobée (2000).)

6. Let $Q : \mathcal{Q} \times \mathcal{P} \to \mathcal{Q}$ and $P : \mathcal{Q} \times \mathcal{P} \to \mathcal{P}$ denote the functions such that $Q\left((q, p)\right) = q$ and $P\left((q, p)\right) = p$ for all $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. There are real numbers c_{cond} and c_{joint} and a positive real number $c_{>0}$ such that

$$I\left(Q = q \,|\, P = p\right) = \left(\text{rdndc}_p\left(q\right) - c_{\text{cond}}\right) c_{>0} \tag{4}$$

$$I\left(Q = q, P = p\right) = \left(\text{rdndc}_p\left(q\right) - c_{\text{joint}}\right) c_{>0} \tag{5}$$

for all $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. (Given that a decision must include a certain source distribution p, incoherence is a linear function of the redundancy with respect to that source distribution. That is fitting since redundancy is a linear function of the difference in ideal codelengths according to equation (1).)

The rationale for each condition appears in parentheses. The conditions are compressed as follows.

Lemma 1. *I and $I\left(\bullet|\bullet\right)$ are an incoherence function and a conditional incoherence function on $\mathfrak{Q} \otimes \mathfrak{P}$ and $\left(\mathfrak{Q} \otimes \mathfrak{P}\right) \times \left(\mathfrak{Q} \otimes \mathfrak{P}\right)$ if and only if, for all $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$, there is a real number $\underline{I}$ such that*

$$I\left(\mathcal{D}_0\right) \geq \underline{I} = I\left(\mathcal{Q} \times \mathcal{P}\right) \tag{6}$$

$$I\left(\emptyset\right) \geq \sup_{d \in \mathcal{Q} \times \mathcal{P}} I\left(d\right) \tag{7}$$

$$I\left(\mathcal{D}_0\right) = \inf\nolimits_{d \in \mathcal{D}_0} I\left(d\right) \tag{8}$$

$$I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = I\left(\mathcal{D}_0 \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} \;\forall\, \mathcal{D}_1 \in \left(\mathfrak{Q} \otimes \mathfrak{P}\right) \backslash \emptyset \tag{9}$$

$$I\left(Q = q \,|\, P = p\right) - \underline{I} \propto \mathrm{rdndc}_p\left(q\right) - \inf\nolimits_{q' \in \mathcal{Q}} \mathrm{rdndc}_p\left(q'\right) \;\forall\, q \in \mathcal{Q}, p \in \mathcal{P}\,. \tag{10}$$

Proof. This proof's condition numbers are those of the list in Definition 1. Consider any $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ and any real value $\underline{I}$.

$(\Longleftarrow)$. Assume equations (6)-(10) for any non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$. Condition 5 follows immediately from equation (9). By equation (9), $I\left(\mathcal{D}_0 \,|\, \mathcal{Q} \times \mathcal{P}\right) = I\left(\mathcal{D}_0 \cap \left(\mathcal{Q} \times \mathcal{P}\right)\right) - I\left(\mathcal{Q} \times \mathcal{P}\right) + \underline{I}$, which is $I\left(\mathcal{D}_0\right)$ by $\mathcal{D}_0 \subset \mathcal{Q} \times \mathcal{P}$ and equation (6). That establishes condition 2.

Equations (9) and (8) generate the conditional counterparts of equations (6), (7), and (8):

$$I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = I\left(\mathcal{D}_0 \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} \geq \underline{I} = I\left(\left(\mathcal{Q} \times \mathcal{P}\right) \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} = I\left(\mathcal{Q} \times \mathcal{P} \,|\, \mathcal{D}_1\right) \tag{11}$$

$$I\left(\emptyset | \, \mathcal{D}_1\right) = I\left(\emptyset \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} \geq \sup_{d \in \mathcal{Q} \times \mathcal{P}} I\left(\{d\} \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} = \sup_{d \in \mathcal{Q} \times \mathcal{P}} I\left(d \,|\, \mathcal{D}_1\right) \tag{12}$$

$$I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = I\left(\mathcal{D}_0 \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} = \inf\nolimits_{d \in \mathcal{D}_0} I\left(\{d\} \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} = \inf\nolimits_{d \in \mathcal{D}_0} I\left(d \,|\, \mathcal{D}_1\right). \tag{13}$$

For every $d \in \mathrm{cohrnt}\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right)$, equation (3) gives $I\left(d \,|\, \mathcal{D}_1\right) = \inf_{d \in \mathcal{D}_0} I\left(d \,|\, \mathcal{D}_1\right)$, which is $I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right)$ according to equation (13), and so condition 1 must hold. If $\mathcal{D}_1 \subset \mathcal{D}_0$, then equations (8) and (13) imply

$$I\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) = \inf\nolimits_{d \in \mathcal{D}_1} I\left(\{d\} \cap \mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I} = I\left(\mathcal{D}_1\right) - I\left(\mathcal{D}_1\right) + \underline{I},$$

resulting in condition 3. By equations (3) and (12), $\mathrm{cohrnt}\left(\mathcal{D}_0 \,|\, \mathcal{D}_1\right) \neq \emptyset$ if $\mathcal{D}_0 \neq \emptyset$, which is what condition 4 requires.

For the $(\Longleftarrow)$ direction of the proof, it now suffices to prove condition 6, which asserts equations (4)

and (5). Equation (4) follows directly from equation (10). From formulas (9) and (10), there is a constant of proportionality $c_\propto > 0$ such that

$$\begin{aligned} I\left(Q=q, P=p\right) &= I\left(Q=q \,|\, P=p\right) - \underline{I} + I\left(P=p\right) \\ &= \left(\mathrm{rdndc}_p\left(q\right) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p\left(q'\right)\right) c_\propto + I\left(P=p\right) \\ &= \left(\mathrm{rdndc}_p\left(q\right) - \left(\inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p\left(q'\right) - \frac{I\left(P=p\right)}{c_\propto}\right)\right) c_\propto, \end{aligned}$$

which reduces to equation (5).

Therefore, $I\left(\bullet\right)$ and $I\left(\bullet|\bullet\right)$ are an incoherence function and a conditional incoherence function on $\mathfrak{Q} \otimes \mathfrak{P}$ and $\left(\mathfrak{Q} \otimes \mathfrak{P}\right) \times \left(\mathfrak{Q} \otimes \mathfrak{P}\right)$.

$(\implies)$. Assume $I\left(\bullet\right)$ and $I\left(\bullet|\bullet\right)$ are an incoherence function and a conditional incoherence function on $\mathfrak{Q} \otimes \mathfrak{P}$ and $\left(\mathfrak{Q} \otimes \mathfrak{P}\right) \times \left(\mathfrak{Q} \otimes \mathfrak{P}\right)$. By equation (3) and by conditions 2 and 1,

$$I\left(\mathcal{D}_0\right) = I\left(\mathcal{D}_0 \,|\, \mathcal{Q} \times \mathcal{P}\right) = \inf_{d \in \mathcal{D}_0} I\left(d \,|\, \mathcal{Q} \times \mathcal{P}\right) = \inf_{d \in \mathcal{D}_0} I\left(d\right),$$

establishing equation (8), from which it follows that $I\left(\mathcal{D}_0\right) \geq I\left(\mathcal{Q} \times \mathcal{P}\right)$ since $\mathcal{D}_0 \subset \mathcal{Q} \times \mathcal{P}$. Thus, since $I\left(\mathcal{Q} \times \mathcal{P}\right) = I\left(\mathcal{Q} \times \mathcal{P} \,|\, \mathcal{Q} \times \mathcal{P}\right) = \underline{I}$ by conditions 2 and 3, equation (6) holds.

If equation (7) were false, then there would be a $d \in \mathcal{Q} \times \mathcal{P}$ such that $I\left(\emptyset\right) < I\left(\{d\}\right)$, which, with condition 2, would imply that $\mathrm{cohrnt}\left(\{d\} \,|\, \mathcal{Q} \times \mathcal{P}\right) = \emptyset$. Since that contradicts condition 4, equation (7) cannot be false.

Condition 6 says there are a real number c_{cond} and a positive number $c_{>0}$ such that equation (4) holds for all $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. Since condition 3 implies that $I\left(\mathcal{Q} \times \mathcal{P} \,|\, P=p\right) = \underline{I}$,

$$\underline{I} = I\left(\mathcal{Q} \times \mathcal{P} \,|\, P=p\right) = I\left(Q \in \mathcal{Q} \,|\, P=p\right) = \left(\inf_{q \in \mathcal{Q}} \mathrm{rdndc}_p\left(q\right) - c_{\mathrm{cond}}\right) c_{>0}\,.$$

Thus, $c_{\mathrm{cond}} = \inf_{q \in \mathcal{Q}} \mathrm{rdndc}_p\left(q\right) - \underline{I} \,/\, c_{>0}$, which substituted into equation (4) gives

$$I\left(Q=q \,|\, P=p\right) = \left(\mathrm{rdndc}_p\left(q\right) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p\left(q'\right)\right) c_{>0} + \underline{I} \tag{14}$$

for all $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. Equation (10) follows with $c_{>0}$ as the constant of proportionality.

A similar expression comes from analogous reasoning with equation (5) instead of equation (4). By equation (6),

$$\underline{I} = I(\mathcal{Q} \times \mathcal{P}) = (\inf_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathrm{rdndc}_p(q) - c_{\mathrm{joint}})\, c_{>0}\,.$$

Then $c_{\mathrm{joint}} = \inf_{q \in \mathcal{Q}, p \in \mathcal{P}} \mathrm{rdndc}_p(q) - \underline{I} \,/\, c_{>0}$, and substitution into equation (5) gives

$$I(Q = q, P = p) = (\mathrm{rdndc}_p(q) - \inf_{q' \in \mathcal{Q}, p' \in \mathcal{P}} E_{p'}(u_{q'}))\, c_{>0} + \underline{I} \tag{15}$$

for all $q \in \mathcal{Q}$ and $p \in \mathcal{P}$. Again applying equation (8),

$$\begin{aligned} I(P = p) &= \inf_{q' \in \mathcal{Q}} I(Q = q', P = p) \\ &= (\inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p(q') - \inf_{q' \in \mathcal{Q}, p' \in \mathcal{P}} E_{p'}(u_{q'}))\, c_{>0} + \underline{I} \end{aligned}$$

according to equation (15). Then, by equations (14) and (15),

$$\begin{aligned} I(Q = q, P = p) - I(Q = q \,|\, P = p) &= (\inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p(q') - \inf_{q' \in \mathcal{Q}, p' \in \mathcal{P}} E_{p'}(u_{q'}))\, c_{>0} \\ &= I(P = p) - \underline{I}\,. \end{aligned}$$

That forces the ϕ in condition 5 to satisfy

$$I(\mathcal{D}_0 \,|\, \mathcal{D}_1) = \phi(I(\mathcal{D}_0 \cap \mathcal{D}_1), I(\mathcal{D}_1)) = I(\mathcal{D}_0 \cap \mathcal{D}_1) - I(\mathcal{D}_1) + \underline{I}$$

for any non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$. That entails equation (9). □

The conditions are simplified further in Section 2.3.

2.3 The incoherence function as a min-plus probability distribution

Two simplifications to the result of Lemma 1 do not affect the set of chosen decisions: first, replace equation (7) with $I(\emptyset) = \infty$; second, let $\underline{I} = 0$. In fact, any functions $\Pi(\bullet)$ and $\Pi(\bullet|\bullet)$ on $\mathfrak{Q} \otimes \mathfrak{P}$ and $(\mathfrak{Q} \otimes \mathfrak{P}) \times (\mathfrak{Q} \otimes \mathfrak{P})$ are known as a *min-plus probability distribution* and a *conditional min-plus probability distribution*, respectively, if they satisfy $I(\emptyset) = \infty$ and equations (6), (8), and (9) with $\underline{I} = 0$. The "min" in "min-plus"

refers to the "inf" in equation (8), where standard probability would instead have addition for marginalization, and the "plus" refers to the "+" in $I(\mathcal{D}_0 \cap \mathcal{D}_1) = I(\mathcal{D}_0 \,|\, \mathcal{D}_1) + I(\mathcal{D}_1)$, where standard conditional probability would instead have multiplication.

The considerable simplification without changing the most coherent decisions warrants working only with incoherence functions that are also min-plus probability distributions.

Theorem 1. *For every $I(\bullet)$ and $I(\bullet|\bullet)$ that are an incoherence function and a conditional incoherence function on $\mathfrak{Q} \otimes \mathfrak{P}$ and $(\mathfrak{Q} \otimes \mathfrak{P}) \times (\mathfrak{Q} \otimes \mathfrak{P})$, let $\Pi(\bullet)$ and $\Pi(\bullet|\bullet)$ be a min-plus probability distribution and a conditional min-plus probability distribution such that*

$$\left\{ d' \in \mathcal{D}_0 : I\left(d' \,|\, \mathcal{D}_1\right) = \inf_{d \in \mathcal{D}_0} I(d \,|\, \mathcal{D}_1) \right\} = \left\{ d' \in \mathcal{D}_0 : \Pi\left(d' \,|\, \mathcal{D}_1\right) = \inf_{d \in \mathcal{D}_0} \Pi(d \,|\, \mathcal{D}_1) \right\} \tag{16}$$

for any $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ and non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$. Then $\Pi(\bullet)$ and $\Pi(\bullet|\bullet)$ are an incoherence function and a conditional incoherence function on $\mathfrak{Q} \otimes \mathfrak{P}$ and $(\mathfrak{Q} \otimes \mathfrak{P}) \times (\mathfrak{Q} \otimes \mathfrak{P})$.

Proof. Since equations (6)-(10) characterize incoherence functions and conditional incoherence functions according to Lemma 1, it suffices to derive those equations for $\Pi(\bullet)$ and $\Pi(\bullet|\bullet)$ using $\underline{\Pi} = 0$ in place of $\underline{I}$. As min-plus probability distributions and conditional min-plus probability distributions, they by definition satisfy equations (6), (8), and (9). Since min-plus probability distributions also satisfy $I(\emptyset) = \infty$, equation (7) holds for $\Pi(\bullet)$.

Thus, equation (16) is only needed to prove that $\Pi(\bullet)$ and $\Pi(\bullet|\bullet)$ satisfy equation (10) with $\underline{\Pi} = 0$ in place of $\underline{I}$. By equations (16) and (9),

$$\begin{gathered}
\arg\inf_{d \in \mathcal{D}_0} I(d \,|\, \mathcal{D}_1) = \arg\inf_{d \in \mathcal{D}_0} \Pi(d \,|\, \mathcal{D}_1) \\
\arg\inf_{d \in \mathcal{D}_0} I(\{d\} \cap \mathcal{D}_1) - I(\mathcal{D}_1) + \underline{I} = \arg\inf_{d \in \mathcal{D}_0} \Pi(\{d\} \cap \mathcal{D}_1) - \Pi(\mathcal{D}_1) + \underline{\Pi} \\
\arg\inf_{d \in \mathcal{D}_0 \cap \mathcal{D}_1} I(d) - I(\mathcal{D}_1) + \underline{I} = \arg\inf_{d \in \mathcal{D}_0 \cap \mathcal{D}_1} \Pi(d) - \Pi(\mathcal{D}_1) + 0 \\
\arg\inf_{d \in \mathcal{D}_0 \cap \mathcal{D}_1} I(d) = \arg\inf_{d \in \mathcal{D}_0 \cap \mathcal{D}_1} \Pi(d) \\
\arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} I(Q = q, P = p) = \arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} \Pi(Q = q, P = p) \\
\arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} I(Q = q \,|\, P = p) + I(P = p) - \underline{I} = \arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} \Pi(Q = q \,|\, P = p) + \Pi(P = p) - \underline{\Pi} \\
\arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} I(Q = q \,|\, P = p) + I(P = p) = \arg\inf_{(q,p) \in \mathcal{D}_0 \cap \mathcal{D}_1} \Pi(Q = q \,|\, P = p) + \Pi(P = p),
\end{gathered}$$

which can only hold for every $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ and non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$ if $q \mapsto \Pi(Q = q \,|\, P = p) + \Pi(P = p)$ is a monotonic increasing, linear function of $q \mapsto I(Q = q \,|\, P = p) + I(P = p)$. Thus, $q \mapsto \Pi(Q = q \,|\, P = p)$ is a monotonic increasing, linear function of $q \mapsto I(Q = q \,|\, P = p)$, which is in turn a monotonic increasing, linear function of $q \mapsto \mathrm{rdndc}_p(q)$ according to equation (10). It follows that $q \mapsto \Pi(Q = q \,|\, P = p)$ is a monotonic increasing, linear function of $q \mapsto \mathrm{rdndc}_p(q)$. That means there are a real number c_0 and a positive number c_1 such that

$$\Pi(Q = q \,|\, P = p) = (\mathrm{rdndc}_p(q) - c_0)\, c_1 \tag{17}$$

for all $q \in \mathcal{Q}$. By equations (11) and (13) ,

$$0 = \underline{\Pi} = \Pi(\mathcal{Q} \times \mathcal{P} \,|\, P = p) = \Pi(Q \in \mathcal{Q} \,|\, P = p) = \inf_{q \in \mathcal{Q}} \Pi(Q = q \,|\, P = p) = (\inf_{q \in \mathcal{Q}} \mathrm{rdndc}_p(q) - c_0)\, c_1,$$

implying that $c_0 = \inf_{q \in \mathcal{Q}} \mathrm{rdndc}_p(q)$. Substituting that into equation (17), with $\underline{\Pi} = 0$, provides formula (10) with c_1 as the constant of proportionality. □

Since min-plus probability distributions are the simplest versions of equivalent incoherence functions, all incoherence functions in the rest of the paper are min-plus probability distributions and for that reason are called *incoherence distributions*. Likewise, every conditional incoherence function that is also a conditional min-plus probability distribution is called a *conditional incoherence distribution*. As a further simplification, the constant of proportionality in equation (10) is set to 1.

Limiting attention to incoherence distributions has another advantage: it facilitates applications of theorems in an extensive theoretical and applied mathematics literature on min-plus probability (e.g., Akian et al., 1994; Quadrat, 1998) and other, equivalent forms of idempotent probability (e.g., Puhalskii, 2001), including max-plus probability (e.g., Akian et al., 1994; Quadrat, 1995; Fleming et al., 2010; Fitzpatrick, 2013). Theories equivalent in wide generality include ranking function theory in the philosophy literature (Spohn, 2012) and, in the case of conditional possibility equivalent to $I(\mathcal{D}_0 \cap \mathcal{D}_1) = I(\mathcal{D}_0 \,|\, \mathcal{D}_1) + I(\mathcal{D}_1)$, possibility theory in the fuzzy logic literature (e.g., De Baets et al., 1999; Lapointe and Bobée, 2000).

3 A derivation of generalized maximum entropy for statistics applications

By equation (4) of Definition 1, minimizing the conditional incoherence distribution given a source distribution results in the same coding distributions as minimizing redundancy with respect to that distribution. If that source distribution is only known to lie in a set $\mathcal{P}_0$, then we instead minimize the conditional incoherence distribution given that the source distribution is in $\mathcal{P}_0$.

Proposition 1. *Assume $p \mapsto I(P = p)$ is constant, that is, $I(P = \bullet) \equiv 0$. For any non-empty $\mathcal{P}_0 \in \mathfrak{P}$,*

$$\arg\inf_{q \in \mathcal{Q}} I(Q = q \mid P \in \mathcal{P}_0) = \arg\inf_{q \in \mathcal{Q}} \inf_{p \in \mathcal{P}_0} \mathrm{rdndc}_p(q).$$

Proof. By equations (8), (9), and (10), there is a real number c such that

$$\begin{aligned} I(Q = q \mid P \in \mathcal{P}_0) &= I(Q = q, P \in \mathcal{P}_0) - I(P \in \mathcal{P}_0) \\ &= \inf_{p \in \mathcal{P}_0} I(Q = q, P = p) - I(P \in \mathcal{P}_0) \\ &= \inf_{p \in \mathcal{P}_0} I(Q = q \mid P = p) + I(P = p) - I(P \in \mathcal{P}_0) \\ &= \inf_{p \in \mathcal{P}_0} I(Q = q \mid P = p) + 0 - I(P \in \mathcal{P}_0) \\ &\propto \inf_{p \in \mathcal{P}_0} \mathrm{rdndc}_p(q) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p(q') + c \end{aligned}$$

□

Instead of minimizing conditional incoherence over coding distributions given a set of source distributions, we can instead minimize conditional incoherence over source distributions given a set of coding distributions.

Example 4. Example 1, continued. To indicate whether the null hypothesis is true, let $\theta = \theta(0) = 0$ if $\psi = 0$ and $\theta = \theta(\psi) = 1$ if $\psi \neq 0$, where ψ is a scalar parameter of interest. The set of posterior distributions of θ satisfying the set of assumptions leading to $\underline{p}(0)$ as the posterior probability's lower bound is $\mathcal{P} = \left\{(\pi_0, 1 - \pi_0) : \pi_0 \in \left[\underline{p}(0), 1\right]\right\}$, where $(\pi_0, 1 - \pi_0)$ is the Bernoulli distribution with probability π_0 or $1 - \pi_0$ of an outcome of 0 or 1.

Let C denote either a confidence density function of ψ or an objective Bayes posterior density function of ψ, where either density is with respect to the Lebesgue measure. C's underlying assumption that $\psi \neq 0$ with probability 1 (Example 1) may be interpreted as its estimating θ to be arbitrarily close to 1. For example, given an arbitrarily small $\varepsilon > 0$, the estimate of θ could be found by minimizing the expected squared error loss over $[0, 1-\varepsilon]$ with respect to C:

$$\arg \min_{\widehat{\theta} \in [0,1-\varepsilon]} E_{\psi \sim C} \left(\left(\widehat{\theta} - \theta(\psi) \right)^2 \right) = \arg \min_{\widehat{\theta} \in [0,1-\varepsilon]} \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta(\psi) \right)^2 C(\psi) \, d\psi$$

$$\begin{aligned} &= \arg \min_{\widehat{\theta} \in [0,1-\varepsilon]} \lim_{\delta \downarrow 0} \int_{-\infty}^{-\delta} \left(\widehat{\theta} - 1 \right)^2 C(\psi) \, d\psi + \int_{-\delta}^{\delta} \left(\widehat{\theta} - 0 \right)^2 C(\psi) \, d\psi + \int_{\delta}^{\infty} \left(\widehat{\theta} - 1 \right)^2 C(\psi) \, d\psi \\ &= \arg \min_{\widehat{\theta} \in [0,1-\varepsilon]} \int_{-\infty}^{\infty} \left(\widehat{\theta} - 1 \right)^2 C(\psi) \, d\psi = \arg \min_{\widehat{\theta} \in [0,1-\varepsilon]} \left(\widehat{\theta} - 1 \right)^2 = 1 - \varepsilon. \end{aligned}$$

Thereby estimating θ to be $1-\varepsilon$ is equivalent to specifying the Bernoulli distribution $(\varepsilon, 1-\varepsilon)$ as q_C, the coding distribution of θ for an arbitrarily small $\varepsilon > 0$. The designation of q_C as a coding distribution and the members of $\mathcal{P}$ as source distributions agrees with Section 2.1's arguments that fiducial distributions in some sense estimate Bayesian posterior distributions.

Which of the fully Bayesian posterior distributions in $\mathcal{P}$ should be used given that $q_C = (\varepsilon, 1-\varepsilon)$ is chosen as a coding distribution? Equation (3) answers, "Those that minimize conditional incoherence given $Q = q_C$":

$$\text{cohrnt}\left(P \in \mathcal{P} \mid Q = q_C\right) = \left\{ p' \in \mathcal{P} : I\left(P = p' \mid Q = q_C\right) = \inf_{p \in \mathcal{P}} I\left(P = p \mid Q = q_C\right) \right\}.$$

Assuming the marginal incoherence of each coding distribution and source distribution is constant, that is,

$I(Q = \bullet) \equiv 0$ and $I(P = \bullet) \equiv 0$, equations (9) and (10) lead to

$$\begin{aligned}
I(P = p \mid Q = q_C) &= I(P = p, Q = q_C) - I(Q = q_C) \\
&= I(Q = q_C \mid P = p) + I(P = p) - I(Q = q_C) \\
&= I(Q = q_C \mid P = p) + 0 - 0 \\
&\propto \operatorname{rdndc}_p(q_C) - \inf_{q' \in \mathcal{Q}} \operatorname{rdndc}_p(q') \\
&= E_p\left(\log \frac{p}{q_C}\right) - \inf_{q' \in \mathcal{Q}} E_p\left(\log \frac{p}{q'}\right) \\
&= E_p\left(\log \frac{p}{(\varepsilon, 1-\varepsilon)}\right) - \inf_{\pi_0' \in [0,1]} E_p\left(\log \frac{p}{(\pi_0', 1-\pi_0')}\right) = E_p\left(\log \frac{p}{(\varepsilon, 1-\varepsilon)}\right).
\end{aligned} \tag{18}$$

Since ε is arbitrarily small, the most coherent distribution is

$$\begin{aligned}
\lim_{\varepsilon \downarrow 0} \operatorname{cohrnt}(P \in \mathcal{P} \mid Q = q_C) &= \lim_{\varepsilon \downarrow 0} \arg \inf_{p \in \mathcal{P}} E_p\left(\log \frac{p}{(\varepsilon, 1-\varepsilon)}\right) \\
&= \lim_{\varepsilon \downarrow 0} \arg \inf_{(\pi_0', 1-\pi_0') \in \left\{(\pi_0, 1-\pi_0): \pi_0 \in \left[\underline{p}(0), 1\right]\right\}} E_{(\pi_0', 1-\pi_0')}\left(\log \frac{(\pi_0', 1-\pi_0')}{(\varepsilon, 1-\varepsilon)}\right) \\
&= \lim_{\varepsilon \downarrow 0} \arg \inf_{(\pi_0', 1-\pi_0') \in \left\{(\pi_0, 1-\pi_0): \pi_0 \in \left[\underline{p}(0), 1\right]\right\}} \pi_0' \log \frac{\pi_0'}{\varepsilon} + (1-\pi_0') \log \frac{1-\pi_0'}{1-\varepsilon} \\
&= \left(\lim_{\varepsilon \downarrow 0} \arg \inf_{\pi_0' \in \left[\underline{p}(0), 1\right]} \pi_0' \log \frac{\pi_0'}{\varepsilon}, 1 - \lim_{\varepsilon \downarrow 0} \arg \inf_{\pi_0' \in \left[\underline{p}(0), 1\right]} \pi_0' \log \frac{\pi_0'}{\varepsilon}\right) \\
&= \left(\underline{p}(0), 1 - \underline{p}(0)\right),
\end{aligned}$$

which is the posterior distribution for which the posterior probability of the null hypothesis is $\underline{p}(0)$, the lower bound. For example, if $\underline{p}(0) = 10^{-7}$, then 10^{-7} is the most coherent choice of a posterior probability of the null hypothesis.

Since $\operatorname{rdndc}_p(q_C)$ is minimized as a function of p rather than as a function of q_C, what is minimized is relative entropy rather than redundancy or reverse relative entropy. Minimizing relative entropy is commonly called *maximum entropy* since $-\operatorname{rdndc}_p(q)$ is the Shannon entropy of p up to a constant if Θ is finite and q is uniform. Other applications of maximum entropy to blend Bayesian and frequentist inference appear in Bickel (2015), which justified maximum entropy on the basis of a minimax framework (Topsøe, 1979) rather than the minimum incoherence framework. ▲

Equation (18) is analogous to Bayes's theorem, with the constraint on the coding distribution ($Q = q_C$ in this case) corresponding to the observed data, $I(P = p \mid Q = q_C)$ to the posterior probability, $I(P = p)$ to the prior probability, and $I(Q = q_C \mid P = p)$ to the probability of the observation given a parameter value. For conditional incoherence distributions, equation (10) ensures that the analog of the likelihood function is, up to a constant term, the redundancy given $Q = q$ as a function of the distribution defining the expectation value:

$$p \mapsto \mathrm{rdndc}_p(q) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p(q').$$

The following corollary and its proof generalize the analogy to conditioning on sets of coding distributions rather than a single coding distribution. The max-plus probability analog of Bayes's theorem has been applied to the control of unmanned vehicles (Fitzpatrick, 2013); see also the possibility analog in Lapointe and Bobée (2000).

Theorem 2. *For any $\mathcal{D}_0 \in \mathfrak{Q} \otimes \mathfrak{P}$ and non-empty $\mathcal{D}_1 \in \mathfrak{Q} \otimes \mathfrak{P}$,*

$$I(\mathcal{D}_0 \mid \mathcal{D}_1) = I(\mathcal{D}_0) + I(\mathcal{D}_1 \mid \mathcal{D}_0) - I(\mathcal{D}_1) \tag{19}$$

and, if $I(P = \bullet) \equiv 0$, then, for any non-empty $\mathcal{P}_0 \in \mathfrak{P}$ and non-empty $\mathcal{Q}_0 \in \mathfrak{Q}$,

$$\arg\inf_{p \in \mathcal{P}_0} I(P = p \mid Q \in \mathcal{Q}_0) = \arg\inf_{p \in \mathcal{P}_0} \inf_{q \in \mathcal{Q}_0} \mathrm{rdndc}_p(q). \tag{20}$$

Proof. Consider any $p \in \mathcal{P}_0$. According to equation (9),

$$I(P = p \mid Q \in \mathcal{Q}_0) = I(P = p, Q \in \mathcal{Q}_0) - I(Q \in \mathcal{Q}_0),$$

from which equation (19) follows. By $I(P = \bullet) \equiv 0$ and equations (8), (9), and (10), there is a real number c such that

$$\begin{aligned} I(P = p \mid Q \in \mathcal{Q}_0) = I(P = p \mid Q \in \mathcal{Q}_0) &= I(Q \in \mathcal{Q}_0 \mid P = p) + I(P = p) - I(Q \in \mathcal{Q}_0) \\ &= \inf_{q \in \mathcal{Q}_0} I(Q = q \mid P = p) + 0 - I(Q \in \mathcal{Q}_0) \\ &\propto \inf_{q \in \mathcal{Q}_0} \mathrm{rdndc}_p(q) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p(q') + c. \end{aligned}$$

□

Equation (20) is the "generalization of the maximum entropy principle" of Csiszár (1985, p. 88), where "maximum entropy" refers to minimum relative entropy, $\inf_{p\in\mathcal{P}_0} \mathrm{rdndc}_p(q)$. Equation (20) is "generalized" from the case that $\mathcal{Q}_0 = \{q\}$ for a single coding distribution q to a $\mathcal{Q}_0$ containing more than one coding distribution. In that sense, generalized maximum entropy applies to indeterminate uncertainty beyond the indeterminate uncertainty already handled by maximum entropy.

The relevance of Theorem 2 to applied statistics is clearest in the $\mathcal{Q}_0 = \{q\}$ case, as in Example 4 and the next example.

Example 5. Let q denote the posterior distribution according to the initial Bayesian model. In light of new information, the scientist has the insight that the Bayesian model should be such that its posterior distribution is a member of some non-empty $\mathcal{P}_0 \in \mathfrak{P}$. Then, since the initial model's posterior distribution is a coding distribution (§2.1), the updated model given the insight is

$$q^* = \arg\inf_{p\in\mathcal{P}_0} I(P = p \mid Q = q) = \arg\inf_{p\in\mathcal{P}_0} \mathrm{rdndc}_p(q)$$

by equation (20). Thus, if the insight added no new information since $q \in \mathcal{P}_0$, then $q = q^*$; otherwise, q^* is an similar to q as possible without violating the insight corresponding to $\mathcal{P}_0$. That maximum-entropy approach to model revision had also been derived from the ideal Cromwell's rule defined under large deviations (Bickel, 2018b). ▲

4 Hierarchical incoherence from source distributions that are coding distributions

Let $\mathcal{R}$ denote a set of standard probability distributions on the same domain as the distributions in $\mathcal{P}$ and $\mathcal{Q}$, and let $\mathfrak{R}$ be the power set of $\mathcal{R}$. Let $R : \mathcal{R}\times\mathcal{Q}\times\mathcal{P} \to \mathcal{R}$, $Q : \mathcal{R}\times\mathcal{Q}\times\mathcal{P} \to \mathcal{Q}$, and $P : \mathcal{R}\times\mathcal{Q}\times\mathcal{P} \to \mathcal{P}$ denote the functions such that $R((r,q,p)) = r$, $Q((r,q,p)) = q$, and $P((r,q,p)) = p$ for all $r \in \mathcal{R}$, $q \in \mathcal{Q}$, and $p \in \mathcal{P}$. Consider I, a min-plus probability distribution on $\mathfrak{R}\otimes\mathfrak{Q}\otimes\mathfrak{P}$, the power set of $\mathcal{R}\times\mathcal{Q}\times\mathcal{P}$, such that $(\mathcal{Q}_0,\mathcal{P}_0) \mapsto I(Q\in\mathcal{Q}_0, P\in\mathcal{P}_0)$ is an incoherence distribution and $(\mathcal{R}_0,\mathcal{Q}_0) \mapsto$

$I\left(R \in \mathcal{R}_0, Q \in \mathcal{Q}_0 \,|\, P \in \mathcal{P}_0\right)$ is an incoherence distribution for every non-empty $\mathcal{P}_0 \in \mathfrak{P}$. Since incoherence distributions are min-plus probability distributions that are also incoherence functions, $I\left(Q \in \bullet, P \in \bullet\right)$ is a min-plus probability distribution such that

$$I\left(Q = q \,|\, P = p\right) = \mathrm{rdndc}_p\left(q\right) - \inf_{q' \in \mathcal{Q}} \mathrm{rdndc}_p\left(q'\right) \tag{21}$$

for all $r \in \mathcal{R}$ and $q \in \mathcal{Q}$, and $I\left(R \in \bullet, Q \in \bullet | \, P \in \mathcal{P}_0\right)$ is a min-plus probability distribution such that

$$I\left(R = r \,|\, Q = q, P \in \mathcal{P}_0\right) = \mathrm{rdndc}_q\left(r\right) - \inf_{r' \in \mathcal{R}} \mathrm{rdndc}_q\left(r'\right), \tag{22}$$

for each $r \in \mathcal{R}$, each $q \in \mathcal{Q}$, and each non-empty $\mathcal{P}_0 \in \mathfrak{P}$, as equation (10) with 1 as the constant of proportionality requires.

According to that hierarchical structure, the source distributions in $\mathcal{P}$ are estimated by the coding distributions in $\mathcal{Q}$, which in turn are source distributions estimated by the coding distributions in $\mathcal{R}$. Minimizing incoherence then guides inference conditional on some of the distributions, as in the next result.

Theorem 3. *Assume* $I\left(P = \bullet\right) \equiv 0$ *and* $I\left(Q = \bullet\right) \equiv 0$. *For any non-empty* $\mathcal{R}_0 \in \mathfrak{R}$ *and non-empty* $\mathcal{P}_0 \in \mathfrak{P}$,

$$\arg\inf_{q \in \mathcal{Q}} I\left(Q = q \,|\, R \in \mathcal{R}_0, P \in \mathcal{P}_0\right) = \arg\inf_{q \in \mathcal{Q}} \left(\inf_{p \in \mathcal{P}_0} \mathrm{rdndc}_p\left(q\right) + \inf_{r \in \mathcal{R}_0} \mathrm{rdndc}_q\left(r\right)\right).$$

Proof. Consider any non-empty $\mathcal{R}_0 \in \mathfrak{R}$ and $\mathcal{P}_0 \in \mathfrak{P}$. Substitutions using equation (8), equation (9), $I\left(P = \bullet\right) \equiv 0$, and $I\left(Q = \bullet\right) \equiv 0$ yield

$$\begin{aligned}
I\left(Q = q \,|\, R \in \mathcal{R}_0, P \in \mathcal{P}_0\right) &= I\left(Q = q \,|\, P \in \mathcal{P}_0\right) + I\left(R \in \mathcal{R}_0 \,|\, Q = q, P \in \mathcal{P}_0\right) + c', \\
&= I\left(Q = q\right) + I\left(P \in \mathcal{P}_0 \,|\, Q = q\right) + I\left(R \in \mathcal{R}_0 \,|\, Q = q, P \in \mathcal{P}_0\right) + c'' \\
&= 0 + \inf_{p \in \mathcal{P}_0} I\left(P = p \,|\, Q = q\right) + \inf_{r \in \mathcal{R}_0} I\left(R = r \,|\, Q = q, P \in \mathcal{P}_0\right) \\
&= \inf_{p \in \mathcal{P}_0} \left(I\left(P = p\right) + I\left(Q = q \,|\, P = p\right)\right) + \inf_{r \in \mathcal{R}_0} I\left(R = r \,|\, Q = q, P \in \mathcal{P}_0\right) + c''' \\
&= \inf_{p \in \mathcal{P}_0} \left(0 + I\left(Q = q \,|\, P = p\right)\right) + \inf_{r \in \mathcal{R}_0} I\left(R = r \,|\, Q = q, P \in \mathcal{P}_0\right) + c''',
\end{aligned}$$

where c', c'', and c''' are real numbers, from the last term in equation (19), that are constant for all $q \in \mathcal{Q}$. By equations (21)-(22), there is a real number c such that, for all $q \in \mathcal{Q}$,

$$I\left(Q = q \,|\, R \in \mathcal{R}_0, P \in \mathcal{P}_0\right) = \inf_{p \in \mathcal{P}_0} \mathrm{rdndc}_p\left(q\right) + \inf_{r \in \mathcal{R}_0} \mathrm{rdndc}_q\left(r\right) + c\,.$$

□

In the simplest cases, $\mathcal{R}_0$ and $\mathcal{P}_0$ consist of one distribution each.

Example 6. Example 2, continued. Let p denote the posterior distribution of θ according to Bayes's theorem with the prior $\mathrm{N}\left(0, \left(1/8\right)^2\right)$. By Theorem 3, the blended posterior distribution is

$$\widetilde{q} \coloneqq \arg\inf_{q \in \mathcal{Q}} I\left(Q = q \,|\, R = \mathrm{N}\left(2,1\right), P = p\right) = \arg\inf_{q \in \mathcal{Q}} \mathrm{rdndc}_p\left(q\right) + \mathrm{rdndc}_q\left(\mathrm{N}\left(2,1\right)\right),$$

where $\mathcal{Q} = \left\{q_{\sigma_0} : \sigma_0 > 0\right\}$ is the set of posterior distributions of θ that Bayes's theorem induces from the priors in $\left\{\mathrm{N}\left(0, \sigma_0^2\right) : \sigma_0 > 0\right\}$. The function plotted in the left-hand side of Figure 2 is

$$\sigma_0 \mapsto I\left(Q = q_{\sigma_0} \,|\, R = \mathrm{N}\left(2,1\right), P = p\right),$$

the minimum of which is achieved at $\widetilde{\sigma}_0$, yielding $\widetilde{q} = q_{\widetilde{\sigma}_0}$. ▲

Example 6's use of the Bayesian posterior distribution from the prior $\mathrm{N}\left(0, \left(1/8\right)^2\right)$ as a source distribution and $\mathrm{N}\left(2,1\right)$ as a coding distribution is justified in Section 2.1, which conversely implies that, in the next example, $\mathrm{N}\left(2,1\right)$ is a source distribution and $\mathrm{N}\left(0, \left(1/8\right)^2\right)$ is a coding distribution. That switch explains the discrepancy seen in Figure 1.

Example 7. Example 3, continued. Let r denote the posterior distribution of θ according to Bayes's theorem with the prior $\mathrm{N}\left(0, \left(1/8\right)^2\right)$. By reasoning analogous to that of Example 6, the updated posterior distribution is

$$q^* \coloneqq \arg\inf_{q \in \mathcal{Q}} I\left(Q = q \,|\, R = r, P = \mathrm{N}\left(2,1\right)\right) = \arg\inf_{q \in \mathcal{Q}} \mathrm{rdndc}_{\mathrm{N}(2,1)}\left(q\right) + \mathrm{rdndc}_q\left(r\right),$$

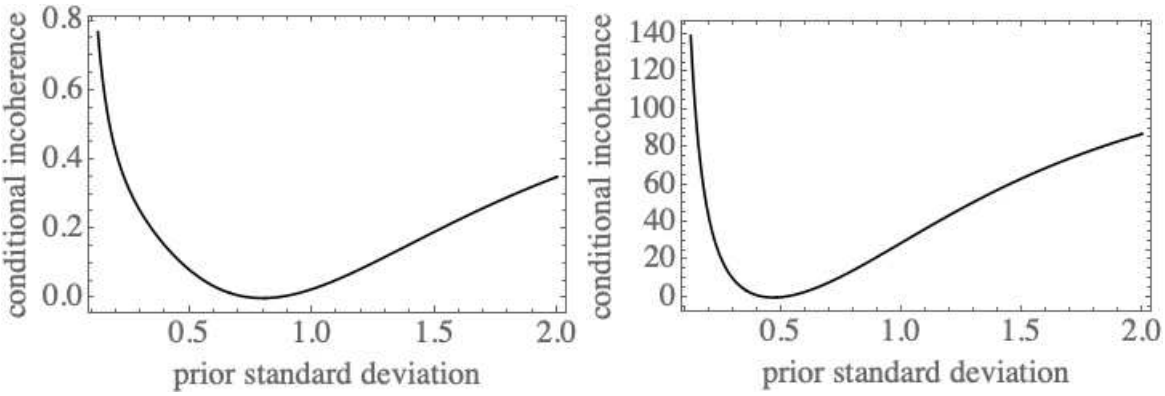

Figure 2: The conditional incoherence $I\left(Q = q_{\sigma_0} \,|\, R = r, P = p\right)$ as a function of σ_0, the standard deviation of θ under the prior. Left plot: p is the fully Bayesian posterior, r is the confidence distribution or objective Bayes posterior, and $I\left(Q = q_{\widetilde{\sigma}_0} \,|\, R = r, P = p\right) = 0$. Right plot: r is the fully Bayesian posterior, p is the objective Bayes posterior, and $I\left(Q = q_{\sigma_0^*} \,|\, R = r, P = p\right) = 0$. Note that $\sigma_0^* < \widetilde{\sigma}_0$, as claimed in Section 1.

where $\mathcal{Q} = \left\{q_{\sigma_0} : \sigma_0 > 0\right\}$ remains the same. The function plotted in the right-hand side of Figure 2 is

$$\sigma_0 \mapsto I\left(Q = q_{\sigma_0} \,|\, R = r, P = \mathrm{N}\left(2, 1\right)\right),$$

the minimum of which is achieved at σ_0^*, yielding $q^* = q_{\sigma_0^*}$. ▲

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009).

A Generalized maximum expected utility under indeterminate uncertainty

What if maximum expected utility or, equivalently, minimum expected loss, replaced minimum redundancy in the framework of Section 2? Consider a set $\mathcal{A}$ of possible actions and $\mathfrak{A}$, its power set. Let Π denote the min-plus probability distribution on $\mathfrak{A} \otimes \mathfrak{P}$, the power set of $\mathcal{A} \times \mathcal{P}$, and let $\Pi\left(\bullet|\bullet\right)$ denote the conditional

min-plus probability distribution on $(\mathfrak{A} \otimes \mathfrak{P}) \times (\mathfrak{A} \otimes \mathfrak{P})$ such that

$$\Pi(A = a \mid P = p) = E_p(\ell_a) - \inf_{a' \in \mathcal{A}} E_p(\ell_{a'}) \; \forall \, a \in \mathcal{A}, p \in \mathcal{P}, \tag{23}$$

where $(\theta, a) \mapsto \ell_a(\theta)$ is a loss function on $\Theta \times \mathcal{A}$.

Minimizing $a \mapsto \Pi(A = a \mid P = p)$ is more elegant than minimizing equation (10), at least in the eyes of those more familiar with expected loss than with redundancy. Indeed, the equivalent of Section 2.1's identification of estimates and coding distributions would not be needed were equation (23) adopted, for optimal estimates are already defined in decision theory as actions that minimize expected loss.

However, replacing redundancy with expected loss suffers from two drawbacks. First, when $\mathcal{A}$ consists of probability density functions on Θ, it does not lead to maximum entropy, even under $\theta \mapsto \ell_a(\theta) = -\log a(\theta)$, the proper scoring rule recommended by Bernardo (1979) for a probability density $a(\theta)$; contrast Theorem 2. That is unfortunate since maximum entropy is highly desirable for its unique satisfaction of important invariance properties (Shore and Johnson, 1980; Johnson and Shore, 1983; Csiszár, 1991; Paris, 1994). The equation behind this drawback is

$$\arg\inf_{p \in \mathcal{P}_0} \Pi(P = p \mid A \in \mathcal{A}_0) = \arg\inf_{p \in \mathcal{P}_0} \inf_{a \in \mathcal{A}_0} E_p\left(\log \frac{1}{a}\right), \tag{24}$$

where $\mathcal{A}_0$ is a set of probability density functions on Θ. Since the argument of the expectation function does not depend on p, equation (24) differs from equation (20). Equation (24) is an immediate consequence, given $\ell_a(\theta) = -\log a(\theta)$ for all $\theta \in \Theta$, of the following counterpart to Theorem 2.

Theorem 4. *For any $\mathcal{D}_0 \in \mathfrak{A} \otimes \mathfrak{P}$ and non-empty $\mathcal{D}_1 \in \mathfrak{A} \otimes \mathfrak{P}$,*

$$\Pi(\mathcal{D}_0 \mid \mathcal{D}_1) = \Pi(\mathcal{D}_0) + \Pi(\mathcal{D}_1 \mid \mathcal{D}_0) - \Pi(\mathcal{D}_1)$$

and, if $\Pi(P = \bullet) \equiv 0$, then, for any non-empty $\mathcal{P}_0 \in \mathfrak{P}$ and non-empty $\mathcal{A}_0 \in \mathfrak{A}$,

$$\arg\inf_{p \in \mathcal{P}_0} \Pi(P = p \mid A \in \mathcal{A}_0) = \arg\inf_{p \in \mathcal{P}_0} \inf_{a \in \mathcal{A}_0} E_p(\ell_a).$$

The proof is essentially the same as that of Theorem 2.

The second drawback is that equation (23) seems to lead to overly optimistic estimates in view of the this analog of Proposition 1.

Proposition 2. *Assume $p \mapsto \Pi(P = p)$ is constant, that is, $\Pi(P = \bullet) \equiv 0$. For any non-empty $\mathcal{P}_0 \in \mathfrak{P}$,*

$$\arg\inf_{a\in\mathcal{A}} \Pi(A = a \mid P \in \mathcal{P}_0) = \arg\inf_{a\in\mathcal{A}} \inf_{p\in\mathcal{P}_0} E_p(\ell_a). \tag{25}$$

Equation (25) takes the action that minimizes expected loss with respect to the distribution that minimizes expected loss. That is the same action produced by the Hurwicz criterion (Hurwicz, 1951) that has 0 as the parameter controlling the degree of pessimism. That *minimin criterion*, called the "maximax criterion" when replacing loss with utility, is typically dismissed for its extreme optimism (e.g., Perakis and Roels, 2008).

References

Aczél, J., Daróczy, Z., 1975. On Measures of Information and Their Characterizations. Mathematics in Science and Engineering. Elsevier Science.

Akian, M., Cohen, G., Gaubert, S., Quadrat, J., Viot, M., 1994. Max-plus algebra and applications to system theory and optimal control. In: Proceedings of the International Congress of Mathematicians. Citeseer, Birkhäuser, Zurich, Switzerland.

Bahamyirou, A., Marchand, T., 2015. On the discrepancy between bayes credibility and frequentist probability of coverage. Statistics and Probability Letters 97, 63–68.

Bernardo, J. M., 1979. Expected information as expected utility. The Annals of Statistics 7 (3), 686–690.

Bernardo, J. M., 2011. Integrated objective bayesian estimation and hypothesis testing. Bayesian statistics 9, 1–68.

Bickel, D. R., 2012a. Coherent frequentism: A decision theory based on confidence sets. Communications in Statistics - Theory and Methods 41, 1478–1496.

Bickel, D. R., 2012b. A frequentist framework of inductive reasoning. Sankhya A 74, 141–169.

Bickel, D. R., 2015. Blending Bayesian and frequentist methods according to the precision of prior information with applications to hypothesis testing. Statistical Methods & Applications 24, 523–546.

Bickel, D. R., 2018a. A note on fiducial model averaging as an alternative to checking Bayesian and frequentist models. Communications in Statistics - Theory and Methods 47, 3125–3137.

Bickel, D. R., 2018b. Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: A large-deviation approach. Statistics 52, 552–570.

Cover, T., Thomas, J., 2006. Elements of Information Theory. John Wiley & Sons, New York.

Cox, R. T., 1961. The Algebra of Probable Inference. Johns Hopkins Press, Baltimore.

Csiszár, I., 1985. An extended maximum entropy principle and a Bayesian justification. In: Bernardo, J., DeGroot, M., Lindley, D. V., Smith, A. (Eds.), Bayesian Statistics 2. Elsevier Inc., Amsterdam, pp. 83–98.

Csiszár, I., 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. 19, 2032–2066.

De Baets, B., Tsiporkova, E., Mesiar, R., 1999. Conditioning in possibility theory with strict order norms. Fuzzy Sets and Systems 106, 221–229.

Efron, B., 1993. Bayes and likelihood calculations from confidence intervals. Biometrika 80, 3–26.

Evans, M., 2015. Measuring Statistical Evidence Using Relative Belief. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, New York.

Evans, M., Jang, G., 2011. A limit result for the prior predictive applied to checking for prior-data conflict. Statistics and Probability Letters 81 (8), 1034–1038.

Fitzpatrick, B. G., 2013. Max plus decision processes in planning problems for unmanned air vehicle teams. In: Recent Advances in Research on Unmanned Aerial Vehicles. Springer, pp. 31–45.

Fleming, W., Kaise, H., Sheu, S.-J., 2010. Max-plus stochastic control and risk-sensitivity. Applied Mathematics and Optimization 62 (1), 81–144.

Fraser, D. A. S., 2011. Is Bayes posterior just quick and dirty confidence? Statistical Science 26, 299–316.

Grundy, P. M., 1956. Fiducial distributions and prior distributions: An example in which the former cannot be associated with the latter. Journal of the Royal Statistical Society, Series B 18, 217–221.

Grünwald, P., Dawid, A. P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics 32, 1367–1433.

Grünwald, P. D., 2007. The Minimum Description Length Principle. MIT Press, London.

Hannig, J., Iyer, H., Lai, R. C., Lee, T. C., 2016. Generalized fiducial inference: A review and new results. Journal of the American Statistical Association 111 (515), 1346–1361.

Held, L., Ott, M., 2018. On p-values and Bayes factors. Annual Review of Statistics and Its Application 5, 393–419.

Hurwicz, L., 1951. The generalized Bayes-minimax principle: a criterion for decision-making under uncertainty. Cowles Commission Discussion Paper 355.

Johnson, R., Shore, J., 1983. Axiomatic derivation of the principle of maximum-entropy and the principle of minimum cross-entropy - comments and correction. IEEE Transactions on Information Theory 29 (6), 942–943.

Knight, F., 2012. Risk, Uncertainty and Profit. Dover Publications.

Lapointe, S., Bobée, B., 2000. Revision of possibility distributions: A Bayesian inference pattern. Fuzzy Sets and Systems 116 (2), 119 – 140.

Lindley, D. V., 1958. Fiducial distributions and Bayes' theorem. Journal of the Royal Statistical Society B 20, 102–107.

Martin, R., Liu, C., 2013. Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference. Journal of the American Statistical Association 108 (501), 301–313.

Mayo, D., 2018. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press, Cambridge.

Paris, J. B., 1994. The Uncertain Reasoner's Companion: A Mathematical Perspective. Cambridge University Press, New York.

Perakis, G., Roels, G., 2008. Regret in the newsvendor model with partial information. Operations Research 56 (1), 188–203.

Picard, J. (Ed.), 2004. learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI - 2001. Lecture notes in mathematics. Springer.

Polansky, A. M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

Puhalskii, A., 2001. Large Deviations and Idempotent Probability. Monographs and Surveys in Pure and Applied Mathematics. CRC Press, New York.

Quadrat, J., 1998. Min-plus probability calculus. Algebre Max-Plus et applications en informatique et automatique, 393–411.

Quadrat, J.-P., 1995. Max-plus algebra and applications to system theory and optimal control. In: Proceedings of the International Congress of Mathematicians. Springer, pp. 1511–1522.

Samaniego, F. J., 2010. A Comparison of the Bayesian and Frequentist Approaches to Estimation (Springer Series in Statistics). Springer, New York.

Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Sellke, T., Bayarri, M. J., Berger, J. O., 2001. Calibration of p values for testing precise null hypotheses. American Statistician 55, 62–71.

Shore, J. E., Johnson, R. W., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory IT-26, 26–37.

Singh, K., Xie, M., Strawderman, W. E., 2007. Confidence distribution (CD) – distribution estimator of a parameter. IMS Lecture Notes Monograph Series 2007 54, 132–150.

Spohn, W., 2012. The Laws of Belief: Ranking Theory and Its Philosophical Applications. Oxford University Press.

Topsøe, F., 1979. Information theoretical optimization techniques. Kybernetika 15 (1), 8–27.

Veronese, P., Melilli, E., 2018a. Fiducial, confidence and objective bayesian posterior distributions for a multidimensional parameter. Journal of Statistical Planning and Inference 195, 153–173.

Veronese, P., Melilli, E., 2018b. Some asymptotic results for fiducial and confidence distributions. Statistics & Probability Letters 134, 98–105.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London.

Walter, G., Augustin, T., 2009. Imprecision and prior-data conflict in generalized bayesian inference. Journal of Statistical Theory and Practice 3 (1), 255–271.

Wilkinson, G. N., 1977. On resolving the controversy in statistical inference (with discussion). Journal of the Royal Statistical Society B 39, 119–171.

Xie, M.-G., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review 81 (1), 3–39.