Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: A large-deviation approach

David R. Bickel

December 31, 2015

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

Learning from model diagnostics that a prior distribution must be replaced by one that conflicts less with the data raises the question of which prior should be used for inference and decision. The same problem arises when a decision maker learns that one or more reliable experts express unexpected beliefs. In both cases, coherence of the solution would be guaranteed by formally stating the problem in terms suitable to using Bayes's theorem to condition on the insight that the prior distribution lies in a closed convex set potentially differing from that of the initial prior.

A readily available distribution of priors needed for such conditioning in the finite-sample setting is the law of the joint empirical distribution of infinitely many independent parameter values drawn from the initial prior. The solution is the prior distribution that minimizes the entropy relative to the initial prior according to the Gibbs conditioning principle from the theory of large deviations. While minimizing relative entropy accommodates the necessity of going beyond the initial prior without departing from it any more than the insight demands, its derivation from the Gibbs conditioning principle ensures the advantages of Bayesian coherence.

This approach is generalized to uncertain constraints by allowing the closed convex set of priors to be random. The distribution of that constraint set may in some cases arise from a confidence distribution such as the one that controls the probability of observing misleading statistical evidence.

Keywords: Bayesian model averaging; confidence distribution; maximum entropy updating; minimum cross entropy; minimum information principle; minimum relative entropy; model assessment; model checking; model criticism; posterior predictive check; prior predictive check

1 Introduction

As idealized agents, Bayesian decision makers minimize expected loss with respect to a distribution according to a Bayesian model that, in the usual statistics setting, includes a prior probability distribution P and a sampling model consisting of a family $\{f(\bullet|\theta) : \theta \in \Theta\}$ of probability functions, where Θ is the parameter space. If x is the observed value of $X \sim f(\bullet|\theta)$ with $\theta \sim P$, then Bayes's rule minimizes the expected loss with respect to the posterior probability distribution $P(\bullet|x)$.

A persistent problem in Bayesian statistics and decision making is that posed by an unexpected insight that requires revision of the prior distribution. The insight here goes beyond the observation that $X = x$ and may include machine learning. An artificial agent perceives a surprising report from one or more experts. A statistician perceives diagnostics indicating that P is obviously inadequate in light of its conflict with x or with other data. A physician perceives through reflection a new way to look at a puzzling disease. How should these agents proceed given the constraints their insights impose on the prior?

Suppose such a new insight requires a change in the prior to one in some set Γ of priors that are in some sense more adequate for inferential and decision-making purposes. The insight only constrains the new prior to the adequate set without in itself indicating one adequate prior over another. Expected loss is now to be minimized with respect to one of the posterior distributions corresponding to one of the priors in Γ. Which should be used for inference and decision?

Example 1. A decision maker (DM) held beliefs about the value of θ that were represented by P until becoming convinced to replace it with a prior probability distribution in the closed convex hull Γ of the prior distributions that represent the beliefs of domain experts. The DM lacks the time needed to formulate from scratch a model treating the experts' priors as observations and thus cannot apply the Bayesian methods reviewed in Cooke (1991, pp. 176-184). An alternative is that the DM does not have the resources to elicit the precise prior distribution of any expert but only some closed convex set Γ of probability mass functions generated from eliciting the relevant buying and selling prices that reflect beliefs held by a single expert (Augustin et al., 2014, Prop. 1.6). In either case, the DM needs to know which probability distribution in Γ to use for computing the expected loss of each contemplated action.

A DM who does not consider the m experts equally reliable as sources of information about θ may assign the ith prior a weight λ_i. The DM then has the insight that the prior suitable for decisions is in the closed

convex hull of the priors meeting a minimal adequacy α, that is, $\Gamma(\alpha) = \text{clco}\{P_i : \lambda_i \geq \alpha\}$. ▲

Example 2. Let $X \sim \text{N}(\theta, \sigma^2)$ denote a sample of n independent normal observations of unknown mean θ and known standard deviation σ. The prior distribution of θ is $\text{N}(\mu_0, \sigma_0^2)$. The posterior distribution of θ is $P_{\mu_0}(\bullet|X = x) = \text{N}(\mu_{\mu_0}(\bar{x}), \sigma^2(\bar{x}))$, where $\bar{x}$ is the observed sample mean, $\mu_{\mu_0}(\bar{x}) = (\mu_0 + n\bar{x}\sigma_0^2/\sigma^2) / (1 + n\sigma_0^2/\sigma^2)$, and $\sigma^2(\bar{x}) = \sigma_0^2/(1 + n\sigma_0^2/\sigma^2)$. Let χ_1^2 denote the χ^2 variate with 1 degree of freedom. A two-sided prior predictive p value testing $\text{N}(\mu_0, \sigma_0^2)$ is

$$p^x(\mu_0) = \text{Prob}_{\bar{X}^{\text{prior}} \sim \text{N}(\mu_0, \sigma^2/n + \sigma_0^2)} \left(\left(\bar{X}^{\text{prior}} - \mu_0\right)^2 \geq (\bar{x} - \mu_0)^2 \right) = \text{Prob}\left(\chi_1^2 \geq \frac{(\bar{x} - \mu_0)^2/\sigma^2}{1 + n\sigma_0^2/\sigma^2} n \right), \quad (1)$$

where x and $\bar{x}$ are the observed sample and its mean, and $\bar{X}^{\text{prior}}$ is the random sample mean of n independent observations drawn from the prior predictive distribution (Bickel, 2015a). In this case, $p^x(\mu_0)$ is equal to the calibrated posterior predictive p value of Hjort et al. (2006).

Rather than only considering $p^x(3.5)$, the p value checking the initial prior distribution $P = \text{N}(3.5, \sigma_0^2)$ (see Walter and Augustin, 2009), the statistician has the insight that the prior distribution is constrained to the closed convex hull of its observed $100(1 - \alpha)\%$ confidence set, an example of an *α-adequate set of prior distributions* (Bickel, 2015b):

$$\Gamma(\alpha; x) = \text{clco}\left\{ \text{N}(\mu_0, \sigma_0^2) : p^x(\mu_0) \geq \alpha, \mu_0 \in \mathbb{R} \right\}. \quad (2)$$

▲

Examples 1 and 2 respectively present subjectively assigned weights and prior-predictive p values as straightforward quantities for generating adequate sets of prior distributions for inference in light of an insight. Adequate sets can alternatively be defined using posterior-predictive p values, Bayes factors, likelihood ratios, proper scoring rules, or other algorithmic means of model assessment or may be specified by human judgment rather than by any formal rule (Bickel, 2015b, §2.2, §A.1). Another possible measure of the adequacy of a prior is how weakly informative it is (see Evans, 2015). Regardless of how a set of distributions comes to be regarded as more adequate than the initial distribution, the problem of subsequent inference and decision remains.

One solution is to apply decision rules from imprecise probability and robust Bayes theory (Bickel,

2015b, §3.2.2, §A.2.2). Unfortunately, rules of that type can lead to inconsistencies in actions that could be avoided if there were only one posterior distribution (Elga, 2010). That suggests combining the prior distributions in Γ into a single distribution for purposes of inference and action (Bickel, 2015b, §3.2.1, §A.2.1). While the combination of distributions may be simple, non-Bayesian methods of combination often run into inconsistencies similar to those of Elga (2010).

This paper presents a general Bayesian solution to the problem of inference and decision after the insight constraining the prior to Γ. Without requiring large samples, this solution takes advantage of the *Gibbs conditioning principle* from the theory of large deviations (e.g. Dembo and Zeitouni, 2009). Mathematically stated as various conditional limit theorems (e.g., Cover and Thomas, 2006), it says a conditional probability distribution, given that an empirical distribution is in Γ, approaches the distribution that minimizes the entropy over Γ relative to the initial distribution. Like the approach of Skyrms (1985), this solution moves in the opposite direction as those deriving conditional probability from minimizing the relative entropy. For the latter direction, see Williams (1980) and Paris (1994, pp. 119-120) on discrete distributions and Diaconis and Zabell (1982), Caticha and Giffin (2006), and Harremoës (2007, Exa. 3) for more general proofs. These approaches, like those based on general classes of loss functions (e.g., Grünwald and Dawid, 2004), do not uniquely lead to minimizing relative entropy (Diaconis and Zabell, 1982).

Section 2 departs from previous Bayesian applications of large-deviation theory (Csiszár, 1985; Grendar Jr. and Grendar, 2004; Halpern, 2003, §11.5) chiefly by distinguishing the sample size, which remains finite, from the large-deviation distribution index, which diverges. This is accomplished by noting that distributions are equivalent for inferential purposes if they may be made arbitrarily close to each other without requiring additional data. Such distributions are practically equivalent in that they yield the same actions when minimizing expected loss.

Based on the notion of practical equivalence with Bayesian conditionalization, this version of minimizing relative entropy (a) complies with the insight that the prior lies in Γ, (b) retains the initial prior in the degenerate case in which nothing new was learned from the insight, and (c) commutes with updating a prior to a posterior, as seen in Section 3. While the first two results resemble those in the literature on desirable properties of maximum entropy (e.g., Shore and Johnson, 1980; Csiszár, 1991; Paris, 1994, pp. 120-122; Caticha and Giffin, 2006; Baez and Fritz, 2014), the third holds for the large-deviation version under broader conditions than the affine-constraint condition required to guarantee the commutativity of

the more direct version of minimizing relative entropy (Williams, 1980; Csiszár, 1991). Other widely used approaches either suffer from a lack of coherence in the strict Bayesian sense or fail to satisfy the featured properties of minimizing relative entropy. First, the arithmetic average of adequate distributions fails to depend on the original prior, replacing it even when it satisfies all the constraints imposed by the insight. Second, the normalized geometric average of distributions is not in general a member of the set of adequate distributions being combined. When that average of adequate distributions is not itself adequate, it is incompatible with the constraints the new insight places on the prior. Some shortcomings are overcome by Bayesian model averaging, given a hyperprior over the priors in Γ, the set of adequate distributions. That, however, can fail to retain the initial prior when it already satisfies the constraint imposed by the insight. Those problems with the above methods of inference on the basis of Γ stem from ignoring the considerable information available in P, the initial prior distribution.

An idealization involved in the proposed method of minimum-relative-entropy updating is the sharp distinction between adequacy and inadequacy on the basis of the choice of the adequacy threshold. Example 1 treats barely adequate expert opinions the same as the most adequate and barely inadequate opinions the same as the most inadequate. Likewise, Example 2 considers priors below the significance threshold as inadequate and all others as adequate even though those just above the threshold may have p values very close to those just below it. As the resulting prior distributions that minimize relative entropy may only poorly approximate the beliefs of the DM or other agent, Section 4 generalizes the proposed approach to reflect uncertainty in the threshold between adequacy and inadequacy. That enables the distribution used for inference and decision to incorporate continuous degrees of adequacy rather than only the two categories, inadequate and adequate.

2 Inference given an insight about the prior

2.1 Preliminary concepts

Consider a metric space Θ and the measurable spaces $(\Theta, \mathfrak{F}_\Theta)$ and $(\mathcal{X}, \mathfrak{F}_\mathcal{X})$, where Θ is called a *parameter space* and $\mathcal{X}$ a *sample space*. For example, $\mathcal{X} = \mathbb{R}^n$ if x is a vector sample of n real observations.

For any set Ω, let $\mathcal{P}_\Omega$ denote a metric space of all probability measures on a measurable space $(\Omega, \mathfrak{F}_\Omega)$ and $\mathcal{C}_\Omega$ the set of all bounded continuous real-valued functions on Ω. $\mathcal{P}_\Theta$ is endowed with the *narrow topology*

(Crauel, 2002) defined as the topology generated by the family of open sets of the form

$$\mathcal{P}_{P'}(h_1,\dots h_k;\varepsilon_1,\dots,\varepsilon_k)=\left\{P''\in\mathcal{P}_\Theta:\left|\int h_i(\theta)\,dP'(\theta)-\int h_i(\theta)\,dP''(\theta)\right|<\varepsilon_i, i=1,\dots,k\right\}$$

for all $P'\in\mathcal{P}_\Theta$, $h_1,\dots h_k\in\mathcal{C}_\Theta$, $k=1,2,\dots$ (Dudley, 2002, p. 40). A sequence of probability measures $\{P_N:N=1,2,\dots\}$ in $\mathcal{P}_\Theta$ is said to *weakly converge* to a probability measure $P_\infty\in\mathcal{P}_\Theta$ $\left(P_N\xrightarrow{\text{weak}}P_\infty\right)$ if

$$\lim_{n\to\infty}\int h_\Theta(\theta)\,dP_N(\theta)=\int h_\Theta(\theta)\,dP_\infty(\theta)$$

for all $h_\Theta\in\mathcal{C}_\Theta$ (e.g., Billingsley, 1999).

For every $\theta\in\Theta$, consider F_θ as a different distribution in $\mathcal{P}_\mathcal{X}$, and $f_\theta=dF_\theta/d\eta$ as the Radon-Nikodym derivative of F_θ with respect to a measure η that dominates $\mathcal{X}$. Suppose the *likelihood function* $\theta\mapsto f_\theta(x)$ is in $\mathcal{C}_\Theta$ for any $x\in\mathcal{X}$.

Fix the *decision space* $\mathcal{D}$ as the set of possible decisions. A *bounded continuous loss function* on Θ and the decision space $\mathcal{D}$ is a bounded function $\ell:\Theta\times\mathcal{D}\to[0,\infty[$ such that $\theta\mapsto\ell(\theta,\Delta)$ is continuous on Θ. The set of all such functions is denoted by $\mathcal{C}_{\Theta,\mathcal{D}}$. Thus, $\{\theta\mapsto\ell(\theta,\Delta):\ell\in\mathcal{C}_{\Theta,\mathcal{D}}\}=\mathcal{C}_\Theta$ for all $\Delta\in\mathcal{D}$.

The Bayesian updating rule is stated as a map transforming a prior distribution.

Definition 1. For any *prior distribution* $P\in\mathcal{P}_\Theta$, the *posterior distribution* given the observation that $x\in\mathcal{X}$, denoted by $P(\bullet|X=x)$, is the distribution in $\mathcal{P}_\Theta$ such that, for any $\mathcal{H}\in\mathfrak{F}_\Theta$,

$$P(\mathcal{H}|X=x)=\frac{\int f_\theta(x)\,dP(\theta|\mathcal{H})}{\int f_\theta(x)\,dP(\theta)}P(\mathcal{H}). \tag{3}$$

The *conditionalization on* $X=x$ is the map $u_x:\mathcal{P}_\Theta\to\mathcal{P}_\Theta$ equal to $P(\bullet|X=x)$ of equation (3) as a function of P, that is, $P\mapsto u_x(P)=P(\bullet|X=x)$. Abusing the notation, the map $u_x:\mathcal{P}_{\mathcal{P}_\Theta}\to\mathcal{P}_{\mathcal{P}_\Theta}$ satisfying

$$\mathcal{P}_{\mathcal{P}_\Theta}\ni Q\mapsto u_x(Q)=\text{Prob}(\theta\in\bullet|X=x)=:Q(\bullet|X=x) \tag{4}$$

is also a conditionalization on $X=x$, where $X\sim f_\theta$ *and* $\theta\sim Q$.

Thus, the conditionalization on $X=x$ is the function that transforms a prior distribution to a posterior distribution according to Bayes's theorem. Consider an index $N=1,2,\dots$, that, unlike the sample size n,

may be increased at will until achieving the desired precision.

Lemma 1. *If $P_1, P_2, \ldots$ and P_∞ are probability measures in $\mathcal{P}_\Theta$ such that $P_N \xrightarrow{weak} P_\infty$, then*

$$\lim_{n\to\infty} \int h_\Theta(\theta)\, dP_N(\theta|X=x) = \int h_\Theta(\theta)\, dP_\infty(\theta|X=x)$$

for all $h_\Theta \in \mathcal{C}_\Theta$ and $x \in \mathcal{X}$.

Proof. Letting h_Θ denote any function in $\mathcal{C}_\Theta$,

$$\lim_{N\to\infty} \int h_\Theta(\theta)\, dP_N(\theta|X=x) = \lim_{N\to\infty} \frac{\int h_\Theta(\theta) f_\theta(x)\, dP_N(\theta)}{\int f_\theta(x)\, dP_N(\theta)} = \lim_{N\to\infty} \frac{\int h_{\Theta,x}(\theta)\, dP_N(\theta)}{\int f_\theta(x)\, dP_N(\theta)},$$

where $\theta \mapsto h_{\Theta,N}(\theta) = h_\Theta(\theta) f_\theta(x)$. By the definition of weak convergence, $P_N \xrightarrow{\text{weak}} P_\infty$ and $(\theta \mapsto f_\theta(x)) \in \mathcal{C}_\Theta$ imply $\int f_\theta(x)\, dP_N(\theta) \to \int f_\theta(x)\, dP_\infty(\theta)$. From $h_\Theta, (\theta \mapsto f_\theta(x)) \in \mathcal{C}_\Theta$, it follows that $h_{\Theta,x} \in \mathcal{C}_\Theta$ and consequently $\int h_{\Theta,x}(\theta)\, dP_N(\theta) \to \int h_{\Theta,x}(\theta)\, dP_\infty(\theta)$ also follows from $P_N \xrightarrow{\text{weak}} P_\infty$. Thus,

$$\lim_{N\to\infty} \int h_\Theta(\theta)\, dP_N(\theta|X=x) = \frac{\lim_{N\to\infty} \int h_{\Theta,x}(\theta)\, dP_N(\theta)}{\lim_{N\to\infty} \int f_\theta(x)\, dP_N(\theta)} = \frac{\int h_\Theta(\theta) f_\theta(x)\, dP_\infty(\theta)}{\int f_\theta(x)\, dP_\infty(\theta)}.$$

□

2.2 Practical equivalence and coherent updating

Revising the prior is equivalent to revising the posterior under the working assumption that the sampling model remains unchanged. An equivalence relation formalizes the idea that prior probability distributions are in practice equivalent for purposes of data analysis if they yield arbitrarily close prior and posterior expected losses.

Definition 2. A sequence $\{P_N : N = 1, 2, \ldots\}$ of probability measures in $\mathcal{P}_\Theta$ is *practically equivalent* to a probability measure $P \in \mathcal{P}_\Theta$ if, for all $\ell \in \mathcal{C}_{\Theta,\mathcal{D}}$,

$$\lim_{N\to\infty} \int \ell(\theta, \Delta)\, dP_N(\theta) = \int \ell(\theta, \Delta)\, dP(\theta) \tag{5}$$

$$\lim_{N\to\infty}\int \ell(\theta,\Delta)\,dP_N(\theta|X=x)=\int \ell(\theta,\Delta)\,dP(\theta|X=x) \tag{6}$$

for every $\Delta\in\mathcal{D}$. Two sequences of probability measures in $\mathcal{P}_\Theta$ are practically equivalent to each other if each is practically equivalent to the same probability measure in $\mathcal{P}_\Theta$.

Recall that N is arbitrarily large without requiring any additional data. Weak convergence is necessary and sufficient for practical equivalence.

Lemma 2. *A sequence $\{P_N : N=1,2,\dots\}$ of probability measures in $\mathcal{P}_\Theta$ weakly converges to a probability measure $P_\infty\in\mathcal{P}_\Theta$ if and only if $\{P_N : N=1,2,\dots\}$ is practically equivalent to P_∞.*

Proof. ($\Longrightarrow$). Suppose $P_N \xrightarrow{\text{weak}} P_\infty$. The following statements hold for every $\Delta\in\mathcal{D}$ and every $\ell\in\mathcal{C}_{\Theta,\mathcal{D}}$. Since the function ℓ_Δ on Θ that is defined by $\theta\mapsto\ell_\Delta(\theta)=\ell(\theta,\Delta)$ is continuous, bounded, and real-valued, the definition of weak convergence yields

$$\lim_{N\to\infty}\int \ell(\theta,\Delta)\,dP_N(\theta)=\lim_{n\to\infty}\int \ell_\Delta(\theta)\,dP_N(\theta)=\int \ell_\Delta(\theta)\,dP_\infty(\theta)=\int \ell(\theta,\Delta)\,dP_\infty(\theta),$$

establishing equation (5). Similarly, equation (6) then follows from Lemma 1.

($\Longleftarrow$). Assume $\{P_N : N=1,2,\dots\}$ is practically equivalent to P_∞. For any $\Delta\in\mathcal{D}$, equation (5) holds for all $\ell(\bullet,\Delta)\in\mathcal{C}_\Theta$. Thus, $P_N \xrightarrow{\text{weak}} P_\infty$ by the definition of weak convergence. □

The concept of coherently learning from a new insight distinguishes between, on one hand, the conditioning that updates the prior given the new insight about which priors are adequate and, on the other hand, the conditioning on x that transforms a prior to a posterior.

Definition 3. Consider a set $\mathfrak{G}\subset\mathfrak{F}_{\mathcal{P}_\Theta}$ of possible adequate sets of prior distributions, each of which consists of probability distributions on $(\Theta,\mathfrak{F}_\Theta)$. The sequence $\{Q_{N,P} : N=1,2,\dots\}$ of probability measures in $\mathcal{P}_{\mathcal{P}_\Theta}$, is a *coherent updater* of P under $\mathfrak{G}$ if $Q_{N,P}$ meets these conditions:

1. *Pre-updating equivalence.* The sequence $\left\{\int \pi(\bullet)\,dQ_{N,P}(\pi) : N=1,2,\dots\right\}$ of mixture distributions is practically equivalent to P.

2. *Updating equivalence.* For every $\Gamma \in \mathfrak{G}$, the sequence $\left\{\int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma) : N = 1, 2, \dots\right\}$ of mixtures of conditional distributions is practically equivalent to some distribution $P_\Gamma \in \mathcal{P}_\Theta$, which is called the *coherent update* of P under Γ and $\{Q_{N,P} : N = 1, 2, \dots\}$.

Every sequence of distributions in $\mathcal{P}_\Theta$ that is practically equivalent to a coherent update of P under Γ is itself a coherent update of P under Γ.

The main idea is that coherence between the initial prior and the insight that inference should instead be based on a more adequate prior is achieved by introducing the coherent updater. By definition, it agrees with the initial prior distribution and posterior distribution when disregarding the insight (Condition 1) and yet, conditional on the insight that only the prior distributions in Γ are adequate, also agrees with with the coherent update's prior distribution and its posterior distribution (Condition 2). With Lemma 2, Conditions 1 and 2 may be abbreviated as $\int \pi(\bullet)\, dQ_{N,P}(\pi) \xrightarrow{\text{weak}} P$ and $\int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma) \xrightarrow{\text{weak}} P_\Gamma$, respectively.

For any $(\theta_1, \dots, \theta_N) \in \Theta^N$, the *type* is the distribution $L_{(\theta_1,\dots,\theta_N)}$ on $(\Theta, \mathfrak{F}_\Theta)$ such that, for all $\mathcal{H} \in \mathfrak{F}_\Theta$,

$$L_{(\theta_1,\dots,\theta_N)}(\mathcal{H}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i}(\mathcal{H}) = \frac{1}{N}\sum_{i=1}^{N} 1_{\mathcal{H}}(\theta_i),$$

where $\theta_1 \mapsto 1_{\mathcal{H}}(\theta_1)$ is the indicator function. Thus, if $(\theta_1, \dots, \theta_N)$ were observed, $L_{(\theta_1,\dots,\theta_N)}$ would be the empirical distribution, but the type is only considered here as a mathematical device.

The following result says updating a sequence by conditioning on a constraint on the type is coherent.

Lemma 3. *For any $P \in \mathcal{P}_\Theta$, let P^N denote the distribution of the vector $(\theta_1, \dots, \theta_N)$, where $\theta_1, \dots, \theta_N \sim P$ are independent. Consider a nonempty $\Gamma \in \mathfrak{F}_{\mathcal{P}_\Theta}$. Assume that the sequence $\left\{\bar{L}_{N,P,\Gamma} : N = 1, 2, \dots\right\}$ that is defined by the mixture distribution*

$$\bar{L}_{N,P,\Gamma}(\bullet) = \int L_{(\theta_1,\dots,\theta_N)}(\bullet)\, d\, P^N\left(\theta_1, \dots, \theta_N | L_{(\theta_1,\dots,\theta_N)} \in \Gamma\right) \tag{7}$$

exists and weakly converges to some distribution $\bar{L}_{\infty,P,\Gamma}$ on $(\Theta, \mathfrak{F}_\Theta)$. Then $\bar{L}_{\infty,P,\Gamma}$ is a coherent update of P under Γ.

Proof. Consider an arbitrary function $h_\Theta \in \mathcal{C}_\Theta$. Let $Q_{N,P}$ denote the distribution of $L_{(\theta_1,\dots,\theta_N)}$, that is,

$$Q_{N,P}(\mathcal{L}) = P^N\left(L_{(\theta_1,\dots,\theta_N)} \in \mathcal{L}\right)$$

for all $\mathcal{L} \in \mathfrak{F}_{\mathcal{P}_\Theta}$ and $N = 1, 2, \ldots$, and let $\bar{\pi}_N = \int \pi(\bullet)\, dQ_{N,P}(\pi)$ and $\bar{\pi}_{N,\Gamma} = \int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma)$. By substitution,

$$\begin{aligned}\int h_\Theta(\theta)\, d\bar{\pi}_N(\theta) &= \int\int h_\Theta(\theta)\, d\pi(\theta)\, dQ_{N,P}(\pi) \\ &= \int\int h_\Theta(\theta)\, dL_{(\theta_1,\ldots,\theta_N)}(\theta)\, d\,P^N(\theta_1,\ldots,\theta_N).\end{aligned}$$

Thus, by the weak law of large numbers and the invariance of $L_{(\theta_1,\ldots,\theta_N)}$ to permutations of $\theta_1, \ldots, \theta_N$ (Harremoës, 2007, pp. 140-141),

$$\lim_{N\to\infty} \int h_\Theta(\theta)\, d\bar{\pi}_N(\theta) = \int h_\Theta(\theta)\, d\,P(\theta).$$

Since h_Θ is arbitrary, it follows that $\{\bar{\pi}_N : N = 1, 2\ldots\}$ weakly converges to P, which with Lemma 2 establishes the practical equivalence of $\{\bar{\pi}_N : N = 1, 2\ldots\}$ to P. That means Condition 1 of Definition 3 is met. Similarly,

$$\begin{aligned}\int h_\Theta(\theta)\, d\bar{\pi}_{N,\Gamma}(\theta) &= \int\int h_\Theta(\theta)\, d\pi(\theta)\, dQ_{N,P}(\pi|\Gamma) \\ &= \int\int h_\Theta(\theta)\, dL_{(\theta_1,\ldots,\theta_N)}(\theta)\, d\,P^N\left(\theta_1,\ldots,\theta_N | L_{(\theta_1,\ldots,\theta_N)} \in \Gamma\right) \\ &= \int h_\Theta(\theta)\, d\bar{L}_{N,P,\Gamma}(\theta).\end{aligned}$$

Thus, $\{\bar{\pi}_{N,\Gamma} : N = 1, 2, \ldots\}$ weakly converges to $\bar{L}_{\infty,P,\Gamma}$, satisfying Condition 2 of Definition 3 according to Lemma 2 and thereby proving $\bar{L}_{\infty,P,\Gamma}$ to be a coherent update of P. □

2.3 Coherence of minimum relative entropy

For any $P', P \in \mathcal{P}_\Theta$, the Kullback-Leibler divergence between them or the *entropy* of P' *relative to* P is

$$D(P'||\,P) = \int \ln\left(\frac{dP'}{d\,P}(\theta)\right) dP'(\theta)$$

if P' is absolutely continuous with respect to P but $D(P'||\,P) = \infty$ otherwise (e.g., Rassoul-Agha and Seppäläinen, 2015, p. 68). Any distribution $P^\star_\Gamma \in \mathcal{P}_\Theta$ is a *minimum-relative-entropy distribution* over some

nonempty $\Gamma \in \mathfrak{F}_{\mathcal{P}_\Theta}$ relative to a $P \in \mathcal{P}_\Theta$ if

$$D\left(P_\Gamma^\star || P\right) = \inf_{P' \in \Gamma} D\left(P' || P\right). \tag{8}$$

The term *maximum entropy* is also used in this context, having originated from the proportionality of $-D\left(P' || P\right)$ to Shannon entropy in the case that Θ is finite and P is the uniform distribution on $\left(\Theta, 2^\Theta\right)$.

Theorem 1. *If $\Gamma \in \mathfrak{F}_{\mathcal{P}_\Theta}$ is closed, convex, and of non-empty interior $\operatorname{int}\Gamma$ such that*

$$\inf_{P' \in \Gamma} D\left(P' || P\right) = \inf_{P' \in \operatorname{int}\Gamma} D\left(P' || P\right) < \infty, \tag{9}$$

then $P_\Gamma^\star$, the minimum-relative-entropy distribution over Γ relative to P, exists, is unique, and is a coherent update of P under the adequate set Γ. Further, $Q_{N,P}\left(\bullet|\Gamma\right) = P^N\left(L_{(\theta_1,\ldots,\theta_N)} \in \bullet | L_{(\theta_1,\ldots,\theta_N)} \in \Gamma\right)$, the conditional distribution on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$ of $L_{(\theta_1,\ldots,\theta_N)}$ given $L_{(\theta_1,\ldots,\theta_N)} \in \Gamma$ with $(\theta_1,\ldots,\theta_N) \sim P^N$, weakly converges to $\delta_{P_\Gamma^\star}$, the Dirac measure on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$ with support at $P_\Gamma^\star$.

Proof. From the stated conditions, Rassoul-Agha and Seppäläinen (2015, pp. 79-80) used the convexity of $P' \mapsto D\left(P' || P\right)$ to prove the existence and uniqueness of the maximum entropy distribution $P_\Gamma^\star$ and used Sanov's theorem from large deviation theory and the portmanteau theorem to prove:

1. $P^N\left(\theta_1 \in \bullet | L_{(\theta_1,\ldots,\theta_N)} \in \Gamma\right)$, the conditional distribution of θ_1 given $L_{(\theta_1,\ldots,\theta_N)} \in \Gamma$ with $(\theta_1,\ldots,\theta_N) \sim P^N$, weakly converges to $P_\Gamma^\star$, that is,

$$\lim_{n\to\infty} \int h_\Theta\left(\theta_1\right) d\, P^N\left(\theta_1,\ldots,\theta_N | L_{(\theta_1,\ldots,\theta_N)} \in \Gamma\right) = \int h_\Theta\left(\theta\right) dP_\Gamma^\star\left(\theta\right) \tag{10}$$

   for all $h_\Theta \in \mathcal{C}_\Theta$.

2. $Q_{N,P}\left(\bullet|\Gamma\right) \xrightarrow{\text{weak}} \delta_{P_\Gamma^\star}$.

Using the invariance of $L_{(\theta_1,\ldots,\theta_N)}$ to permutations of $\theta_1,\ldots,\theta_N$ (Harremoës, 2007, pp. 140-141), equation (10) establishes that $\left\{\bar{L}_{N,P,\Gamma} : N = 1, 2, \ldots\right\}$, as defined by equation (7), also weakly converges to $P_\Gamma^\star$, abbreviated by $\bar{L}_{N,P,\Gamma} \xrightarrow{\text{weak}} P_\Gamma^\star$. Thus, by Lemma 3, $P_\Gamma^\star$ is a coherent update of P. □

Example 3. Theorem 1 applies to Example 2 since its $\Gamma(\alpha; x)$ satisfies $\operatorname{int}\Gamma(\alpha; x) \neq \emptyset$ and is a closed convex set of distributions (2) with finite relative entropy from the initial prior distribution, $P = \mathrm{N}\left(3.5, \sigma_0^2\right)$. By equation (1),

$$\begin{aligned} P^{\star}_{\Gamma(\alpha;x)} &= \arg \inf_{P' \in \Gamma(\alpha;x)} D\left(P' || \,\mathrm{N}\left(3.5, \sigma_0^2\right)\right) \\ &= \begin{cases} \mathrm{N}\left(\mathrm{CDF}_{\bar{x}}^{-1}\left(\alpha/2\right), \sigma_0^2\right) & \text{if } \mathrm{CDF}_{\bar{x}}^{-1}\left(\alpha/2\right) > 3.5 \\ \mathrm{N}\left(\mathrm{CDF}_{\bar{x}}^{-1}\left(1-\alpha/2\right), \sigma_0^2\right) & \text{if } \mathrm{CDF}_{\bar{x}}^{-1}\left(1-\alpha/2\right) < 3.5\,, \\ \mathrm{N}\left(3.5, \sigma_0^2\right) & \text{if } p^{x}\left(\mathrm{N}\left(3.5, \sigma_0^2\right)\right) \geq \alpha \end{cases} \end{aligned} \tag{11}$$

where $\mathrm{CDF}_{\bar{x}}$ is the distribution function of $\mathrm{N}\left(\bar{x}, \sigma^2/n + \sigma_0^2\right)$, which originates from $\mathrm{N}\left(\mu_0, \sigma^2/n + \sigma_0^2\right)$ as the law of $\bar{X}^{\mathrm{prior}}$ (Bickel, 2015a). ▲

3 Comparisons to other updating methods

3.1 Updating methods

Recalling $\mathfrak{G}$ from Definition 3, an *updater* of the initial prior P is a sequence $Q_{N,P}$ of probability measures on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$ that is equipped with an *updating method*, a function $v_{\bullet}(\bullet) : \mathfrak{G} \times \mathcal{P}_{\mathcal{P}_\Theta} \to \mathcal{P}_{\mathcal{P}_\Theta}$ that transforms a constraint set $\Gamma \in \mathfrak{G}$ and each member of the updater to a probability measure on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$. If $\left\{\int \pi(\bullet)\, d\left(v_\Gamma\left(Q_{N,P}\right)\right)(\pi) : N = 1, 2, \dots\right\}$ is practically equivalent to some $P_\Gamma \in \mathcal{P}_\Theta$, then P_Γ is called the *update* of P under Γ.

For example, the coherent updater $\{Q_{N,P} : N = 1, 2, \dots\}$ of Definition 3 is an updater equipped with the updating method

$$(\Gamma, Q_{N,P}) \mapsto v_\Gamma\left(Q_{N,P}\right) = Q_{N,P}\left(\bullet|\Gamma\right), \tag{12}$$

which generates the coherent update as the update. That $(\Gamma, Q_{N,P}) \mapsto v_\Gamma\left(Q_{N,P}\right)$ is called a *coherent updating method*.

Other methods of combining distributions (Genest and Zidek, 1986) also qualify as updating methods even though they do not depend on P. A simple updater is the constant sequence $\{\lambda : N = 1, 2, \dots\}$ equipped

with this generalization of the *linear opinion pool* (McConway, 1981):

$$(\Gamma, \lambda) \mapsto v_\Gamma(\lambda) = \lambda(\bullet|\Gamma), \tag{13}$$

where λ is a probability measure on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$. The corresponding generalization of the m-expert logarithmic opinion pool (McConway, 1981; Genest et al., 1986) uses the constant updater $\{\delta_P : N = 1, 2, \dots\}$ equipped with the updating method

$$(\Gamma, \delta_P) \mapsto v_\Gamma(\delta_P) = \delta_{P_\Gamma^{\log}}; \tag{14}$$

$$P_\Gamma^{\log} = \frac{\prod_{1,\dots,m} \pi_i^{\lambda_i}}{\int d\left(\prod_{1,\dots,m} \pi_i^{\lambda_i}\right)(\theta)}$$

where $\delta_{P_\Gamma^{\log}}$ is the Dirac measure on $(\mathcal{P}_\Theta, \mathfrak{F}_{\mathcal{P}_\Theta})$ with support at $P_\Gamma^{\log}$, π_i is the ith of the m probability distributions in Γ corresponding to expert opinions, and $\lambda_i \in [0, 1]$ is its weight $\left(\sum_{1,\dots,m} \lambda_i = 1\right)$.

An updater $\{Q_{N,P} : N = 1, 2, \dots\}$ equipped with an updating method $v_\bullet(\bullet)$ *minimizes the relative entropy* if $v_\Gamma(Q_{N,P})$ is practically equivalent to $\delta_{P_\Gamma^\star}$ for all $\Gamma \in \mathfrak{G}$, with $\delta_{P_\Gamma^\star}$ defined as in Theorem 1. The simplest example is the direct version of minimizing relative entropy, in which the constant updater $\{\delta_P : N = 1, 2, \dots\}$ is equipped with the updating method

$$(\Gamma, \delta_P) \mapsto v_\Gamma(\delta_P) = \delta_{P_\Gamma^\star}. \tag{15}$$

Desirable properties of an updating method are invariance to the order in which information is processed (§3.2), agreement with the insight that the adequate priors are in Γ (§3.3), and agreement with the initial prior P when $P \in \Gamma$ (§3.4). All properties describe the version of minimum relative entropy derived from large-deviation theory (§3.5) as opposed to the version in equation (15).

3.2 Commutativity between observations and insights

An important property of a method of updating a prior is that it not depend on whether the new insight about which priors are adequate is applied before or after conditioning on the observation x to transform a prior distribution to a posterior distribution. This commutativity property requires that the insight-updated posterior is practically equivalent to the posterior corresponding to an insight-updated prior.

Let $\{Q_{N,P} : N = 1, 2, \dots\}$ denote an updater equipped with some $v_\bullet(\bullet)$ as its updating method. In terms of Definition 1, $v_\bullet(\bullet)$ *commutes* with a conditionalization u_x if $u_x(v_\Gamma(Q_{N,P})) = v_\Gamma(u_x(Q_{N,P}))$ for all $x \in \mathcal{X}$ and $\Gamma \in \mathfrak{G}$.

Commutativity is satisfied by the direct minimization of relative entropy (15) if and only if

$$\delta_{u_x\left(P^\star_\Gamma\right)} = u_x\left(\delta_{P^\star_\Gamma}\right) = u_x\left(v_\Gamma\left(\delta_P\right)\right) = v_\Gamma\left(u_x\left(\delta_P\right)\right) = v_\Gamma\left(\delta_{u_x(P)}\right) = \delta_{(u_x(P))^\star_\Gamma}$$

according to equation (4). This notion of commutativity thus reduces in this case to the $u_x\left(P^\star_\Gamma\right) = \left(u_x\left(P\right)\right)^\star_\Gamma$ proven to hold under affine constraints (Csiszár, 1991), which are stronger than the requirement that Γ be closed and convex (Williams, 1980).

Likewise, the logarithmic opinion pool (14) satisfies commutativity if and only if

$$\delta_{u_x\left(P^{\log}_\Gamma\right)} = u_x\left(\delta_{P^{\log}_\Gamma}\right) = u_x\left(v_\Gamma\left(\delta_P\right)\right) = v_\Gamma\left(u_x\left(\delta_P\right)\right) = v_\Gamma\left(\delta_{u_x(P)}\right) = \delta_{(u_x(P))^{\log}_\Gamma},$$

equivalent to $u_x\left(P^{\log}_\Gamma\right) = \left(u_x\left(P\right)\right)^{\log}_\Gamma$, which is met under broad conditions (Genest et al., 1986). While the linear opinion pool (13) only satisfies $\int \left(u_x\pi\right)(\bullet)\, d\lambda\left(\pi|\mathcal{P}_\Theta\right) = u_x\left(\int \pi\left(\bullet\right) d\lambda\left(\pi|\mathcal{P}_\Theta\right)\right)$ in degenerate cases (Genest and Zidek, 1986), it nonetheless generally commutes with conditionalization:

$$u_x\left(v_\Gamma\left(\lambda\right)\right) = u_x\left(\lambda\left(\bullet|\Gamma\right)\right) = \lambda\left(\bullet|\Gamma, X = x\right) = \left(u_x\left(\lambda\right)\right)\left(\bullet|\Gamma\right) = v_\Gamma\left(u_x\left(\lambda\right)\right).$$

3.3 Insight-receptive updating

A method of updating a prior in light of the insight that the adequate priors are in Γ should result in a prior lying in Γ for all practical purposes. Accordingly, an updating method $v_\bullet(\bullet)$ is *insight-receptive* if $P_\Gamma \in \operatorname{Limpts}\Gamma$ for all $\Gamma \in \mathfrak{G}$, where P_Γ is the update from $v_\bullet(\bullet)$, and $\operatorname{Limpts}\Gamma$ is the set of all limit points of Γ. For example, the linear opinion pool (13) is insight-receptive given the convexity of Γ. The policy of retaining the initial prior P even when $P \notin \operatorname{Limpts}\Gamma$ is obviously not insight-receptive.

That the logarithmic opinion pool can also fail to be insight-receptive is seen from a counterexample.

Example 4. Applying the updating method of equation (14) to Example 1, suppose $\Theta = \{0, 1, 2\}$, $m = 2$, $\pi_1\left(\{0\}\right) = \pi_2\left(\{0\}\right) = 1/3$, $\pi_1\left(\{1\}\right) = 0$, $\pi_2\left(\{1\}\right) = 2/3$, $\pi_1\left(\{2\}\right) = 2/3$, $\pi_2\left(\{2\}\right) = 0$, $\lambda_1 = \lambda_2 = 1/2$, and

$\Gamma = \operatorname{clco}\{\pi_1, \pi_2\}$, which is the closed convex hull of π_1 and π_2. Combining the two distributions that each assigns probability $1/3$ to $\theta = 0$ yields a very different probability that $\theta = 0$:

$$\begin{aligned} P_\Gamma^{\log}(\{0\}) &= \frac{\sqrt{\pi_1(\{0\})\,\pi_2(\{0\})}}{\sqrt{\pi_1(\{0\})\,\pi_2(\{0\})} + \sqrt{\pi_1(\{1\})\,\pi_2(\{1\})} + \sqrt{\pi_1(\{2\})\,\pi_2(\{2\})}} \\ &= \frac{1/3}{1/3 + 0 + 0} = 1 \notin [1/3, 1/3] = [\pi_1(\{0\}), \pi_2(\{0\})]. \end{aligned}$$

From $P_\Gamma^{\log}(\{0\}) \notin [\pi_1(\{0\}), \pi_2(\{0\})]$, it is clear that $P_\Gamma^{\log} \notin \Gamma$. It follows that $P_\Gamma^{\log} \notin \operatorname{Limpts}\Gamma$. ▲

3.4 Prior-persistent updating

A method of updating a prior in light of the insight that the adequate priors are in Γ should retain the initial prior P if $P \in \Gamma$ since the insight provides no new information. As Paris (2014) put it, "For surely I should not be in the position of having to change my beliefs on receipt of something I already believed?!"

An updating method $v_\bullet(\bullet)$ that equips an updater $\{Q_{N,P} : N = 1, 2, \dots\}$ is *prior-persistent* if $v_\Gamma(Q_{N,P}) = Q_{N,P}$ for all $N = 1, 2, \dots$ and $P \in \Gamma$. This rules out all updating methods that do not depend on the initial prior P, including the linear opinion pool defined by equation (13) (Paris, 2014, p. 6191), the logarithmic opinion pool defined by equation (14), and the usual approaches to Bayesian model averaging.

3.5 Satisfaction of commutativity, insight-receptiveness and prior-persistence

The next result says prior persistence, insight-receptiveness, and commutativity are weaker than practical coherence, which is weaker than minimizing relative entropy.

Theorem 2. *If $\mathfrak{G}$ is a class of the closed, convex sets in $\mathfrak{F}_{\mathcal{P}_\Theta}$ that are of non-empty interior and that satisfy equation* (9) *such that $u_x\Gamma \in \mathfrak{G}$ for all $\Gamma \in \mathfrak{G}$, then $v_\bullet(\bullet)$ is a coherent updating method. If $v_\bullet(\bullet)$ is a coherent updating method, then it is prior-persistent and insight-receptive, and it commutes with the conditionalization u_x for any $x \in \mathcal{X}$.*

Proof. With $Q_{N,P} = P^N\left(L_{(\theta_1,\dots,\theta_N)} \in \bullet \middle| L_{(\theta_1,\dots,\theta_N)} \in \Gamma\right)$ and under the stated assumptions about $\mathfrak{G}$, Theorem 1 indicates that $(\Gamma, Q_{N,P}) \mapsto v_\Gamma(Q_{N,P})$ both is practically equivalent to $P_\Gamma^\star$ and is a coherent updating method.

For the other results not requiring those assumptions about $\mathfrak{G}$, consider any $P \in \mathcal{P}_\Theta$ and $\Gamma \in \mathfrak{G}$ and any coherent updating method $(\Gamma, Q_{N,P}) \mapsto v_\Gamma(Q_{N,P})$ and its coherent update $v_\Gamma(P) = P_\Gamma$ (see Definition 3). In terms of Definition 2, $P \mapsto \bar{\pi}_{N,P} \coloneqq \int \pi(\bullet)\, dQ_{N,P}(\pi)$, and $P \mapsto \bar{\pi}_{N,P,\Gamma} \coloneqq \int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma)$, Conditions 1 and 2 of Definition 3 state the following for any $\ell \in \mathcal{C}_{\Theta,\mathcal{D}}$ and $\Delta \in \mathcal{D}$:

$$\lim_{N\to\infty} \int \ell(\theta, \Delta)\, d\bar{\pi}_{N,P}(\theta) = \int \ell(\theta, \Delta)\, dP(\theta) \tag{16}$$

$$\lim_{N\to\infty} \int \ell(\theta, \Delta)\, d\bar{\pi}_{N,P,\Gamma}(\theta) = \int \ell(\theta, \Delta)\, dP_\Gamma(\theta) \tag{17}$$

If $P \in \Gamma$, then the definition of $\bar{\pi}_{N,P,\Gamma}$ as a mixture of π with respect to a conditional distribution given $\pi \in \Gamma$ implies that $\int \ell(\theta, \Delta)\, d\bar{\pi}_{N,P,\Gamma}(\theta) \to \int \ell(\theta, \Delta)\, dP(\theta)$ at least as fast as the $\int \ell(\theta, \Delta)\, d\bar{\pi}_{N,P}(\theta|X = x) \to \int \ell(\theta, \Delta)\, dP(\theta)$ indicated by equation (16). A comparison with equation (17) then indicates that $P_\Gamma = P$. Thus, $(\Gamma, Q_{N,P}) \mapsto v_\Gamma(Q_{N,P})$ is prior-persistent.

Equation (17) also establishes insight-receptiveness: $\bar{\pi}_{N,P,\Gamma} \in \Gamma$ for all $N = 1, 2, \ldots$ with the practical equivalence of $\bar{\pi}_{N,P,\Gamma}$ and P_Γ implies that $P_\Gamma \in \operatorname{Limpts}\Gamma$.

By equations (4) and (12),

$$u_x(v_\Gamma(Q_{N,P})) = u_x(Q_{N,P}(\bullet|\Gamma)) = Q_{N,P}(\bullet|\Gamma, X = x) = (u_x(Q_{N,P}))(\bullet|\Gamma) = v_\Gamma(u_x(Q_{N,P})),$$

from which it follows that every coherent updating method $(\Gamma, Q_{N,P}) \mapsto v_\Gamma(Q_{N,P})$ commutes with conditionalization. □

4 Inference given an uncertain insight about a prior

4.1 Extension of coherent updating to uncertain insights

The concept of coherent updating given an adequate set of priors generalizes to coherent updating given a distribution of such sets of distributions in $\mathcal{P}_\Theta$. This generalization applies when the adequate set of priors is not known with certainty.

Definition 4. Let $\mathfrak{F}_\mathfrak{G}$ denote a σ-field of subsets of $\mathfrak{G}$ and G a probability distribution on the measurable

space $(\mathfrak{G}, \mathfrak{F}_{\mathfrak{G}})$. The distribution $\bar{P}_G \in \mathcal{P}_\Theta$ is a *coherent update* of a prior distribution $P \in \mathcal{P}_\Theta$ under G if there is a function $P' \mapsto \{Q_{N,P'} : N = 1, 2, \dots\}$ that meets Condition 1 of Definition 3 (*pre-updating equivalence)* and this generalization of Condition 2:

- *Updating equivalence.* The sequence $\left\{\int\int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma)\, dG(\Gamma) : N = 1, 2, \dots\right\}$ of mixtures of conditional distributions is practically equivalent to $\bar{P}_G$.

Conditioning on a random set also has frequentist applications (e.g., Couso et al., 2014, p. 30).

Proposition 1. *If P_Γ is a coherent update of a prior distribution $P \in \mathcal{P}_\Theta$ under the adequate set Γ for all $\Gamma \in \mathfrak{F}_{\mathcal{P}_\Theta}$, then the mixture distribution $\bar{P}_G = \int P_\Gamma(\bullet)\, dG(\Gamma)$ is a coherent update of P under any distribution G on $(\mathfrak{G}, \mathfrak{F}_{\mathfrak{G}})$.*

Proof. Let $P_N = \int \pi(\bullet)\, dQ_{N,P}(\pi)$ and $\bar{P}_N = \int\int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma)\, dG(\Gamma)$. By Lemma 2 and Definition 3,

$$\begin{aligned} \int \pi(\bullet)\, dQ_{N,P}(\pi) &\xrightarrow{\text{weak}} P \\ \int\int \pi(\bullet)\, dQ_{N,P}(\pi)\, dG(\Gamma) = \left(\int dG(\Gamma)\right)\left(\int \pi(\bullet)\, dQ_{N,P}(\pi)\right) &\xrightarrow{\text{weak}} \left(\int dG(\Gamma)\right) P = P \end{aligned}$$

satisfying the pre-updating equivalence condition of Definition 4 according to Lemma 2. Analogous steps establish satisfaction of the updating equivalence condition:

$$\begin{aligned} \int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma) &\xrightarrow{\text{weak}} P_\Gamma \\ \int\int \pi(\bullet)\, dQ_{N,P}(\pi|\Gamma)\, dG(\Gamma) &\xrightarrow{\text{weak}} \bar{P}_G. \end{aligned}$$

□

The special case of averaging a maximum entropy distribution over the distribution of a quantity in a linear constraint was rejected by Jaynes (1989) but recommended by Cheeseman and Stutz (2005).

4.2 Set estimation for coherent updating

4.2.1 Uncertain insights based on adequate sets

A random adequate set may be simplified by making it a function of a random level of adequacy and of the observed data. For some real number α in an interval $\mathcal{A}$ and for an observation y that may differ from x, a set $\Gamma(\alpha; y)$ of prior distributions is a *prior α-adequate constraint set* if it consists of the closed convex hull of a *prior α-adequate set*, the prior distributions with some data-dependent measure of adequacy of at least α:

$$\Gamma(\alpha; y) = \operatorname{clco}\{P' \in \mathcal{P}_\Theta : a(P'; y) \geq \alpha\},$$

where $P' \mapsto a(P'; y)$ is a real-valued function on $\mathcal{P}_\Theta$. Data-dependent measures of adequacy include Bayesian p values, proper scoring rules, Bayes factors, and likelihood ratios (Bickel, 2015b). A random scalar A then defines the distribution G by $\Gamma(A; y) \sim G$.

Proposition 1 says the coherent update of P is

$$\begin{aligned} P_G &= \int P_{\Gamma(\alpha;y)}(\bullet)\, dA(\alpha) \\ &= \int_{\inf \mathcal{A}}^{a(P;y)} P_{\Gamma(\alpha;y)}(\bullet)\, dA(\alpha) + \int_{a(P;y)}^{\sup \mathcal{A}} P_{\Gamma(\alpha;y)}(\bullet)\, dA(\alpha) \\ &= A(\{\alpha \leq a(P; y) : \alpha \in \mathcal{A}\})\, P(\bullet) + \int_{a(P;y)}^{\sup \mathcal{A}} P_{\Gamma(\alpha;y)}(\bullet)\, dA(\alpha), \end{aligned} \tag{18}$$

equation (18) following from

$$\alpha \leq a(P; y) \implies P(\bullet) \in \Gamma(\alpha; y) \implies P_{\Gamma(\alpha;y)}(\bullet) = P(\bullet)$$

since coherent updating methods are prior-persistent (Theorem 2). $P_{\Gamma(\alpha;y)}(\bullet)$ by definition minimizes the divergence over $\Gamma(\alpha; y)$ with respect to $P(\bullet)$.

While using different data for checking the prior and for moving from the prior to the posterior ($x \neq y$) would seem ideal, it is often not feasible in practice (Bickel, 2015b). Thus, $y \coloneqq x$ in the remainder of the paper without loss of generality.

4.2.2 Sets with specified coverage probabilities

Choosing $\mathrm{U}(0,1)$ as the distribution of the threshold A is a natural default when $1-\alpha$ is a probability that $P' \in \Gamma(\alpha; X)$ for a fixed prior P' or a probability that $\pi' \in \Gamma(\alpha; x)$ for a random prior π' according to a hierarchical model. That coverage probability $1-\alpha$ may be the confidence level, fiducial probability, or posterior probability with $\Gamma(\alpha; X)$ as a confidence set, with $\Gamma(\alpha; x)$ as a fiducial set, or with $\Gamma(\alpha; x)$ as a Bayesian credible set, respectively. In any case, the uncertain adequate set corresponding to the observation x is $\Gamma(A; x)$, where either $A \sim \mathrm{U}(0,1)$, by default, or the distribution of A models the betting policy of a real or hypothetical agent. The default is a special case of the randomized closed sets of Nguyen (2006, §5.2), provided that $\Gamma(\alpha; x)$ is upper semicontinuous (see Bickel, 2015b, §2.1). It simplifies equation (18) to

$$P_G(\bullet) = a(P; x)\, P(\bullet) + (1 - a(P; x)) \left(\frac{1}{1 - a(P; x)} \int_{a(P;y)}^{1} P_{\Gamma(\alpha;x)}(\bullet)\, d\alpha \right). \tag{19}$$

The confidence-set case arises from measuring adequacy by a prior predictive p value:

$$a(P'; x) = p^x(P') \coloneqq \mathrm{Prob}_{\theta \sim P', X \sim f_\theta} \left(t_{P'}(X) \geq t_{P'}(x) \right), \tag{20}$$

where $t_{P'}(\bullet) : \mathcal{X} \to \mathbb{R}$ transforms the data to the statistic used to check prior distribution P'. Thus, $p^X(P') \sim \mathrm{U}(0,1)$ when sampling from the prior (i.e., $\theta \sim P'$ and $X \sim f_\theta$). It follows that

$$\mathcal{P}(\alpha; X) \coloneqq \left\{ P' \in \mathcal{P}_\Theta : p^X(P') \geq \alpha \right\} \tag{21}$$

is a $(1-\alpha)\,100\%$ confidence set for P' in the sense that

$$1 - \alpha = \mathrm{Prob}_{\theta \sim P', X \sim f_\theta} \left(P' \in \Gamma(\alpha; X) \right) \tag{22}$$

for any $P' \in \mathcal{P}_\Theta$ and $\alpha \in [0,1]$ (Bickel, 2015a). The same applies to the calibrated posterior predictive p values of Hjort et al. (2006) since they follow equation (20) for some function $t_{P'}(\bullet)$. The α-adequate constraint set in this case is the closed convex set $\Gamma(\alpha; x) = \mathrm{clco}\, \mathcal{P}(\alpha; x)$.

Example 5. Applying $A \sim \mathrm{U}(0,1)$ to Examples 2 and 3, substituting $P_{\Gamma(\alpha;x)}(\bullet)$ from equation (11) into

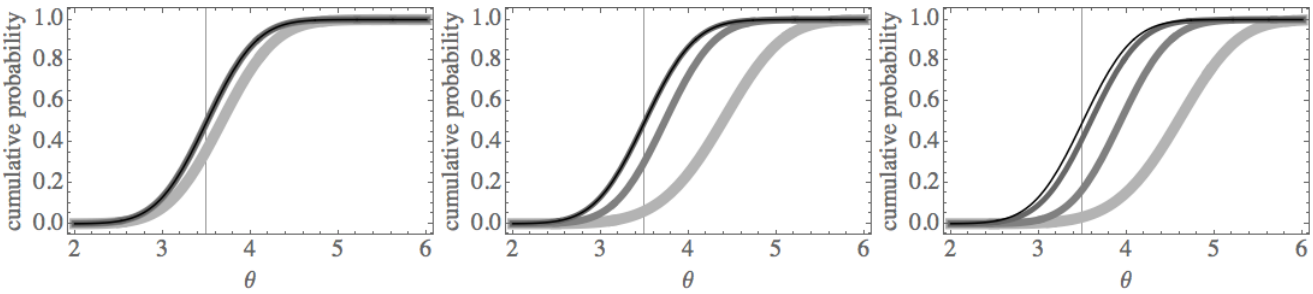

Figure 1: Normal-normal prior distribution functions of the normal mean θ, in order of increasing line thickness and lightness: initial prior distribution (P); prior conditional on a constraint fixed at the $\alpha = 1\%$ significance level $\left(P^{\star}_{\Gamma(0.01;x)}\right)$; prior conditional on a constraint fixed at the $\alpha = 5\%$ significance level $\left(P^{\star}_{\Gamma(0.05;x)}\right)$; prior conditional on a random constraint ($P^{\star}_{G}$). The sample mean increases from left to right: $\bar{x} = 4$ in the left panel, where $P = P^{\star}_{\Gamma(0.01;x)} = \left(P^{\star}_{\Gamma(0.05;x)}\right)$; $\bar{x} = 4.8$ in the center panel, where $P = P^{\star}_{\Gamma(0.01;x)}$; $\bar{x} = 5$ in the right panel. The vertical line is at $\mu_0 = 3.5$, the mean of the initial prior, P.

equation (19), with $p^x(3.5) = p^x\left(\mathrm{N}\left(3.5, \sigma_0^2\right)\right)$, yields

$$P^{\star}_{G}(\bullet) = p^x(3.5)\,\mathrm{N}\left(3.5, \sigma_0^2\right)(\bullet) + \left(1 - p^x(3.5)\right)\left(\frac{1}{1 - p^x(3.5)}\int_{p^x(3.5)}^{1} P_{\Gamma(\alpha;x)}(\bullet)\,d\alpha\right)$$

$$= \begin{cases} p^x(3.5)\,\mathrm{N}\left(3.5, \sigma_0^2\right)(\bullet) + \left(1 - p^x(3.5)\right)\pi^x_{<}(\bullet) & \text{if } \bar{x} < 3.5 \\ p^x(3.5)\,\mathrm{N}\left(3.5, \sigma_0^2\right)(\bullet) + \left(1 - p^x(3.5)\right)\pi^x_{>}(\bullet) & \text{if } \bar{x} \geq 3.5 \end{cases}$$

$$\pi^x_{>} := \left(\frac{\int_{p^x(3.5)}^{1} \mathrm{N}\left(\mathrm{CDF}^{-1}_{\bar{x}}\left(\alpha/2\right), \sigma_0^2\right)d\alpha}{1 - p^x(3.5)}\right).$$

and similarly for $\pi^x_{<}$.

Figure 1 displays the distribution functions of P, and $P^{\star}_{\Gamma(0.01;x)}$, $P^{\star}_{\Gamma(0.05;x)}$, and $P^{\star}_{G}$ using $\bar{x} = 4.8, 5$ and these settings from Bickel (2015b,a), following Walter and Augustin (2009): $\bar{x} = 4$, $n = 10$, $\sigma^2 = 1$, and $\sigma_0^2 = 1/5$. In the left panel, $P = P^{\star}_{\Gamma(0.01;x)} = \left(P^{\star}_{\Gamma(0.05;x)}\right)$ because $p^x(3.5) = 0.36 > 0.05$ when $\bar{x} = 4$. However, $0.01 < p^x(3.5) = 0.02 < 0.05$ when $\bar{x} = 4.8$ (center panel) and $p^x(3.5) = 6 \times 10^{-3} < 0.01$ when $\bar{x} = 5$ (right panel). ▲

Remark 1. Since $a(P; x) = p^x(P)$, equation (19) treats the p value of P as the prior probability that $\theta \sim P$.

By contrast, a fiducial argument leads to treating a p value as a posterior probability that a null hypothesis is true rather than the prior probability. That argument leads to making inferences and decisions on the basis of

$$\Pi^x \coloneqq p_0^x \Pi_0^x + (1 - p^x(P))\,\Pi_1^x,$$

where Π_0^x and Π_1^x are coherent fiducial distributions, probability measures derived from confidence distributions, and p_0^x is a p value testing a hypothesis consistent with the sole use of Π_0^x (Bickel and Padilla, 2014). That fiducial approach conflicts with the present default approach if $p_0^x = p^x(P)$ since p_0^x is a fiducial posterior probability that $\theta \sim \Pi_0^x$, whereas $p^x(P)$ is a prior probability in equation (19).

4.2.3 Integrated likelihood sets

For $\alpha \leq 0$, an *α-adequate integrated likelihood set*,

$$\widetilde{\mathcal{P}}(\alpha; x) = \left\{ P' \in \mathcal{P}_\Theta : \int f_\theta(x)\, dP'(\theta) \geq 2^\alpha \arg\sup_{P' \in \mathcal{P}_\Theta} \int f_\theta(x)\, dP'(\theta) \right\}, \tag{23}$$

has been proposed for determining which priors conflict with x (Bickel, 2015b). A case of this $\widetilde{\mathcal{P}}(\alpha; x)$ for non-Bayesian inference was suggested by Fisher (1973, pp. 75-76) with $2^\alpha = 1/2, 1/5, 1/15$ and with the equivalent of each P' a Dirac measure at a different value of θ. The corresponding α-adequate constraint set is the convex closure $\widetilde{\Gamma}(\alpha; x) = \operatorname{clco} \widetilde{\mathcal{P}}(\alpha; x)$.

The choice of a distribution for A for conditioning on the insight that $P' \in \widetilde{\Gamma}(A; y)$ may be guided by the following considerations. An α-adequate integrated likelihood set is considered unreliable if $\widetilde{\mathcal{P}}(\alpha; X)$ has a high probability of failing to include a fixed prior distribution P' under $\theta \sim P'$ and $X \sim f_\theta$. That probability is the *probability of a misleading insight*,

$$\varpi_{P'}(\alpha) = \operatorname{Prob}_{\theta \sim P', X \sim f_\theta} \left(P' \notin \widetilde{\mathcal{P}}(\alpha; X) \right), \tag{24}$$

which is a model-checking generalization of the probability of observing misleading evidence defined by Bickel (2012). That probability is itself a composite-hypothesis extension of the probability of observing misleading likelihood-ratio evidence defined by Royall (2000) and developed by Blume (2008) and others. Baskurt (2014), following Royall and Tsou (2003), considers satisfaction of an upper bound on the latter

probability as a necessary property that a pseudolikelihood ratio must possess to be a valid measure of statistical evidence (cf. Bickel, 2013, §3.2).

In the simple case corresponding to current practice, $A \sim \delta_\alpha$ for some value α, perhaps -3 , -5, or -7 (Bickel, 2015b), with the result that the insight is that $P' \in \widetilde{\Gamma}(\alpha; y)$ almost surely. While agreeing with the likelihood principle, that approach is not reliable for all models and sample sizes. It would be considered reliable if $\varpi_{P'}(\alpha)$ is less than 0.01, 0.05, another conventional value, or the upper bound mentioned above. Otherwise, the analysis could be postponed until additional data arrive, but if that is not practical, α may have to be lowered (Hodge et al., 2011). This suggests a two-stage approach:

1. Check whether the specified likelihood-ratio threshold 2^α is low enough that the misleading adequacy probability $\varpi(\alpha, P')$ is below some threshold such as 0.05.

2. If so, the insight is that $\widetilde{\Gamma}(\alpha; y)$ is the adequate set of prior distributions for application of Theorem 1. Otherwise, return to Stage 1 with a lower value of α.

4.2.4 Likelihood-confidence sets

The two-stage approach of Section 4.2.3 may be replaced by a single-stage approach involving, for each $\alpha' \in [0, 1]$, the $\varpi_{P'}^{-1}(\alpha')$-adequate integrated likelihood set $\widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(\alpha'); x\right)$. Combining equations (23) and (24),

$$\begin{aligned} 1 - \alpha' = 1 - \varpi_{P'}\left(\varpi_{P'}^{-1}(\alpha')\right) &= 1 - \text{Prob}_{\theta \sim P', X \sim f_\theta}\left(P' \notin \widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(\alpha'); X\right)\right) \\ &= \text{Prob}_{\theta \sim P', X \sim f_\theta}\left(P' \in \widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(\alpha'); X\right)\right). \end{aligned}$$

A comparison with equation (22) demonstrates that the likelihood set $\widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(\alpha'); X\right)$ is a $(1 - \alpha')\,100\%$ confidence set for P'. In fact, $\widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(\alpha'); X\right) = \mathcal{P}(\alpha'; X)$, where $\mathcal{P}(\alpha'; X)$ is the α'-adequate confidence set of equation (21) with the likelihood-ratio test statistic

$$x \mapsto t_{P'}(x) = \int f_\theta(x)\, dP'(\theta) \,/\, \arg\sup_{P' \in \mathcal{P}_\Theta} \int f_\theta(x)\, dP'(\theta)$$

in equation (20) to define the prior predictive p value for testing the null hypothesis that the prior is P'. Section 4.2.2 indicates the natural default adequate set is $\widetilde{\mathcal{P}}\left(\varpi_{P'}^{-1}(A); x\right) = \mathcal{P}(A; X)$ with $A \sim \text{U}(0, 1)$.

The special case in which $\mathcal{P}_\Theta$ is a set of Dirac measures on Θ makes further connections to earlier work. First, reliance on the significance level α' rather than on the likelihood-ratio threshold α provides the calibration that some models require for good frequentist performance (see Kalbfleisch (2000) and Severini (2000, p. 102)). Second, $\Gamma(\alpha'; X)$, as both a likelihood set and a confidence set, is called a *likelihood-confidence set* after the likelihood-confidence interval studied by Sprott (2000, §5.3). Third, a large-sample limit of the $t_{P'}$-based prior predictive p value is a p value from a likelihood-ratio test under broad regularity conditions, which is closely related to the *s value* of Patriota (2013).

Acknowledgments

A discussion with Lisa Strug on measuring evidence is gratefully acknowledged. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, by Agriculture and Agri-Food Canada, and by the Faculty of Medicine of the University of Ottawa.

References

Augustin, T., Coolen, F., de Cooman, G., Troffaes, M. (Eds.), 2014. Introduction to Imprecise Probabilities. Wiley Series in Probability and Statistics. Wiley.

Baez, J., Fritz, T., 2014. A Bayesian characterization of relative entropy. Theory and Applications of Categories 29, 422–456.

Baskurt, Z., 2014. Two avenues of inference: The evidential paradigm and relative belief theory. Ph.D. thesis, University of Toronto.

Bickel, D. R., 2012. The strength of statistical evidence for composite hypotheses: Inference to the best explanation. Statistica Sinica 22, 1147–1198.

Bickel, D. R., 2013. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81, 188–206.

Bickel, D. R., 2015a. Fiducial model averaging for Bayesian and frequentist inference. Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/32313.

Bickel, D. R., 2015b. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. International Journal of Approximate Reasoning 66, 53–72.

Bickel, D. R., Padilla, M., 2014. A prior-free framework of coherent inference and its derivation of simple shrinkage estimators. Journal of Statistical Planning and Inference 145, 204–221.

Billingsley, P., 1999. Convergence of Probability Measures. Wiley Series in Probability and Statistics. Wiley, New York.

Blume, J. D., 2008. How often likelihood ratios are misleading in sequential trials. Communications in Statistics - Theory and Methods 37, 1193–1206.

Caticha, A., Giffin, A., 2006. Updating probabilities. arXiv preprint physics/0608185 (DOI: 10.1063/1.2423258).

Cheeseman, P., Stutz, J., 2005. Generalized maximum entropy. AIP Conference Proceedings 803 (1), 374–381.

Cooke, R. M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.

Couso, I., Dubois, D., Sánchez, L., 2014. Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables: An Introduction for Ph.D. Students and Practitioners. SpringerBriefs in Applied Sciences and Technology. Springer, New York.

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

Crauel, H., 2002. Random Probability Measures on Polish Spaces. Stochastics Monographs. CRC Press, London.

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.

Dembo, A., Zeitouni, O., 2009. Large Deviations Techniques and Applications. Stochastic Modelling and Applied Probability. Springer, Berlin.

Diaconis, P., Zabell, S. L., 1982. Updating subjective probability. Journal of the American Statistical Association 77, 822–830.

Dudley, R., 2002. Real Analysis and Probability. Cambridge University Press, Cambridge.

Elga, A., 2010. Subjective Probabilities should be Sharp. Philosophers Imprint 10 (5), 1–11.

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

Fisher, R. A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Genest, C., Mcconway, K., Schervish, M., 1986. Characterization of externally Bayesian pooling operators. Annals of Statistics 14 (2), 487–501.

Genest, C., Zidek, J. V., 1986. Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science 1, 114–135.

Grendar Jr., M., Grendar, M., 2004. Maximum probability and maximum entropy methods: Bayesian interpretation. AIP Conference Proceedings 707, 490–494.

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

Halpern, J., 2003. Reasoning about Uncertainty. MIT Press, Cambridge.

Harremoës, P., 2007. Information topologies with applications. In: Csiszár, I., Katona, G. O. H., Tardos, G., Wiener, G. (Eds.), Entropy, Search, Complexity. Vol. 16 of Bolyai Society Mathematical Studies. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 113–150.

Hjort, N., Dahl, F., Steinbakk, G., 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association 101 (475), 1157–1174.

Hodge, S. E., Baskurt, Z., Strug, L. J., 2011. Using parametric multipoint lods and mods for linkage analysis requires a shift in statistical thinking. Human Heredity 72 (4), 264–275.

Jaynes, E. T., 1989. Where do we stand on maximum entropy? (1978). In: Rosenkrantz, R. (Ed.), E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics. Vol. 158 of Synthese Library. Springer Netherlands, pp. 210–314.

Kalbfleisch, J. D., 2000. Comment on R. Royall, "On the probability of observing misleading statistical evidence". Journal of the American Statistical Association 95, 770–771.

McConway, K. J., 1981. Marginalization and linear opinion pools. Journal of the American Statistical Association 76, 410–414.

Nguyen, H., 2006. An Introduction to Random Sets. CRC Press, New York.

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

Paris, J. B., 2014. What you see is what you get. Entropy 16 (11), 6186–6194.

Patriota, A. G., 2013. A classical measure of evidence for general null hypotheses. Fuzzy Sets and Systems 233, 74 – 88.

Rassoul-Agha, F., Seppäläinen, T., 2015. A Course on Large Deviations with an Introduction to Gibbs Measures. Graduate Studies in Mathematics. American Mathematical Society, Providence.

Royall, R., 2000. On the probability of observing misleading statistical evidence. Journal of the American Statistical Association 95, 760–768.

Royall, R., Tsou, T.-S., 2003. Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. Journal of the Royal Statistical Society. Series B 65, 391–404.

Severini, T., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.

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.

Skyrms, B., 1985. Maximum entropy inference as a special case of conditionalization. Synthese 63, 55–74.

Sprott, D. A., 2000. Statistical Inference in Science. Springer, New York.

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

Williams, P. M., 1980. Bayesian conditionalisation and the principle of minimum information. The British Journal for the Philosophy of Science 31, 131–144.