Electronic Journal of Statistics
Vol. 6 (2012) 686–709
ISSN: 1935-7524
DOI: 10.1214/12-EJS689

Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes*

David R. Bickel†

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

Abstract: In statistical practice, whether a Bayesian or frequentist approach is used in inference depends not only on the availability of prior information but also on the attitude taken toward partial prior information, with frequentists tending to be more cautious than Bayesians. The proposed framework defines that attitude in terms of a specified amount of caution, thereby enabling data analysis at the level of caution desired and on the basis of prior information. The caution parameter represents the attitude toward partial prior information in much the same way as a loss function represents the attitude toward risk. When there is very little prior information and nonzero caution, the resulting inferences correspond to those of the candidate confidence intervals and p-values that are most similar to the credible intervals and hypothesis probabilities of the specified Bayesian posterior. On the other hand, in the presence of a known physical distribution of the parameter, inferences are based only on the corresponding physical posterior. In those extremes of either negligible prior information or complete prior information, inferences do not depend on the degree of caution. Partial prior information between those two extremes leads to intermediate inferences that are more frequentist to the extent that the caution is high and more Bayesian to the extent that the caution is low.

AMS 2000 subject classifications: Primary 62A01; secondary 62A99.

Keywords and phrases: Ambiguity, blended inference, conditional Gamma-minimax, confidence distribution, confidence posterior, Ellsberg paradox, imprecise probability, maximum entropy, maxmin expected utility, minimum cross entropy, minimum divergence, minimum information for discrimination, minimum relative entropy, observed confidence level, robust Bayesian analysis.

Received September 2011.

*This is an original research paper.

†The author is grateful to an anonymous reviewer for insightful comments that led to many improvements in presentation and to the addition of Example 2. This research was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

1. Introduction

Since traditional arguments for the rationality of Bayesian methods and the objectivity of frequentist methods have failed to achieve a consensus among statisticians, Samaniego (2010) and others seek to determine under what circumstances one approach performs better than another according to neutral decision-theoretic criteria. However, the controversy between Bayesianism and frequentism may be irresolvable inasmuch as it reflects honest differences in personal attitudes of statisticians rather than differences in their rationality or knowledge of performance comparisons. Efron (2005) argued,

> The Bayesian-frequentist debate reflects two different attitudes about the process of doing science, both quite legitimate. Bayesian statistics is well suited to individual researchers, or a research group, trying to use all of the information at its disposal to make the quickest possible progress. In pursuing progress, Bayesians tend to be aggressive and optimistic with their modeling assumptions. Frequentist statisticians are more cautious and defensive. One definition says that a frequentist is a Bayesian trying to do well, or at least not too badly, against any possible prior distribution. The frequentist aims for universally acceptable conclusions, ones that will stand up to adversarial scrutiny.

On one hand, methodology reflecting extreme caution in the form of the minimax-like attitude attributed to frequentists and, on the other hand, methodology reflecting the extreme reliance on modeling assumptions attributed to Bayesians both play useful roles in statistical inference. Bayesian methods can outperform frequentist methods to the extent that the priors are close to the truth, but call for caution since priors far from the truth can lead to severe bias. Building on that premise, the idea motivating this paper is that methodology for moderate amounts of caution also has a place in practical data analysis. The extent of such caution will be formally defined in order to facilitate making statistical inferences at the level of caution appropriate to the situation.

The mathematical definition will build on previous work to formalize caution in the face of uncertainty. Attitudes toward uncertainty have long been mathematically modeled in the economics literature. Ellsberg (1961) identified two distinct types of uncertainty: *risk* is the variability in an unknown quantity that threatens assets, whereas *ambiguity* is ignorance about the extent of such variability. The same agent may be much more cautious toward risk than toward ambiguity or vice versa. A utility or loss function can model an agent's attitude toward risk but not its attitude toward ambiguity. Because frequentist actions can differ from Bayesian actions given the same loss function, the attitude toward risk is not relevant to the problem of representing and balancing the two basic approaches to statistical inference. The attitude toward ambiguity is much more pertinent to the concept of caution toward relying on a prior distribution.

Ellsberg (1961) distinguished "pessimism" from "conservatism": the former is an excessive belief that worst-case scenarios will materialize, whereas the latter only involves cautiously acting as if they will. In other words, the attitude of "hoping for the best, preparing for the worst" is consistent with conservatism but not pessimism. While that attitude does motivate much of frequentist statistics, "conservatism" already has technical meanings in the statistics literature, e.g.,

TABLE 1
Utility function for actions I-II and the three possible states of nature

	Red drawn	Black drawn	Yellow drawn
action I	$100	$0	$0
action II	$0	$100	$0

TABLE 2
Utility function for actions III-IV and the three possible states of nature

	Red drawn	Black drawn	Yellow drawn
action III	$100	$0	$100
action IV	$0	$100	$100

conservative confidence intervals have higher-than-nominal coverage rates. For that reason, the term "caution" will be used when assigning an operational definition to the degree of conservatism toward ambiguity in the sense of Ellsberg (1961).

Example 1. *Ellsberg Paradox.* In an example from Ellsberg (1961), a ball is randomly drawn from an urn of 90 balls, each of one of three possible colors: red, black, and yellow. Nothing is known about the distributions of the balls in the urn except that exactly 30 are red. Thus, there is ambiguity in the distribution of black and yellow balls. The agent would gain a reward of $0 or $100 based on its taking action I or action II according to utility function displayed as Table 1 in setting 1. In setting 2, the agent would instead gain $0 or $100 based on its taking action III or action IV according to the utility function displayed as Table 2. Agents cautious toward ambiguity would choose action I over action II in setting 1 but would take action IV over action III in setting 2, against subjective Bayesian concepts of coherence but without requiring the extreme caution of a minimax strategy (Ellsberg, 1961).

In the absence of ambiguity, the axiomatic system of von Neumann and Morgenstern (1953, §3.6) and later generalizations prescribe choosing the action that maximizes expected utility. By forcefully applying such a system to conditional expectations given observed data, Savage (1954) revitalized Bayesian statistics. The action that maximizes expected utility with respect to a Bayesian posterior is called the *posterior Bayes action*. Ambiguity about the posterior is usually modeled in terms of a set $\dot{\mathcal{P}}$ of multiple posteriors in place of a single posterior. A multiplicity of posteriors may arise from insufficient elicitation of subjective prior opinions (e.g., Berger, Insua and Ruggeri, 2000), from a spread in a gamble's buying and selling prices (e.g., Walley, 1991), or, more objectively, from ignorance as to which prior distribution in a set describes the physical variability of a parameter. The last source accords best with the notion of ambiguity as used in Ellsberg (1961), Jaffray (1989b), Jaffray (1989a), and Gajdos, Tallon and Vergnaud (2004).

In the Bayesian statistics literature, the most studied decision-theoretic approach for sets of priors is the (marginal) Γ-minimax strategy (e.g., Berger, Insua and Ruggeri, 2000), which formulates the problem in terms of minimax risk in

the frequentist sense of Wald (1961). The closely related conditional Γ-minimax strategy (e.g., Betrò and Ruggeri, 1992) takes the action that minimizes the expected loss maximized over all of the posterior distributions in a set $\dot{\mathcal{P}}$, each member of which corresponds to a prior distribution in a set traditionally denoted by Γ. That statistical strategy is a special case of the maxmin expected utility strategy (Hurwicz, 1951b; Gilboa and Schmeidler, 1989), which takes the action that maximizes the expected utility minimized over a set of distributions. Both "robust Bayesian" strategies are reviewed in Vidakovic (2000).

The following equation extends the conditional Γ-minimax strategy to the problem of conducting statistical inference at a specified degree of *caution* κ and with respect to a Bayesian posterior $\dot{P} \in \dot{\mathcal{P}}$ that is not generally the true physical distribution of the parameter θ. For any $\kappa \in [0,1]$, the *κ-conditional-Γ (κCG) action* is defined as

$$\dot{a}_\kappa = \arg\inf_{a\in\mathcal{A}} \left(\kappa \sup_{P'\in\dot{\mathcal{P}}} \int L(\theta,a)\, dP'(\theta) + (1-\kappa) \int L(\theta,a)\, d\dot{P}(\theta) \right), \qquad (1)$$

with the conventions that $\kappa \times \infty = 0$ if $\kappa = 0$ and $(1-\kappa)\times\infty = 0$ if $\kappa = 1$. The κCG action reduces to the conditional Γ-minimax action under complete caution ($\kappa = 1$) and to the posterior Bayes action in the complete absence of caution ($\kappa = 0$). For discrete θ, this κ is isomorphic to quantities used by Ellsberg (1961), Gajdos, Tallon and Vergnaud (2004), and Tapking (2004) and is similar in spirit to the quantity in Hurwicz (1951a) and Jaffray (1989b) that Augustin (2002) calls "caution." Gärdenfors and Sahlin (1982) and Gajdos, Tallon and Vergnaud (2004) stressed the equivalent of the rearrangement of equation (1) as

$$\dot{a}_\kappa = \arg\inf_{a\in\mathcal{A}} \left(\sup_{P\in\left\{\kappa P' + (1-\kappa)\dot{P} : P'\in\dot{\mathcal{P}}\right\}} \int L(\theta,a)\, dP(\theta) \right). \qquad (2)$$

The κCG strategy has two drawbacks that will prevent its use in many applications. First, under standard loss functions, the conditional Γ-minimax (1CG) strategy requires either that $\dot{\mathcal{P}}$ impose strict bounds on the parameter space (Bayati Eshkaftaki and Parsian, 2011) or that $\mathcal{A}$ be severely restricted (Betrò and Ruggeri, 1992), and the κCG strategy with $0 < \kappa < 1$ has the same limitation. Second, since the 1CG strategy is not necessarily a frequentist procedure, the κCG framework does not fulfill the above goal of formulating procedures that reduce to frequentist procedures given complete caution.

Following the preliminary notation and definitions of Section 2, an information-theoretic framework will be introduced in Section 3 to overcome the identified limitations of the κCG framework. Simple examples demonstrating the wide applicability of the information-theoretic framework will appear in Section 4. Extensions are described in Section 5 in terms of exchanging roles of frequentist and Bayesian procedures as appropriate for particular applications. Section 6 closes the paper with a brief discussion.

2. Bayesian and frequentist posterior distributions

2.1. Preliminary concepts

The observed data vector $x \in \mathcal{X}$ is modeled as a realization of a random variable X of probability space $(\mathcal{X}, \mathfrak{X}, P_{\theta_*,\lambda_*})$, which for some parameter set $\Theta_* \times \Lambda_*$ is indexed by an *interest parameter* $\theta_* \in \Theta_*$ and potentially also by a *nuisance parameter* $\lambda_* \in \Lambda_*$. The family $\{P_{\theta_*,\lambda_*} : \theta_* \in \Theta_*, \lambda_* \in \Lambda_*\}$ will be called the *sampling model* for x.

Inferences will be made about the *focus parameter* $\theta = \theta(\theta_*)$, a subparameter of the interest parameter, in a set Θ. In the simplest case, $\theta = \theta_*$ and $\Theta = \Theta_*$, but there are many other possibilities. For example, when testing the null hypothesis that $\theta_* = 0$ against the alternative hypothesis that $\theta_* \neq 0$ for $\Theta_* = \mathbb{R}$, it is convenient to define the focus parameter by $\theta = 0$ if $\theta_* = 0$ and $\theta = 1$ if $\theta_* \neq 0$, in which case $\Theta = \{0, 1\}$. Let $\mathcal{H}$ denote a σ-field that allows any physically meaningful hypothesis about θ to be expressed as "θ is in $\Theta^\dagger$," where $\Theta^\dagger \in \mathcal{H}$.

2.2. Bayesian posteriors

In the Bayesian setting, the above sampling model is understood as conditional on the parameter values with respect to some prior distribution. For notational simplicity, the distributions written in this subsection are marginal over the nuisance parameter, i.e., λ_* is eliminated by integrating with respect to some prior.

Let $\mathcal{P}$ denote the set of all probability distributions on $(\Theta, \mathcal{H})$. Before observing data, knowledge about the observable vector and focus parameter is represented by $\mathcal{P}^{\text{plaus}}$, a set of plausible joint distributions on $(\mathcal{X} \times \Theta, \mathfrak{X} \otimes \mathcal{H})$. Accordingly, every member of $\mathcal{P}^{\text{plaus}}$ is called a *plausible joint distribution*. The integral of every plausible joint distribution over the data is a marginal distribution on $(\Theta, \mathcal{H})$ and is called a *plausible prior distribution*.

The Bayesian approach yields inferences about the focus parameter on the basis of a single distribution, $\dot{P}^{\text{plaus}} \in \mathcal{P}^{\text{plaus}}$. If $\langle \dot{X}, \dot{\theta} \rangle \sim \dot{P}^{\text{plaus}}$, then the *working prior* distribution is the marginal distribution of $\dot{\theta}$. It follows that the working prior is one of the plausible priors.

The *working Bayesian posterior* $\dot{P}$ and the *knowledge base* $\dot{\mathcal{P}}$ (Topsøe, 2004) are defined such that

$$\dot{P} = \dot{P}^{\text{plaus}}\left(\bullet | \dot{X} = x\right); \tag{3}$$

$$\dot{\mathcal{P}} = \left\{P'\left(\bullet | \dot{X} = x\right) : P' \in \mathcal{P}^{\text{plaus}}\right\}. \tag{4}$$

$\dot{P}$ is simply the Bayesian posterior distribution corresponding to the working prior, and $\dot{\mathcal{P}}$ is likewise the set of Bayesian posteriors in $\mathcal{P}$ that correspond to plausible prior distributions. To prevent confusion with $\dot{P}$, members of $\dot{\mathcal{P}}$ will be referred to as *plausible posteriors* since they are the parameter distributions

consistent with the mathematical representation either of a physical system or of a belief system (cf. Topsøe, 1979, 2004). Thus, the posterior that would be used in purely Bayesian inference is one of the plausible posteriors ($\dot{P} \in \dot{\mathcal{P}}$).

As equation (3) indicates, the working prior is updated to its posterior just as if it were a physical distribution. For that reason, many authors have criticized pure Bayesianism for treating mental probabilities exactly as if they were limiting relative frequencies of events in a realistic model of the external world (e.g., Kardaun et al., 2003; Fraser, 2011; Bickel, 2011d). However, equation (3) results from the same information-theoretic framework of constrained inference as the moderate-posterior approach to be introduced in Section 3. In particular, $\dot{P}^{\text{plaus}}(\bullet|\dot{X} = x)$ is the distribution that minimizes the information divergence with respect to $\dot{P}^{\text{plaus}}(\bullet)$ under the sole constraint that $\dot{X} = x$ (Harremoës, 2007, Example 3), as Williams (1980) proved in the discrete case. That fact leads to $\dot{P}^{\text{plaus}}(\bullet|\dot{X} = x)$ as the solution to a minimax problem under broad conditions (Topsøe, 2007). See also Csiszár (1985).

2.3. Confidence posteriors

The sampling model of Section 2.1 admits not only system constraints and Bayesian inference (§2.2) but also frequentist inference in the form of confidence intervals and p-values. Let $\mathcal{H}_*$ denote $\mathcal{B}(\Theta_*)$, the Borel set of Θ_*. Given some $\alpha \in [0, 1]$, if the function $\widehat{\Theta}(1-\alpha, \bullet) : \mathcal{X} \to \mathcal{H}_*$ satisfies

$$P_{\theta_*,\lambda_*}\left(\theta_* \in \widehat{\Theta}(1-\alpha; X)\right) = 1 - \alpha \tag{5}$$

for all $\theta_* \in \Theta_*$ and $\lambda_* \in \Lambda_*$, then $\widehat{\Theta}(1-\alpha; X)$ is called a $100(1-\alpha)\,\%$ *confidence set* for θ_*. Suppose $\widehat{\Theta} : [0, 1] \times \mathcal{X} \to \Theta_*$ defines nested confidence sets in the sense that $\widehat{\Theta}(1-\alpha; X)$ is a $100(1-\alpha)\,\%$ confidence set for θ_* given any confidence level $1 - \alpha \in [0, 1]$ and such that

$$0 \leq 1 - \alpha_1 < 1 - \alpha_2 \leq 1 \implies \widehat{\Theta}(1-\alpha_1; X) \subset \widehat{\Theta}(1-\alpha_2; X)$$

almost surely. A *confidence posterior distribution* is a distribution $\ddot{P}_*$ on $(\Theta_*, \mathcal{H}_*)$ for which

$$\ddot{P}_*\left(\Theta^\dagger\right) = \ddot{P}_*\left(\ddot{\theta}_* \in \Theta^\dagger\right) \in \left\{1 - \alpha : \widehat{\Theta}(1-\alpha; x) = \Theta^\dagger, 0 \leq \alpha \leq 1\right\} \tag{6}$$

for all $x \in \mathcal{X}$ and $\Theta^\dagger \in \mathcal{H}_*$, where $\ddot{\theta}_*$ is a random variable of distribution $\ddot{P}_*$. Polansky (2007) called $\ddot{P}_*\left(\Theta^\dagger\right)$ the *observed confidence level* of the hypothesis that $\theta_* \in \Theta^\dagger$. Confidence posteriors for which θ_* is a real scalar ($\Theta_* \subseteq \mathbb{R}$) and the σ-field is Borel ($\mathcal{H}_* = \mathcal{B}(\Theta_*)$) are usually called *confidence distributions*, each of which encodes confidence intervals of all confidence levels and hypothesis tests of all simple null hypotheses (Efron, 1993). Various devices extend confidence posteriors to cases in which their posterior probabilities only approximately match confidence levels (Schweder and Hjort, 2002; Singh, Xie and Strawderman, 2005; Polansky, 2007; Bickel, 2012a).

The identity between confidence posterior probabilities and levels of confidence (6) clears up the misunderstanding that confidence levels and p-values cannot be interpreted as epistemological probabilities of hypotheses given the observed data. In fact, since $\ddot{P}_*$ is a Kolmogorov probability measure on parameter space, decisions made using various loss functions by the confidence posterior action

$$\ddot{a} = \arg\inf_{a \in \mathcal{A}} \int L(\theta_*, a)\, d\ddot{P}_*(\theta_*)$$

for each loss function L are coherent with each other in the senses usually associated with Bayesian inference, whether or not $\ddot{P}_*$ can be derived from some prior via Bayes's theorem (Bickel, 2011b, 2012a).

Let $\ddot{\mathcal{P}}_*$ denote the set of confidence posteriors on $(\Theta_*, \mathcal{H}_*)$ that are under consideration. For example, $\ddot{\mathcal{P}}_*$ could be the set of a single confidence posterior, the set of all distributions on $(\Theta_*, \mathcal{H}_*)$ that satisfy equation (6), or, as in Bickel (2012a), the set of two approximate confidence posteriors or the convex set of all mixtures of the two.

The set $\ddot{\mathcal{P}}$ will represent the set of distributions of $\theta(\ddot{\theta}_*)$ for all $\ddot{P}_* \in \ddot{\mathcal{P}}_*$:

$$\ddot{\mathcal{P}} = \left\{ P'' \in \mathcal{P} : \theta\left(\ddot{\theta}_*\right) \sim P'', \ddot{\theta}_* \sim \ddot{P}_* \in \ddot{\mathcal{P}}_* \right\}.$$

Thus, for any $\ddot{P}_* \in \ddot{\mathcal{P}}_*$, there is a random parameter $\ddot{\theta} = \theta(\ddot{\theta}_*)$ of distribution $\ddot{P} \in \ddot{\mathcal{P}}$ such that $\ddot{P}(\ddot{\theta} \in \{\theta(\theta''_*) : \theta''_* \in \Theta^\dagger_*\}) = \ddot{P}_*(\ddot{\theta}_* \in \Theta^\dagger_*)$ for all $\Theta^\dagger_* \in \mathcal{H}_*$. $\ddot{\mathcal{P}}$ will be considered as a set of confidence posterior distributions of the focus parameter even though more literally they are not necessarily confidence posteriors but rather fiducial-like distributions derived from the set $\ddot{\mathcal{P}}_*$ of confidence posteriors by the laws of probability. (Hannig (2009) provides a recent review of fiducial inference.) In the simplest case of $\ddot{\theta} = \ddot{\theta}_*$, $(\Theta, \mathcal{H}) = (\Theta_*, \mathcal{H}_*)$ and $\ddot{\mathcal{P}} = \ddot{\mathcal{P}}_*$. While confidence distributions are used here to represent frequentism in the form of hypothesis tests and confidence intervals, $\ddot{\mathcal{P}}$ can be a set of any distributions on $(\Theta, \mathcal{H})$ to use as benchmarks with respect to which the posterior introduced in the next section is intended as an improvement.

3. Framework of moderate inference

3.1. Moderate posteriors

Let P and Q denote probability distributions on $(\Theta, \mathcal{H})$. The *information divergence of P with respect to Q* is defined as

$$I(P||Q) = \int dP \log\left(\frac{dP}{dQ}\right) \tag{7}$$

if Q is absolutely continuous with respect to P and $I(P||Q) = \infty$ if not, where $0 \log(0) = 0$ and $0 \log(0/0) = 0$. $I(P||Q)$ goes by many names in literature, including "Kullback-Leibler information" and "cross entropy." Viewing $I(P||Q)$

as information leads to the concept of how much information for statistical inference would be gained by replacing a confidence posterior $P'' \in \ddot{\mathcal{P}}$ with another posterior $Q \in \mathcal{P}$ if the plausible posterior $P' \in \dot{\mathcal{P}}$ were the physical distribution of the parameter θ. Specifically,

$$I\left(P' || P'' \rightsquigarrow Q\right) = I\left(P' || P''\right) - I\left(P' || Q\right),$$

as a special case of "information gain" (Pfaffelhuber, 1977), is called the *inferential gain of Q relative to P'' given P'* (Bickel, 2011a). (The $\rightsquigarrow$ notation is borrowed from Topsøe (2007) to express the inferential gain as a function of three probability distributions while conveniently recalling their unique contributions.)

In analogy with equation (2), the *caution* $\kappa \in [0,1]$ is then the extent to which a "worst-case" plausible posterior $P' \in \dot{\mathcal{P}}$ is used for inference as opposed to the working Bayesian posterior $\dot{P}$ in this definition of the *κ-inferential gain of Q relative to P'' given P' and $\dot{P}$*:

$$I\left(P', \dot{P} || P'' \rightsquigarrow Q; \kappa\right) = I\left(\kappa P' + (1-\kappa)\,\dot{P} || P'' \rightsquigarrow Q\right). \tag{8}$$

The posterior distribution that has the highest κ-inferential gain in the following sense will be used for making inferences and decisions. The *moderate posterior distribution with caution κ relative to $\ddot{\mathcal{P}}$ given $\dot{\mathcal{P}}$ and $\dot{P}$* is denoted by $\widetilde{P}_\kappa$ and defined by

$$\inf_{P'' \in \ddot{\mathcal{P}}} \inf_{P' \in \dot{\mathcal{P}}} I\left(P', \dot{P} || P'' \rightsquigarrow \widetilde{P}_\kappa; \kappa\right) = \inf_{P'' \in \ddot{\mathcal{P}}} \sup_{Q \in \mathcal{P}} \inf_{P' \in \dot{\mathcal{P}}} I\left(P', \dot{P} || P'' \rightsquigarrow Q; \kappa\right). \tag{9}$$

Less technically, $\widetilde{P}_\kappa$ is the posterior distribution that maximizes the worst-case inferential gain relative to the confidence posterior P'', which is in turn chosen to minimize the maximum worst-case gain. In the case that equation (9) does not have a unique solution, the moderate posterior is defined to be as close as possible to the working Bayesian posterior:

$$\widetilde{P}_\kappa = \arg \inf_{P''' \in \widetilde{\mathcal{P}}_\kappa} I\left(\dot{P} || P'''\right), \tag{10}$$

where the set $\widetilde{\mathcal{P}}_\kappa$ of *candidate moderate posteriors* is defined as the set of all distributions in $\mathcal{P}$ such that every member of $\widetilde{\mathcal{P}}_\kappa$ solves equation (9). By letting

$$J\left(\dot{P} || P'' \rightsquigarrow Q; \kappa\right) = \inf_{P' \in \dot{\mathcal{P}}} I\left(P', \dot{P} || P'' \rightsquigarrow Q; \kappa\right) \tag{11}$$

for any $P'' \in \ddot{\mathcal{P}}$, that set may be written as

$$\widetilde{\mathcal{P}}_\kappa = \left\{ P \in \mathcal{P} : \inf_{P'' \in \ddot{\mathcal{P}}} J\left(\dot{P} || P'' \rightsquigarrow P; \kappa\right) = \inf_{P'' \in \ddot{\mathcal{P}}} \sup_{Q \in \mathcal{P}} J\left(\dot{P} || P'' \rightsquigarrow Q; \kappa\right) \right\}. \tag{12}$$

The *moderate posterior action with caution* κ is

$$\widetilde{a}_\kappa = \arg\inf_{a\in\mathcal{A}} \int L(\theta, a)\, d\widetilde{P}_\kappa(\theta), \tag{13}$$

which defines making decisions on the basis of the moderate posterior as taking actions that minimize its expected loss. For example, if $\ddot{P}$ is the only confidence posterior under consideration, then $\ddot{\mathcal{P}} = \{\ddot{P}\}$ and

$$\widetilde{P}_\kappa = \arg\sup_{Q\in\mathcal{P}} \left(\inf_{P\in\dot{\mathcal{P}}_\kappa} I\left(P||\ddot{P} \rightsquigarrow Q\right)\right); \tag{14}$$

$$\dot{\mathcal{P}}_\kappa = \left\{\kappa P' + (1-\kappa)\,\dot{P} : P' \in \dot{\mathcal{P}}\right\}, \tag{15}$$

which recalls equation (2). Since $\dot{P} \in \dot{\mathcal{P}}_\kappa \subseteq \dot{\mathcal{P}}$, $\dot{\mathcal{P}}_0 = \{\dot{P}\}$, and $\dot{\mathcal{P}}_1 = \dot{\mathcal{P}}$, the effect of $\kappa < 1$ as opposed to $\kappa = 1$ is to replace the knowledge base $\dot{\mathcal{P}}$ with a subset $\dot{\mathcal{P}}_\kappa$ containing the working Bayesian posterior $\dot{P}$ (cf. Gajdos, Tallon and Vergnaud, 2004).

The two extreme cases of caution reduce decision making to previous frameworks. A complete lack of caution ($\kappa = 0$) leads to the sole use of the working Bayesian posterior for the minimization of posterior expected loss: $\widetilde{P}_0 = \dot{P}$. On the other hand, complete caution ($\kappa = 1$) leads to ignoring the working Bayesian posterior and, in the case of a single confidence posterior, to the framework of Bickel (2011a), in which $\widetilde{P}_1$ is called the *blended posterior*.

3.2. Minimum information divergence

The following fact is from Topsøe (1979); see also Pfaffelhuber (1977), Harremoës (2007), Bickel (2011a), and especially Topsøe (2007).

Lemma 1. *Given a distribution* $\ddot{P}$ *on* $(\Theta, \mathcal{H})$*, if* $\dot{\mathcal{P}}$ *is convex and* $I(P||\ddot{P}) < \infty$ *for all* $P \in \dot{\mathcal{P}}$*, then*

$$\sup_{Q\in\mathcal{P}} \inf_{P'\in\dot{\mathcal{P}}} I\left(P'||\ddot{P} \rightsquigarrow Q\right) = \inf_{Q\in\dot{\mathcal{P}}} I\left(Q||\ddot{P}\right).$$

The resulting theorem is useful for finding $\widetilde{P}_\kappa$:

Theorem 1. *For any* $\kappa \in [0,1]$*, if* $\dot{\mathcal{P}}$ *is convex and* $I(P'||P'') < \infty$ *for all* $P' \in \dot{\mathcal{P}}$ *and* $P'' \in \ddot{\mathcal{P}}$*, then the moderate posterior* $\widetilde{P}_\kappa$ *is given by equations* (10) *and* (15) *with*

$$\widetilde{\mathcal{P}}_\kappa = \left\{P \in \mathcal{P} : \inf_{P''\in\ddot{\mathcal{P}}} I(P||P'') = \inf_{P''\in\ddot{\mathcal{P}}} \inf_{Q\in\dot{\mathcal{P}}_\kappa} I(Q||P'')\right\}. \tag{16}$$

Proof. For any $\kappa \in [0,1]$ and $P'' \in \ddot{\mathcal{P}}$, equations (11) and (14), with Lemma 1, imply

$$\begin{aligned}\sup_{Q\in\mathcal{P}} J\left(\dot{P}||P'' \rightsquigarrow Q; \kappa\right) &= \sup_{Q\in\mathcal{P}} \inf_{P\in\dot{\mathcal{P}}_\kappa} I(P||P'' \rightsquigarrow Q), \\ &= \inf_{Q\in\dot{\mathcal{P}}_\kappa} I(Q||P'')\end{aligned}$$

and thus

$$\inf_{P''\in\ddot{\mathcal{P}}}\sup_{Q\in\mathcal{P}} J\left(\dot{P}||P''\rightsquigarrow Q;\kappa\right)=\inf_{P''\in\ddot{\mathcal{P}}}\inf_{Q\in\dot{\mathcal{P}}_\kappa} I\left(Q||P''\right).$$

Equation (12) thereby reduces to equation (16). □

An immediate consequence is

Corollary 1. *Given $\ddot{\mathcal{P}}=\{\ddot{P}\}$, if $\dot{\mathcal{P}}$ is convex and $I(P'||\ddot{P})<\infty$ for all $P'\in\dot{\mathcal{P}}$, then the moderate posterior is*

$$\widetilde{P}_\kappa=\arg\inf_{Q\in\dot{\mathcal{P}}_\kappa} I\left(Q||\ddot{P}\right). \tag{17}$$

Inference is also simplified when at least one of the confidence posteriors is sufficiently close to the working Bayesian posterior:

Corollary 2. *If $\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa$ is nonempty, $\dot{\mathcal{P}}$ is convex, and $I\left(P'||P''\right)<\infty$ for all $P'\in\dot{\mathcal{P}}$ and $P''\in\ddot{\mathcal{P}}$, then the moderate posterior is the confidence posterior that is closest to the working Bayesian posterior in the sense that*

$$\widetilde{P}_\kappa=\arg\inf_{P''\in\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa} I\left(\dot{P}||P''\right). \tag{18}$$

Proof. For any $P''\in\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa$, equation (7) implies both $\inf_{Q\in\dot{\mathcal{P}}_\kappa} I\left(Q||P''\right)=I\left(P''||P''\right)=0$ and $I\left(P||P''\right)>0$ for all $P\in\mathcal{P}\backslash(\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa)$ (e.g., Cover and Thomas, 2006). Thus, by Theorem 1,

$$\widetilde{\mathcal{P}}_\kappa=\left\{P\in\mathcal{P}:I\left(P||P''\right)=0,P''\in\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa\right\}=\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa,$$

which was assumed to be nonempty. Equation (18) then follows from equation (10). □

Remark 1. Unless $\kappa=0$, the condition that $\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa$ be nonempty holds whenever the plausible posteriors are sufficiently unrestricted. The most important such setting for applications is a complete lack of constraints ($\dot{\mathcal{P}}=\mathcal{P}$), in which case

$$\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa=\ddot{\mathcal{P}}\cap\left\{\kappa P+(1-\kappa)\dot{P}:P\in\mathcal{P}\right\}$$

and, if $\mathcal{P}$ is convex and unbounded, then $\ddot{\mathcal{P}}\cap\dot{\mathcal{P}}_\kappa=\ddot{\mathcal{P}}\cap\mathcal{P}=\ddot{\mathcal{P}}$ for any $\kappa\in(0,1]$.

4. Examples

The first example of this section addresses the Ellsberg Paradox of the Introduction.

Example 2. Example 1, continued. Suppose the agent may choose an action after observing the colors of one or more balls independently drawn from the

urn with replacement. Let x and n denote the numbers of black and non-red balls drawn, respectively. Since the proportion of balls that are red is known to be $1/3$, nothing is learned from any red balls drawn, and $x = 0, 1, \ldots$ is an outcome of P_θ, the $\langle n, \theta \rangle$-binomial distribution, where θ is the proportion of the non-red balls that are black. Based on standard confidence intervals, the set $\ddot{\mathcal{P}}$ of confidence posteriors is the set of distributions on $([0,1], \mathcal{B}([0,1]))$ such that every $\ddot{P} \in \ddot{\mathcal{P}}$ and $\ddot{\theta} \sim \ddot{P}$ satisfies

$$P_\theta(X > x) \leq \ddot{P}\left(\ddot{\theta} \leq \theta\right) \leq P_\theta(X \geq x)$$

for all $\theta \in [0,1]$ (cf. Bickel, 2012a). Let $\text{beta}(\beta_1, \beta_2)$, the $\langle \beta_1, \beta_2 \rangle$-beta distribution, be the working prior of θ. That is a conjugate prior, and $\dot{P}$ is the $\langle \beta_1 + x, \beta_2 + n - x \rangle$-beta distribution. With $\dot{\mathcal{P}} = \mathcal{P}$ as the set of all distribution on $([0,1], \mathcal{B}([0,1]))$, Theorem 1 applies, implying that the moderate posterior is $\dot{P}$ if, for all $\theta \in [0,1]$,

$$\left\{P_\theta(X > x) \leq \kappa P'\left(\ddot{\theta} \leq \theta\right) + (1-\kappa)\dot{P}\left(\ddot{\theta} \leq \theta\right) \leq P_\theta(X \geq x) : P' \in \dot{\mathcal{P}}\right\} \neq \emptyset.$$

Otherwise, the moderate posterior is the member of $\ddot{\mathcal{P}}$ that is closest to $\dot{P}$ in the sense of Theorem 1.

A simpler approach first transforms the set of confidence posteriors into a single posterior for application of Corollary 1. Any function carrying out such a transform from a set of distributions to a single, representative distribution is called an *inference process* (Paris, 1994; Bickel, 2011a), *representation* (Weichselberger, 2001, p. 258; Augustin, 2002), or *reduction* (Bickel, 2012a). The one that corresponds to the usual mid-p-values yields

$$\ddot{\mathcal{P}} = \left\{\left(\ddot{P}_> + \ddot{P}_\geq\right)/2\right\} \tag{19}$$

as the single-member set of confidence posteriors to be used in place of $\ddot{\mathcal{P}}$, where $\ddot{P}_>$ and $\ddot{P}_\geq$ are defined by

$$\ddot{P}_>\left(\ddot{\theta} \leq \theta\right) = P_\theta(X > x);$$
$$\ddot{P}_\geq\left(\ddot{\theta} \leq \theta\right) = P_\theta(X \geq x)$$

for all $\theta \in [0,1]$. For concreteness, suppose the physical distribution of urns used in this type of game is such that there is symmetry between black and yellow balls and that extreme proportions cannot be more frequent than less extreme proportions. Translated to the beta-binomial model, those constraints are equivalent to using $\{\text{beta}(\beta', \beta'), \beta' \geq 1\}$ as the set of plausible prior distributions of the binomial parameter. Let the working prior distribution of the binomial parameter be $\text{beta}(5,5)$. Corollary 1 gives $\dot{P}_\kappa = \dot{P}_{\kappa,\beta(\kappa)}$, where

$$\dot{P}_{\kappa,\beta'} = \kappa\,\text{beta}(\beta' + x, \beta' + n - x) + (1-\kappa)\,\text{beta}(5 + x, 5 + n - x);$$
$$\beta(\kappa) = \arg\inf_{\beta' \geq 1} I\left(\dot{P}_{\kappa,\beta'} || \left(\ddot{P}_> + \ddot{P}_\geq\right)/2\right).$$

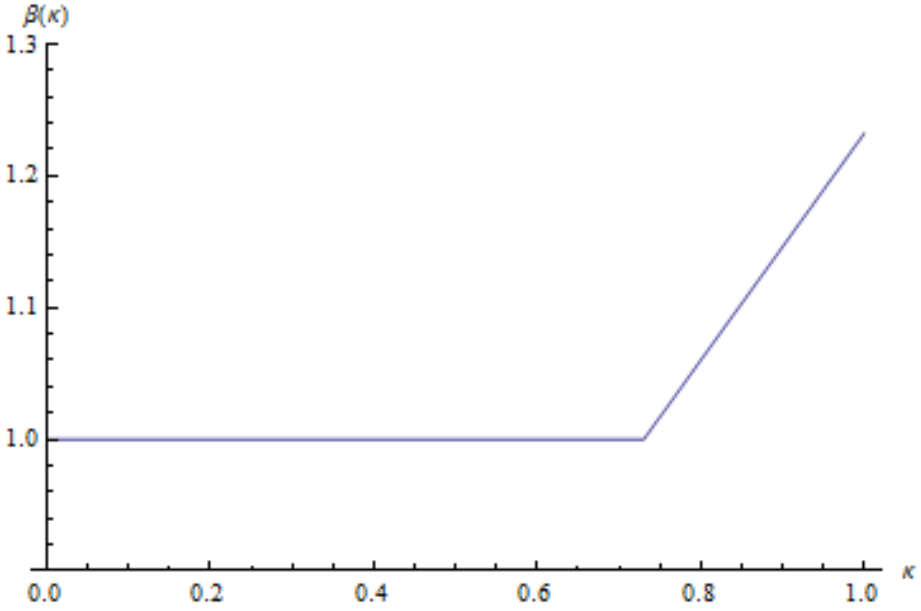

FIG 1. *Optimal strength $\beta(\kappa)$ of the physical prior versus κ, the caution parameter. See Example 2.*

Thus, $\beta(\kappa)$ measures the strength of the first component's optimal prior. Assuming further that a single black ball is drawn ($n = x = 1$), $\beta(\kappa)$ is plotted as a function of κ in Figure 1. The phase transition seen at $\kappa \approx 0.73$ is due to the $\beta' \geq 1$ constraint. The actions that maximize the expected utility are found from the moderate-posterior predictive probability that the next ball drawn will also be black, which is $2/3$ of $\int \theta d\widetilde{P}_\kappa(\theta)$, the expectation value of the binomial parameter, with all expectation values defined with respect to the moderate posterior $\widetilde{P}_\kappa$ (Figure 2). Since that probability is uniformly greater than the predictive probability that the next ball is yellow, actions II and IV are optimal for all $\kappa \in [0, 1]$.

The next two examples involve the continuous, scalar parameters typical of point and interval estimation ($\Theta = \mathbb{R}$). For simplicity, each uses only a single confidence posterior ($\ddot{\mathcal{P}} = \{\ddot{P}\}$).

Example 3. $P_{\theta,1}$ is the normal distribution of mean θ and variance 1, i.e., $X \sim \mathrm{N}(\theta, 1)$, and $X = x$ is observed. Further, $\theta \sim \mathrm{N}(\mu, \sigma^2)$ with unknown μ and σ of known lower and upper bounds: $\mu \in [\underline{\mu}, \overline{\mu}]$; $\sigma \in [\underline{\sigma}, \overline{\sigma}]$. For generality, the bounds are extended real numbers: $\underline{\mu} \in \{-\infty\} \cup \mathbb{R}$, $\underline{\sigma} \in [0, \infty)$, $\overline{\mu} = \mathbb{R} \cup \{\infty\}$, $\overline{\sigma} = [0, \infty) \cup \{\infty\}$. The intervals $[\underline{\mu}, \overline{\mu}]$ and $[\underline{\sigma}, \overline{\sigma}]$ are open only as required to ensure that $\mu, \log\sigma \in \mathbb{R}$ in the presence of infinite bounds or $\underline{\sigma} = 0$, e.g., $\underline{\mu} = 0, \overline{\mu} = \infty \implies [\underline{\mu}, \overline{\mu}] = [0, \infty)$. The working prior is $\mathrm{N}(\dot{\mu}, \dot{\sigma}^2)$ for some given

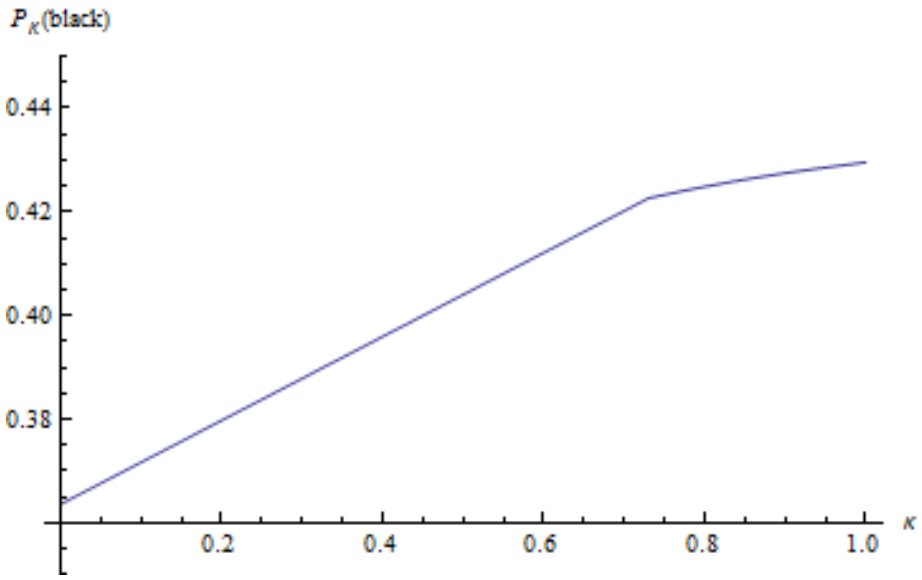

FIG 2. *Moderate-posterior probability of drawing a black ball versus κ, the caution parameter. See Example 2.*

$\dot{\mu} \in \left[\underline{\mu}, \overline{\mu}\right]$ and $\dot{\sigma} \in [\underline{\sigma}, \overline{\sigma}]$. By Bayes's theorem (e.g., Carlin and Louis, 2009),

$$\dot{P} = \mathrm{N}\left(\frac{\dot{\mu} + \dot{\sigma}^2 x}{1+\dot{\sigma}^2}, \frac{\dot{\sigma}^2}{1+\dot{\sigma}^2}\right);$$

$$\dot{\mathcal{P}} = \left\{\mathrm{N}\left(\frac{\mu + \sigma^2 x}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right) : \mu \in \left[\underline{\mu}, \overline{\mu}\right], \sigma \in [\underline{\sigma}, \overline{\sigma}]\right\}.$$

By contrast, $\ddot{P}$ is $\mathrm{N}(x, 1)$, not depending on any prior. This $\ddot{P}$ is a genuine confidence posterior (§2.3), as can be verified from the fact that $\ddot{P}(\ddot{\theta} \leq \theta) = P_{\theta,1}(X \geq x)$ for all $\theta \in \mathbb{R}$ and $x \in \mathbb{R}$. The performance of any estimator $\widehat{\theta}$ of θ may be quantified by its squared-error prediction loss: $L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$. By equation (1), the κCG estimate is

$$\begin{aligned}
\dot{a}_\kappa &= \arg\inf_{\widehat{\theta}\in\mathbb{R}} \left(\kappa \sup_{P' \in \dot{\mathcal{P}}} \int \left(\widehat{\theta} - \theta\right)^2 dP'(\theta) + (1-\kappa) \int \left(\widehat{\theta} - \theta\right)^2 d\dot{P}(\theta)\right) \\
&= \arg\inf_{\widehat{\theta}\in\mathbb{R}} \left(\kappa \left[\int \left(\widehat{\theta} - \theta\right)^2 dP_{\mu(\widehat{\theta}),\underline{\sigma}}(\theta)\right] + (1-\kappa) \int \left(\widehat{\theta} - \theta\right)^2 d\dot{P}(\theta)\right),
\end{aligned}$$

where $\mu(\widehat{\theta}) = \underline{\mu}$ if $|\widehat{\theta} - \underline{\mu}| > |\widehat{\theta} - \overline{\mu}|$ and $\mu(\widehat{\theta}) = \overline{\mu}$ otherwise; $P_{\mu(\widehat{\theta}),\underline{\sigma}} = \mathrm{N}(\mu(\widehat{\theta}), \underline{\sigma}^2)$. If $\underline{\sigma} = 0$, $P_{\mu(\widehat{\theta}),0}$ is the Dirac measure at $\mu(\widehat{\theta})$, implying that

$$\dot{a}_\kappa = \arg\inf_{\widehat{\theta}\in\mathbb{R}} \left(\kappa \left(\widehat{\theta} - \mu\left(\widehat{\theta}\right)\right)^2 + (1-\kappa) \int \left(\widehat{\theta} - \theta\right)^2 d\dot{P}(\theta)\right),$$

which only has a solution if $\underline{\mu} > -\infty$ and $\overline{\mu} < \infty$. Those restrictions are not needed for the estimate based on the moderate posterior. Since

$$\dot{\mathcal{P}}_\kappa = \left\{ \kappa \, \mathrm{N}\left(\frac{\mu + \sigma^2 x}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2} \right) + (1-\kappa)\, \dot{P} : \mu \in \left[\underline{\mu}, \overline{\mu}\right], \sigma \in [\underline{\sigma}, \overline{\sigma}] \right\}, \quad (20)$$

Corollary 1 entails

$$\begin{aligned} \widetilde{P}_\kappa &= \arg \inf_{\mu \in [\underline{\mu}, \overline{\mu}], \sigma \in [\underline{\sigma}, \overline{\sigma}]} I\left(\mathrm{N}\left(\frac{\mu + \sigma^2 x}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2} \right) || \, \mathrm{N}(x, 1) \right) \\ &= \arg \inf_{\mu \in [\underline{\mu}, \overline{\mu}], \sigma \in [\underline{\sigma}, \overline{\sigma}]} \left(\log \frac{1+\sigma^2}{\sigma^2} + \frac{\sigma^2}{1+\sigma^2} + \left(\frac{\mu - x}{1+\sigma^2} \right)^2 \right), \end{aligned}$$

with the second equality from, e.g., Kullback (1968, p. 189). Substituting $\widetilde{P}_\kappa$ into equation (13) gives the moderate-posterior-mean as the estimate of θ:

$$\widetilde{a}_\kappa = \arg \inf_{\widehat{\theta} \in \mathbb{R}} \int \left(\widehat{\theta} - \theta \right)^2 d\widetilde{P}_\kappa(\theta) = \int \theta d\widetilde{P}_\kappa(\theta),$$

which is unique even if $\underline{\mu} = -\infty$, $\overline{\mu} = \infty$, $\underline{\sigma} = 0$, and $\overline{\sigma} = \infty$.

The next example drops the parametric assumptions about the plausible prior distributions.

Example 4. $X \sim \mathrm{N}(\theta, 1)$ with no information about θ except that $\theta \in \mathbb{R} = \Theta$, that $X = x$ is observed, and that $\dot{P}$ is the working Bayesian posterior distribution of θ. It follows that $\dot{\mathcal{P}}$ is the set of all distributions on the Borel space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Again under quadratic loss, by equation (1), the κCG estimate is

$$\dot{a}_\kappa = \arg \inf_{\widehat{\theta} \in \mathbb{R}} \left(\kappa \sup_{P' \in \dot{\mathcal{P}}} \int \left(\widehat{\theta} - \theta \right)^2 dP'(\theta) + (1-\kappa) \int \left(\widehat{\theta} - \theta \right)^2 d\dot{P}(\theta) \right),$$

which is the posterior mean $\int \theta d\dot{P}(\theta)$ if $\kappa = 0$ but which has no unique value for any other value of κ since $\sup_{P' \in \dot{\mathcal{P}}} \int (\widehat{\theta} - \theta)^2 dP'(\theta) = \infty$ for any $\widehat{\theta}$. By contrast, equation (13) specifies the unique moderate-posterior estimate given $\ddot{P} = \mathrm{N}(x, 1)$:

$$\widetilde{a}_\kappa = \arg \inf_{\widehat{\theta} \in \mathbb{R}} \int \left(\widehat{\theta} - \theta \right)^2 d\widetilde{P}_\kappa(\theta) = \int \theta d\widetilde{P}_\kappa(\theta),$$

where, provided that $\kappa > 0$, $\widetilde{P}_\kappa = \ddot{P}$ according to Corollary 2 since $\ddot{P} \in \mathcal{P} = \dot{\mathcal{P}}_\kappa$, leading to

$$\widetilde{a}_\kappa = \int \theta d\ddot{P}(\theta),$$

the frequentist posterior mean.

The last example involves a discrete focus parameter, as is typical of hypothesis testing and model selection applications.

Example 5. Consider the indicator parameter θ defined such that $\theta = 0$ if the null hypothesis about θ_{**} is true ($\theta_{**} = 0$) and $\theta = 1$ if the alternative hypothesis about θ_{**} is true ($\theta_{**} \neq 0$). Equivalently, in terms of $\theta_* = |\theta_{**}|$, $\theta = 0$ if $\theta_* = 0$ and $\theta = 1$ if $\theta_* > 0$. If $\dot{P}$ is a working Bayesian posterior for θ_{**}, then $\dot{P}(\dot{\theta} = 0)$ is the corresponding working Bayesian posterior probability that the null hypothesis is true. Let $p^{(1)}$ and $p^{(2)}$ denote observed p-values of the one-sided test of $\theta_* = 0$ versus $\theta_* > 0$ and thus of the two-sided test of $\theta_{**} = 0$ versus $\theta_{**} \neq 0$. In this example, $p^{(1)}(x) \leq p^{(2)}(x)$, perhaps because $p^{(2)}(x)$ is based on a test that makes weaker parametric assumptions than that of $p^{(1)}(x)$. For $i = 1, 2$, let $\ddot{P}_*^{(i)}$ denote the confidence posterior for θ_* defined given some $x \in \mathcal{X}$ such that

$$\ddot{P}_*^{(i)}\left(\ddot{\theta}_* \leq \theta_*\right) = P_{\theta_*,\lambda_*}\left(p^{(i)}(X) \leq p^{(i)}(x)\right)$$

for all $\theta_* \in \Theta_*$ and $\lambda_* \in \Lambda_*$, where the dependence of $\ddot{P}_*^{(i)}$ on x is suppressed, in Section 2.3. Since $p^{(i)}(X) \sim \mathrm{U}(0,1)$ under the null hypothesis that $\theta_* = 0$, it follows that

$$\ddot{P}_*^{(i)}\left(\ddot{\theta}_* \leq 0\right) = \ddot{P}_*^{(i)}\left(\ddot{\theta}_* = 0\right) = P_{0,\lambda_*}\left(p^{(i)}(X) \leq p^{(i)}(x)\right) = p^{(i)}(x), \quad (21)$$

i.e., the confidence posterior probability of the null hypothesis is equal to the p-value (Bickel, 2011c,a); cf. van Berkum, Linssen and Overdijk (1996). With $\ddot{\theta} = 0$ if $\ddot{\theta}_* = 0$ and $\ddot{\theta} = 1$ if $\ddot{\theta}_* \neq 0$, equation (21) yields $\ddot{P}^{(i)}(\ddot{\theta} = 0) = p^{(i)}(x)$. From widely applicable conditions for two-sided hypothesis testing (Sellke, Bayarri and Berger, 2001; Bickel, 2011a) and with some $\underline{\dot{P}}_\emptyset^{\text{plaus}} \in (0,1)$ given as the lower bound of the prior probabilities of the null hypothesis and the restriction that no such probability is 1, the knowledge base is

$$\dot{\mathcal{P}} = \left\{P' \in \mathcal{P} : \underline{\dot{P}}\left(\underline{\dot{\theta}} = 0\right) = \underline{\dot{P}}(\{0\}) \leq P'(\{0\}) < 1\right\},$$

the set of plausible posteriors, the distributions on $\left(\{0,1\}, 2^{\{0,1\}}\right)$ with

$$\underline{\dot{P}}_\emptyset = \underline{\dot{P}}\left(\underline{\dot{\theta}} = 0\right) = \left(1 + \left(\frac{1 - \underline{\dot{P}}_\emptyset^{\text{plaus}}}{\underline{\dot{P}}_\emptyset^{\text{plaus}} e p^{(2)}(x) \log\left[1/p^{(2)}(x)\right]}\right)\right)^{-1} \wedge \underline{\dot{P}}_\emptyset^{\text{plaus}}$$

as the lower bound of the plausible posterior probability of the null hypothesis, where $\underline{\dot{\theta}} \sim \underline{\dot{P}}$. That lower bound is the greater of the two lower bounds found by separately applying the methodology of Sellke, Bayarri and Berger (2001) to $p^{(1)}(x)$ and $p^{(2)}(x)$. (The binary operator $\wedge$ in the above equation means "the minimum of," and $\vee$ will similarly stand for "the maximum of.") Since Theorem 1 applies, the moderate posterior $\widetilde{P}_\kappa$ is given by equation (10) with

$$\begin{aligned}\widetilde{\mathcal{P}}_\kappa &= \left\{P \in \mathcal{P} : I\left(P||\ddot{P}^{(1)}\right) \wedge I\left(P||\ddot{P}^{(2)}\right) = I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) \wedge I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right)\right\} \\ &= \left\{\widetilde{P}_\kappa^{(i)} : I\left(\widetilde{P}_\kappa^{(i)}||\ddot{P}^{(i)}\right) = I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) \wedge I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right), i \in \{1,2\}\right\},\end{aligned}$$

where $\widetilde{P}_\kappa^{(i)} = \arg\inf_{Q\in\dot{\mathcal{P}}_\kappa} I\left(Q||\ddot{P}^{(i)}\right)$; $\dot{\mathcal{P}}_\kappa = \{\kappa P' + (1-\kappa)\dot{P} : P' \in \mathcal{P}, \underline{\dot{P}}_\emptyset \leq P'(\theta = 0) < 1\}$. More simply,

$$\widetilde{P}_\kappa = \begin{cases} \widetilde{P}_\kappa^{(1)} & \text{if } I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) < I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right) \\ \widetilde{P}_\kappa^{(2)} & \text{if } I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) > I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right) \\ \widetilde{P}_\kappa^{(1)} & \text{if } I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) = I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right) \text{ and } I\left(\dot{P}||\widetilde{P}_\kappa^{(1)}\right) \leq I\left(\dot{P}||\widetilde{P}_\kappa^{(2)}\right) \\ \widetilde{P}_\kappa^{(2)} & \text{if } I\left(\widetilde{P}_\kappa^{(1)}||\ddot{P}^{(1)}\right) = I\left(\widetilde{P}_\kappa^{(2)}||\ddot{P}^{(2)}\right) \text{ and } I\left(\dot{P}||\widetilde{P}_\kappa^{(1)}\right) \geq I\left(\dot{P}||\widetilde{P}_\kappa^{(2)}\right). \end{cases}$$

Letting $\dot{P}_\emptyset = \dot{P}(\dot{\theta} = 0)$ and letting $\widetilde{\theta}$ denote the focus parameter according to the moderate posterior ($\widetilde{\theta} \sim \widetilde{P}_\kappa$),

$$\begin{aligned} &\widetilde{P}_\kappa^{(i)} = \arg\inf_{Q\in\mathcal{P}_\kappa} \sum_{j=0,1} Q(\theta = j) \log \frac{Q(\theta = j)}{\ddot{P}^{(i)}(\theta = j)} \\ &\widetilde{P}_\kappa^{(i)}\left(\widetilde{\theta} = 0\right) \\ &\quad = \arg\inf_{Q_\emptyset \in \left\{\kappa P_\emptyset + (1-\kappa)\dot{P}_\emptyset : P_\emptyset \in \left[\underline{\dot{P}}_\emptyset, 1\right)\right\}} Q_\emptyset \log \frac{Q_\emptyset}{p^{(i)}(x)} + (1 - Q_\emptyset) \log \frac{1 - Q_\emptyset}{1 - p^{(i)}(x)} \\ &\quad = \arg\inf_{Q_\emptyset \in \left[\kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset, \kappa + (1-\kappa)\dot{P}_\emptyset\right)} Q_\emptyset \log \frac{Q_\emptyset}{p^{(i)}(x)} + (1 - Q_\emptyset) \log \frac{1 - Q_\emptyset}{1 - p^{(i)}(x)} \\ &\quad = \begin{cases} \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset & \text{if } p^{(i)}(x) < \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \\ p^{(i)}(x) & \text{if } \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \leq p^{(i)}(x) \leq \kappa + (1-\kappa)\dot{P}_\emptyset \\ \kappa + (1-\kappa)\dot{P}_\emptyset & \text{if } p^{(i)}(x) > \kappa + (1-\kappa)\dot{P}_\emptyset. \end{cases} \end{aligned} \tag{22}$$

Since $\widetilde{P}_\kappa \in \widetilde{\mathcal{P}}_\kappa$,

$$\begin{aligned} &\widetilde{P}_\kappa\left(\widetilde{\theta} = 0\right) \\ &\quad \in \begin{cases} \left\{\kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset\right\} & \text{if } p^{(1)}(x) \leq p^{(2)}(x) < \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \\ \left\{p^{(2)}(x)\right\} & \text{if } p^{(1)}(x) < \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \leq p^{(2)}(x) \leq \kappa + (1-\kappa)\dot{P}_\emptyset \\ \left\{p^{(1)}(x), p^{(2)}(x)\right\} & \text{if } \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \leq p^{(1)}(x) \leq p^{(2)}(x) \leq \kappa + (1-\kappa)\dot{P}_\emptyset \\ \left\{p^{(1)}(x)\right\} & \text{if } \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset \leq p^{(1)}(x) \leq \kappa + (1-\kappa)\dot{P}_\emptyset < p^{(2)}(x) \\ \left\{\kappa + (1-\kappa)\dot{P}_\emptyset\right\} & \text{if } p^{(2)}(x) \geq p^{(1)}(x) > \kappa + (1-\kappa)\dot{P}_\emptyset, \end{cases} \end{aligned} \tag{23}$$

from which the extreme condition $p^{(1)}(x) < \kappa\underline{\dot{P}}_\emptyset + (1-\kappa)\dot{P}_\emptyset < \kappa + (1-\kappa)\dot{P}_\emptyset < p^{(2)}(x)$ is omitted for brevity. In the case of no caution, the working Bayesian posterior probability is recovered: $\widetilde{P}_0(\widetilde{\theta} = 0) = \dot{P}(\dot{\theta} = 0)$, which does not depend

on $p\,(x)$. More interestingly, the case of complete caution leads to

$$\widetilde{P}_1\left(\widetilde{\theta}=0\right)\in\begin{cases}\left\{\underline{\dot{P}}\left(\underline{\dot{\theta}}=0\right)\right\} & \text{if } p^{(1)}\,(x)\leq p^{(2)}\,(x)<\underline{\dot{P}}_{\emptyset}\\ \left\{p^{(2)}\,(x)\right\} & \text{if } p^{(1)}\,(x)<\underline{\dot{P}}_{\emptyset}\leq p^{(2)}\,(x)\\ \left\{p^{(1)}\,(x)\,,p^{(2)}\,(x)\right\} & \text{if } \underline{\dot{P}}_{\emptyset}\leq p^{(1)}\,(x)\leq p^{(2)}\,(x)\,,\end{cases}\tag{24}$$

which has no dependence on $\dot{P}(\dot{\theta}=0)$. The simplifying effect of considering only a single p-value is evident from using $p^{(1)}\,(x)=p^{(2)}\,(x)$ in the formulas (23) and (24). For example, expression (24) results in a unique $\widetilde{P}_1(\widetilde{\theta}=0)$ equal to the blended posterior probability of Bickel (2011a). When formulas (23) and (24) say no more than $\widetilde{P}_\kappa(\widetilde{\theta}=0)\in\left\{p^{(1)}\,(x)\,,p^{(2)}\,(x)\right\}$, equation (10) ensures the uniqueness of the moderate posterior probability by equating it with the p-value closest to the working Bayesian posterior probability:

$$\begin{aligned}\widetilde{P}_\kappa &= \arg\inf_{P'''\in\left\{p^{(1)}(x),p^{(2)}(x)\right\}} I\left(\dot{P}||P'''\right)=\ddot{P}^{(\widetilde{\imath})};\\ \widetilde{\imath} &= \arg\inf_{i\in\{1,2\}}\left(\dot{P}\left(\dot{\theta}=0\right)\log\frac{\dot{P}\,(\theta=0)}{p^{(i)}\,(x)}+\dot{P}\left(\dot{\theta}=1\right)\log\frac{\dot{P}\left(\dot{\theta}=1\right)}{1-p^{(i)}\,(x)}\right),\end{aligned}$$

which is a special case of Corollary 2. In this way, the caution parameter, the working Bayesian posterior, and the constraints on the plausible posteriors together overcome the dilemma of whether to use the more conservative p-value or the less conservative p-value.

5. Extending the caution framework

5.1. Variations of the framework

The above framework for balancing Bayesian and frequentist approaches to inference does not apply to all situations encountered in applications. The various permutations of the Bayesian and confidence posteriors as the *working posterior* $\dot{P}$, used exclusively in the absence of caution, and a *benchmark posterior* $\ddot{P}$, over which inference will be improved as much as possible, in equations (14) and (17) lead to four versions of the proposed approach:

1. $\dot{P}$ is a Bayesian posterior in $\dot{\mathcal{P}}$, and $\ddot{P}$ is a confidence posterior. This version yields the balance between Bayesian and frequentist inference defined in Section 3 and illustrated in Section 4.
2. $\dot{P}$ is a confidence posterior, and $\ddot{P}$ is a Bayesian posterior in $\dot{\mathcal{P}}$. The potential uses of this reversal are unclear since it would paradoxically lead to dependence on a single Bayesian posterior to the extent of the caution.
3. $\dot{P}=\ddot{P}$, where $\dot{P}$ is a Bayesian posterior in $\dot{\mathcal{P}}$. Using the same Bayesian posterior as both the working posterior and the benchmark posterior is attractive in the absence of reliable confidence intervals or p-values from which a

TABLE 3
Settings for three versions of the proposed framework

Setting	$\dot{P}$	$\ddot{P}$ or $\ddot{\mathcal{P}}$
Bayes and frequentist approaches apply	Bayes posterior	confidence posterior
no confidence intervals or p-values	Bayes posterior	
continuous θ & only a set of Bayes posteriors	confidence posterior	

confidence posterior could be constructed. Thus, this version extends the scope of the framework across the domains to which Bayesian methods apply. However, this version becomes trivial whenever equation (17) holds according to Corollary 1, for in that case, $\widetilde{P}_\kappa = \arg\inf_{Q\in\dot{\mathcal{P}}_\kappa} I(Q||\dot{P}) = \dot{P}$ for all $\kappa \in [0,1]$ since $\dot{P} \in \dot{\mathcal{P}}_\kappa$ necessarily. In other words, the Bayesian posterior would be used for inference irrespective of the degree of caution and the knowledge base.

4. $\dot{P} = \ddot{P}$, where $\ddot{P}$ is a confidence posterior. Using the same confidence posterior as both the working posterior and the benchmark posterior is useful when a set $\dot{\mathcal{P}}$ of plausible posteriors can be specified but when no member of that set can be singled out as special. In many cases involving a continuous parameter θ, no such member can be derived from the knowledge base $\dot{\mathcal{P}}$ without imposing arbitrary procedures such as averaging over the members with respect to some measure chosen for convenience. That will be explained in Section 5.2, where the case of two unequal confidence posteriors will also be considered.

For simplicity, the versions are described as if $\ddot{\mathcal{P}} = \{\ddot{P}\}$, but they also pertain to a set $\dot{\mathcal{P}}$ of multiple benchmark posteriors that define the moderate posterior $\widetilde{P}_\kappa$ according to equation (10). The three most applicable of those versions are summarized in Table 3.

Generalizing beyond those versions, the working posterior $\dot{P}$ can be any posterior distribution that would be used exclusively in the absence of caution, whereas when there is caution, inferences are made with the goal that they improve upon those that would be made using any other posterior distribution in $\ddot{\mathcal{P}}$. The application at hand can help determine which of those distributions is a Bayesian posterior and which is some other type of distribution such as a confidence distribution. For example, given a working posterior $\dot{P}$ from a proper prior, a posterior from an improper prior could be used as the benchmark posterior $\ddot{P}$ in the absence of a suitable confidence posterior (cf. Bickel, 2011a).

5.2. Inference without a working Bayesian posterior

The more interesting of the two widely applicable variations of the framework is that in which both $\dot{P}$ and $\ddot{P}$ are the same confidence posterior. Thus, some implications of using a single confidence posterior simultaneously as the working posterior and as the benchmark posterior ($\dot{P} = \ddot{P}$) merit noting. First, complete caution ($\kappa = 1$) leads to ignoring the role of the confidence posterior as a working posterior and thereby collapses to the blended inferences of Bickel (2011a).

Second, when $\kappa < 1$ and the confidence posterior is not a plausible posterior ($\ddot{P} \notin \dot{\mathcal{P}}$), the moderate posterior may not be a plausible posterior ($\widetilde{P}_\kappa \notin \dot{\mathcal{P}}$). In fact, whenever sufficient conditions for Corollary 1 are met, $\widetilde{P}_\kappa$ will be plausible only if $\ddot{P}$ is plausible. That suggests the use of $\kappa = 1$ in the absence of a working Bayesian posterior in order to avoid excessive dependence on $\ddot{P}$ at the expense of $\dot{\mathcal{P}}$, the knowledge base. On the other hand, allowing $\widetilde{P}_\kappa \notin \dot{\mathcal{P}}$ makes $\widetilde{P}_\kappa$ less dependent on the precise borders of $\dot{\mathcal{P}}$, and this may be desirable to the extent that such borders are uncertain or subjectively specified. Third, when $\ddot{P} \in \dot{\mathcal{P}}$, the moderate posterior is simply equal to the confidence posterior ($\widetilde{P}_\kappa = \ddot{P}$) under the sufficient conditions for equation (17) given by Corollary 1. The following examples illustrate the second and third implications.

Example 6 (Variation of Example 4)**.** This example is trivial since $\dot{P} \in \mathcal{P}$, $\dot{P} = \ddot{P}$, and Corollary 1 entail $\widetilde{P}_\kappa = \ddot{P}$.

More generally, $\ddot{P} \in \dot{\mathcal{P}}$ and $\dot{P} = \ddot{P}$ imply $\widetilde{P}_\kappa = \ddot{P}$ by Corollary 2.

Example 7 (Variation of Example 5)**.** Because $\ddot{P}(\ddot{\theta} = 0) = p\,(x)$, the identity $\dot{P} = \ddot{P}$ yields $\dot{P}_\emptyset = p\,(x)$, reducing equation (22) to

$$\widetilde{P}_\kappa\left(\widetilde{\theta} = 0\right) = \begin{cases} \kappa \underline{\dot{P}}_\emptyset + (1-\kappa)\, p\,(x) & \text{if } p\,(x) < \underline{\dot{P}}_\emptyset \\ p\,(x) & \text{if } p\,(x) \geq \underline{\dot{P}}_\emptyset. \end{cases}$$

Re-expressing this as $\widetilde{P}_\kappa(\widetilde{\theta} = 0) = [\kappa \underline{\dot{P}}_\emptyset + (1-\kappa)\, p\,(x)] \vee p\,(x)$ facilitates comparison with the equation $\widetilde{P}_1(\widetilde{\theta} = 0) = \underline{\dot{P}}_\emptyset \vee p\,(x)$ used in the blended inference framework (Bickel, 2011a). Therefore, $\kappa \in [0, 1)$ entails that $\widetilde{P}_\kappa(\widetilde{\theta} = 0)$ is less than the lower bound $\underline{\dot{P}}_\emptyset$ whenever $p\,(x) < \underline{\dot{P}}_\emptyset$. That would be clearly unacceptable if $\underline{\dot{P}}_\emptyset^{\text{plaus}}$ is scientifically established, but if $\underline{\dot{P}}_\emptyset^{\text{plaus}}$ is instead highly uncertain or subjectively assessed, then $\widetilde{P}_\kappa(\widetilde{\theta} = 0)$ can bypass $\underline{\dot{P}}_\emptyset$ as warranted.

An alternative to the above approach in the absence of a specified $\dot{P}$ is to apply the strategy of Section 3 with $\dot{P}$ as a function of $\dot{\mathcal{P}}$, following Gajdos (2008). Examples of functions that transform a set of distributions to a single distribution include the Steiner point (Gajdos, 2008), the arithmetic mean ("center of mass," e.g., (19)), and the maximum entropy distribution (Paris, 1994). In the continuous-parameter case, such functions require a base measure for partitioning.

There is no need to impose an arbitrary base measure if two different confidence posteriors $\dot{P}$ and $\ddot{P}$ are under consideration ($\dot{P} \neq \ddot{P}$). Using them as the working posterior and the benchmark posterior in equations (14) and (17) would be most appropriate when $\dot{P}$ represents a newer or relatively untested procedure and when $\ddot{P}$ corresponds to a better established or more thoroughly tested procedure. More generally, equation (10) specifies how to apply a working confidence posterior $\dot{P}$ with a set $\ddot{\mathcal{P}}$ of benchmark confidence posteriors.

6. Discussion

The featured moderate-posterior methodology has been contrasted with the simpler κCG methodology. As Examples 3 and 4 illustrated under quadratic loss, the former can yield unique actions in a wide variety of settings in which the latter cannot. Using CG minimaxity ($\kappa = 1$), uniqueness has been achieved under quadratic loss by restricting the action space to finite bounds (Betrò and Ruggeri, 1992) and by similarly restricting the parameter space Θ (Bayati Eshkaftaki and Parsian, 2011). The moderate-posterior estimators did not require such restrictions.

The main advantage of the moderate-posterior framework is that it provides first principles from which a statistician may derive a Bayesian analysis, a frequentist analysis, or a combination of the two, depending on the chosen level of caution and on the quality of prior information. This allows the caution level to be precisely reported with the resulting statistical inferences. In addition, the caution level may be determined by the needs of an organization or collaborating scientist rather than by the personal attitude of the statistician.

Various factors may be considered in choosing the level of caution. For example, more caution with Bayesian inference may be warranted when the confidence posterior represents a frequentist procedure that has stood the test of time than when it represents a new frequentist procedure based on questionable assumptions. The caution level could then be interpreted as the pre-data degree of reluctance an agent has in modifying the frequentist procedures encoded in the confidence posterior.

The moderate-posterior framework of Section 3 is general enough to incorporate conflicting frequentist approaches, as seen in Example 5. For additional generality, Section 5.2 provides ways to modify the framework for situations in which any dependence on a subjective or guessed Bayesian prior would be undesirable.

Another general approach for such situations is that of a three-player game against a player that chooses the true distribution from $\dot{\mathcal{P}}$ and against a player that chooses a competing posterior from $\ddot{\mathcal{P}}$ (Bickel, 2012b). Since the framework of the present paper takes a minimax strategy, it can be formulated as a two-player game against the former player (see Topsøe, 1979) in order to relate it to the three-player strategy.

In other situations, any dependence of inference on the level of caution would be undesirable. Provided that there is at least a little caution, the use of a sufficiently broad set of plausible posteriors under the unmodified framework (§3) eliminates any other dependence on the degree of caution (Remark 1).

That most applications require more complicated models than those of the examples raises the question of implementing algorithms for computing the moderate posterior. Fortunately, under the usual conditions, those specified in Theorem 1, the moderate posterior is among those that the maximize the differential entropy $-I(Q||P'')$ with respect to a confidence posterior or other benchmark posterior as the base measure P''; see Grünwald and Dawid (2004). Thus, in most cases, the moderate posterior can be found by one of the many strategies

available in the vast literature on maximum entropy and minimum information divergence. (A key difference between the maximum entropy approach of Theorem 1 and that of generating default priors (Jaynes, 2003) is that the moderate posterior must be defined in terms of posteriors rather than priors since not all confidence posteriors have corresponding priors.) For cases in which the conditions of Theorem 1 do not hold, strategies proposed by Topsøe (1979) and Topsøe (2007) for finding $\sup_{Q\in\mathcal{P}} \inf_{P'\in\dot{\mathcal{P}}} I(P'||\ddot{P} \rightsquigarrow Q)$ may prove useful.

Acknowledgments

I am grateful to an anonymous reviewer for insightful comments that led to many improvements in presentation and to the addition of Example 2. This research was partially supported by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa.

References

Augustin, T. (2002). Expected utility within a generalized concept of probability - a comprehensive framework for decision making under ambiguity. *Statistical Papers* **43** 5-22. MR1883831

Bayati Eshkaftaki, A. and Parsian, A. (2011). Robust Bayes estimation. *Communications in Statistics - Theory and Methods* **40** 929–941. MR2762935

Berger, J. O., Insua, D. R. and Ruggeri, F. (2000). *Bayesian Robustness. Robust Bayesian Analysis* 1 1-32. Springer, New York.

Betrò, B. and Ruggeri, F. (1992). Conditional Γ-minimax actions under convex losses. *Communications in Statistics - Theory and Methods* **21** 1051–1066. MR1173305

Bickel, D. R. (2011a). Blending Bayesian and frequentist methods according to the precision of prior information with an application to hypothesis testing. *Technical Report, Ottawa Institute of Systems Biology*, arXiv:1107.2353.

Bickel, D. R. (2011b). Estimating the null distribution to adjust observed confidence levels for genome-scale screening. *Biometrics* **67** 363-370. MR2829005

Bickel, D. R. (2011c). Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison. *Technical Report, Ottawa Institute of Systems Biology*, arXiv:1104.0341.

Bickel, D. R. (2011d). The strength of statistical evidence for composite hypotheses: Inference to the best explanation. *Statistica Sinica* DOI: 10.5705/ss.2009.125 (online ahead of print).

Bickel, D. R. (2012a). Coherent frequentism: A decision theory based on confidence sets. To appear in *Communications in Statistics - Theory and Methods*; 2009 version available from arXiv:0907.0139.

Bickel, D. R. (2012b). Game-theoretic probability combination with applications to resolving conflicts between statistical methods. To appear in *International Journal of Approximate Reasoning*; 2011 version available from arXiv:1111.6174.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, Third Edition*. Chapman & Hall/CRC, New York. MR2442364

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley and Sons, New York. MR2239987

Csiszár, I. (1985). An extended maximum entropy principle and a Bayesian justification. In *Bayesian Statistics 2* (J. M. Bernardo, M. B. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 83–98. Elsevier Inc., Amsterdam. MR0862485

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3-26. MR1225211

Efron, B. (2005). Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association* **100** 1–5. MR2166064

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics* **75** pp. 643-669.

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

Gajdos, T., Hayashi, T., Tallon, J. M. and Vergnaud, J. C. (2008). Attitude toward imprecise information. *Journal of Economic Theory* **140** 27-65.

Gajdos, T., Tallon, J. M. and Vergnaud, J. C. (2004). Decision making with imprecise probabilistic information. *Journal of Mathematical Economics* **40** 647-681. MR2070961

Gärdenfors, P. and Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese* **53** 361-386. 10.1007/BF00486156. MR0691636

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* **18** 141-153. MR1000102

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics* **32** 1367-1433. MR2089128

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica* **19** 491-544. MR2514173

Harremoës, P. (2007). Information topologies with applications. In *Entropy, Search, Complexity*, (I. Csiszár, G. O. H. Katona, G. Tardos and G. Wiener, eds.). *Bolyai Society Mathematical Studies* **16** 113–150. Springer Berlin Heidelberg, Berlin, Heidelberg. MR2427795

Hurwicz, L. (1951a). Optimality criteria for decision making under ignorance. *Cowles Commission Discussion Paper* **370**.

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

Jaffray, J. Y. (1989a). Généralisation du critère de l'utilité espérée aux choix dans l'incertain régulier. *RAIRO: Recherche opérationnelle* **23** 237–267. MR1025078

Jaffray, J.-Y. (1989b). Linear utility theory for belief functions. *Operations Research Letters* **8** 107-112. MR0995970

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. MR1992316

Kardaun, O. J. W. F., Salomi, D., Schaafsma, W., Steerneman, A. G. M., Willems, J. C. and Cox, D. R. (2003). Reflections on Fourteen Cryptic Issues concerning the Nature of Statistical Inference. *International Statistical Review / Revue Internationale de Statistique* **71** 277-303.

Kullback, S. (1968). *Information Theory and Statistics*. Dover, New York. MR1461541

Paris, J. B. (1994). *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge University Press, New York. MR1314199

Pfaffelhuber, E. (1977). Minimax Information Gain and Minimum Discrimination Principle. In *Topics in Information Theory* (I. Csiszár and P. Elias, eds.). *Colloquia Mathematica Societatis János Bolyai* **16** 493–519. János Bolyai Mathematical Society and North-Holland. MR0459931

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

Samaniego, F. J. (2010). *A Comparison of the Bayesian and Frequentist Approaches to Estimation (Springer Series in Statistics)*. Springer, New York. MR2664350

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York. MR0063582

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29** 309-332. MR1909788

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician* **55** 62-71. MR1818723

Singh, K., Xie, M. and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Annals of Statistics* **33** 159-183. MR2157800

Tapking, J. (2004). Axioms for preferences revealing subjective uncertainty and uncertainty aversion. *Journal of Mathematical Economics* **40** 771–797. MR2094688

Topsøe, F. (1979). Information theoretical optimization techniques. *Kybernetika* **15** 8-27. MR0529888

Topsøe, F. (2004). Information theory and complexity in probability and statistics. In *Soft Methodology and Random Information Systems* (Lopez-Diaz, M. and Gil, M. A. and Grzegorzewski, P. and Hryniewicz, O. and Lawry, J. , ed.). *Advances in Soft Computing* 363-370. MR2118118

Topsøe, F. (2007). Information Theory at the Service of Science. In *Entropy, Search, Complexity*, (I. Csiszár, G. O. H. Katona, G. Tardos and G. Wiener, eds.). *Bolyai Society Mathematical Studies* 179-207. Springer Berlin Heidelberg. MR2427798

van Berkum, E. E. M., Linssen, H. N. and Overdijk, D. A. (1996). Inference rules and inferential distributions. *Journal of Statistical Planning and Inference* **49** 305-317. MR1381161

VIDAKOVIC, B. (2000). *Gamma-minimax: A paradigm for conservative robust Bayesians. Robust Bayesian Analysis* 241–260. Springer, New York. MR1795219

VON NEUMANN, J. and MORGENSTERN, O. (1953). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

WALD, A. (1961). *Statistical Decision Functions*. John Wiley and Sons, New York. MR0036976

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London. MR1145491

WEICHSELBERGER, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Physica-Verlag, Heidelberg.

WILLIAMS, P. M. (1980). Bayesian Conditionalisation and the Principle of Minimum Information. *The British Journal for the Philosophy of Science* **31** 131-144. MR0582834