Blending Bayesian and frequentist methods according to the precision of prior information with applications to hypothesis testing

David R. Bickel

Abstract The proposed minimax procedure blends strict Bayesian methods with \( p \) values and confidence intervals or with default-prior methods. Two applications to hypothesis testing bring some implications to light. First, the blended probability that a point null hypothesis is true is equal to the \( p \) value or a lower bound of an unknown posterior probability, whichever is greater. As a result, the \( p \) value is reported instead of any posterior probability in the case of complete prior ignorance but is ignored in the case of a fully known prior. In the case of partial knowledge about the prior, the possible posterior probability that is closest to the \( p \) value is used for inference. The second application provides guidance on the choice of methods used for small numbers of tests as opposed to those appropriate for large numbers. Whereas statisticians tend to prefer a multiple comparison procedure that adjusts each \( p \) value for small numbers of tests, large numbers instead lead many to estimate the local false discovery rate (LFDR), a posterior probability of hypothesis truth. Each blended probability reduces to the LFDR estimate if it can be estimated with sufficient accuracy or to the adjusted \( p \) value otherwise.

Keywords Confidence distribution · Fiducial inference · Imprecise probability · Maximum entropy · Multiple hypothesis testing · Multiple comparison procedure · Robust Bayesian analysis
1 Introduction

1.1 Motivation

In Bayesian inference, a posterior distribution encodes all the relevant information about the value of a parameter. Given a posterior distribution, the posterior mean and posterior median are point estimates, highest-density credible regions are interval estimates, and every hypothesis about the parameter has a posterior probability.

It is now generally acknowledged that Bayesian methods have both strengths and weaknesses. On one hand, the specification of prior distributions enables the statistically rigorous incorporation of information beyond that of the data set at hand. On the other hand, since specified priors can vary substantially from one researcher to another, disagreement on whether and how to do so is a major obstacle to more widespread adoption of Bayesian methods.

The problem has been formalized in robust Bayesian statistics by considering a set of reasonable prior distributions (e.g., Jozani et al. 2012). Multiple prior distributions then arise because different statisticians use different default prior distributions (see Kass and Wasserman 1996) or because different domain experts have different prior probability assessments (Genest and Zidek 1986; Kracík 2011). Alternatively, the set of priors may represent the partially ordered preferences of a single agent or, more objectively, different states of the physical system under study (Sect. 7.1.1).

While robust Bayesian statistics has enjoyed considerable success, its limitations with respect to applied data analysis have also become clear. If the set of plausible priors is small enough, the sensitivity of a posterior to the prior in the set (Berger 1990; Lavine 1991; Liu and Aitkin 2008) can attest to the reliability of the results when that sensitivity is negligible but provides no recourse for interpreting the data when it is higher (Berger 1984). Worse, the set of plausible priors may be so large that the Bayesian approach is abandoned in favor of a frequentist method without the guidance from theory that could clarify the choice between approaches and even merge them as warranted. The main goals of this paper are to provide a theoretical framework that can play that role and to illustrate its application to common problems in hypothesis testing.

1.2 The strategy of this paper

A generally recognized strength of Bayesian statistics is its ability to incorporate previous information into data analysis by specifying a prior distribution. On the other hand, when the prior distribution is only known very roughly, inferences can be more accurate if they instead rely on a frequentist method that does not depend on a prior distribution. In the intermediate case in which previous knowledge places moderate constraints on the prior, the safest strategy from a minimax perspective is to rely on the frequentist method as much as possible without violating the constraints.
This paper proposes a general information-theoretic framework in order to blend Bayesian and frequentist inference according to these criteria:

1. **Complete knowledge of the prior** If the prior is known, the corresponding posterior is used for inference. Among statisticians, this principle is almost universally acknowledged. However, it is rarely the case that the prior is known for all practical purposes.

2. **Negligible knowledge of the prior** If there is no reliable knowledge of a prior, inference is based on methods that do not require such knowledge. This principle motivates not only the development of confidence intervals (CIs) and $p$ values but also Bayesian posteriors derived from improper and data-dependent priors. Accordingly, blended inference must allow the use of such methods when applicable.

3. **Continuum between extremes** Inference relies on the prior to the extent that it is known while relying on the other methods to the extent that it is not known. Thus, there is a gradation of methodology between the above two extremes. The premise of this paper is that this intermediate scenario calls for a careful balance between pure Bayesian methods on one hand and impure Bayesian or non-Bayesian methods on the other hand.

### 1.3 Additional background

Various compromises between Bayesian and frequentist approaches to statistical inference represent first attempts to combining attractive aspects of each approach (Good 1983). The problem may be considered in terms of a set of plausible priors, as in interval probability (Weichselberger 2000; Augustin 2002, 2004) and robust Bayesian (Berger 1984) approaches. Both objective and subjective interpretations of the set of plausible priors are possible (Sects. 1.1, 7.1.1).

Employing a set of plausible priors, the knowledge of the prior has been framed in terms of empirical Bayes CIs of the parameter of interest (Carlin and Louis 2009), in terms of hierarchical fiducial distributions of hyperparameter values corresponding to the plausible prior distributions (Fisher 1973; Dawid and Stone 1982; Padilla and Bickel 2012; Bickel 2013), and in terms of optimal point estimation (Yuan 2009). The “hybrid inference” approach (Yuan 2009) succeeded in leveraging Bayesian point estimators with maximum likelihood estimators, reducing to the former or the latter in the presence or absence of a reliably estimated prior on all parameters. Thus, it may in an important sense satisfy the criteria of Sect. 1.2 in the case of point estimation given a suitable loss function.

The most well known decision-theoretic approaches for the inference problem involving a set of plausible priors are the minimax Bayes risk (“$\Gamma$-minimax”) practice of minimizing the maximum Bayes risk (Robbins 1951; Good 1952; Berger 1985; Vidakovic 2000) and the maxmin expected utility (“conditional $\Gamma$-minimax”) practice of maximizing the minimum posterior expected payoff or, equivalently, minimizing the maximum posterior expected loss (Gärdenfors and Sahlin 1982; Gilboa and Schmeidler 1989; DasGupta and Studden 1989; Vidakovic 2000; Augustin 2002, 2004). Augustin (2004) reviews both methods in terms of interval probabilities. With typical
loss functions, the former method meets the above criteria for classical minimax alternatives to Bayesian methods but does not apply to other attractive alternatives. For example, several CIs, $p$ values, and objective-Bayes posteriors routinely used in biostatistics are not minimax optimal. (Fraser and Reid 1990 and Fraser 2004 argued that requiring the optimality of frequentist procedures can lead to trade-offs between hypothetical samples that potentially mislead scientists or yield pathological procedures). Optimality in the classical sense is not required of the frequentist methods or other alternative procedures under the framework outlined below. It instead represents such procedures in terms of a benchmark distribution to be used in place of the Bayesian posteriors to the degree warranted by the unavailability of prior information.

1.4 Heuristic overview

First, the building blocks of the framework are laid in Sect. 2.

To define a general theory of blended inference that meets a formal statement of the three criteria, Sect. 3 then introduces a variation of a zero-sum game of Topsøe (1979, 2007) and Harremoës and Topsøe (2001). (The discrete version of the game also appeared in Pfaffelhuber 1977). The “nature” opponent selects a prior consistent with the available knowledge as the “statistician” player selects a posterior distribution with the aim of maximizing the minimum information gained relative to one or more alternative methods. Such benchmark methods may be CI procedures, frequentist hypothesis tests, or other techniques that are not necessarily Bayesian. Under a convexity assumption, the optimal move of the statistician maximizes the entropy, subject to the constraints imposed by the prior knowledge, with respect to a parameter distribution representing the benchmark methods. Reliance on those methods distinguishes this approach from maximizing entropy with respect to a uniform distribution or other default prior (Jaynes 2003; Kass and Wasserman 1996; Williamson 2010).

From that theory, Sect. 4 derives a widely applicable framework for testing hypotheses. For concreteness, the motivating results are heuristically summarized here. Consider the problem of testing $H_0: \theta^* = 0$, the hypothesis that a real-valued parameter $\theta^*$ of interest is equal to the point 0 on the real line $\mathbb{R}$. Let $p(x)$ denote the $p$ value resulting from a statistical test.

It has long been recognized that the $p$ value for a simple (point) null hypothesis is often smaller than Bayesian posterior probabilities of the hypothesis (Lindley 1957; Berger and Sellke 1987). Suppose $\theta^*$ has an unknown prior distribution according to which the prior probability of $H_0$ is $\pi_0$. While $\pi_0$ is unknown, it is assumed to be no less than some known lower bound denoted by $\pi_0$.

Following the methodology of Berger et al. (1994), Sellke et al. (2001) found a generally applicable lower bound on the Bayes factor. As Sect. 4.1 will explain, that bound immediately leads to

$$\Pr(H_0|p(X) = p(x)) = \left(1 - \left(\frac{1 - \pi_0}{\pi_0 ep(x) \log p(x)}\right)\right)^{-1}$$ (1)
as a lower bound on the unknown posterior probability of the null hypothesis for $p(x) < 1/e$ and to $\pi_0$ as a lower bound on the probability if $p(x) \geq 1/e$.

In addition to $\Pr(H_0|p(X) = p(x))$, the unknown Bayesian posterior probability of $H_0$, there is a frequentist posterior probability of $H_0$ that will guide selection of a posterior probability for inference based on $\pi_0 \geq \pi_0$ and other constraints summarized by $\Pr(H_0|p(X) = p(x)) \geq \Pr(H_0|p(X) = p(x))$. While it is incorrect to interpret the $p$ value $p(x)$ as a Bayesian posterior probability, it will be seen in Sect. 4.2 that $p(x)$ is a confidence posterior probability that $H_0$ is true.

With the confidence posterior as the benchmark, the optimal move of the statistician player in the zero-sum game described above gives the blended posterior probability that the null hypothesis is true. It is simply the maximum of the $p$ value and the lower bound on the Bayesian posterior probability:

$$\Pr(H_0; p(x)) = p(x) \lor \Pr(H_0|p(X) = p(x)),$$

where $\lor$ denotes maximization. By plotting $\Pr(H_0; p(x))$ as a function of $p(x)$ and $\pi_0$, Figs. 1 and 2 illustrate each of the above criteria for blended inference:

1. **Complete knowledge of the prior** In this example, the prior is only known when $\pi_0$, the lower bound of $\pi_0$, is equal to the upper bound of $\pi_0$. Since the upper bound is 1, complete prior knowledge corresponds to $\pi_0 = 1$, in which case

$$\Pr(H_0; p(x)) = \Pr(H_0|p(X) = p(x)) = 1$$

for any $p(x)$. Thus, the $p$ value is ignored in the presence of a known prior.

---

**Fig. 1** Blended posterior probability that the null hypothesis is true versus the $p$ value. The curves correspond to lower bounds of prior probabilities ranging in 5% increments from 0% on the bottom to 100% on the top.
Fig. 2  Blended posterior probability that the null hypothesis is true versus the $p$ value and the lower bound of the prior probability that the null hypothesis is true. The top plot displays the full domain, half of which is shown in the bottom plot.

2. Negligible knowledge of the prior There is no knowledge of the prior when $\pi_0 = 0$ and negligible knowledge when $\pi_0$ is so low that $\Pr(H_0|p(X) = p(x)) \leq p(x)$. In such cases, $\Pr(H_0; p(x)) = p(x)$, and the Bayesian posteriors are ignored.

3. Continuum between extremes When $\pi_0$ is of intermediate value in the sense that $\Pr(H_0|p(X) = p(x))$ is exclusively between $p(x)$ and 1,

$$
\Pr(H_0; p(x)) = \Pr(H_0|p(X) = p(x)) < 1.
$$

Consequently, $\Pr(H_0; p(x))$ increases gradually from $p(x)$ to 1 as $\pi_0$ increases (Figs. 1, 2). In this case, the blended posterior lies in the set of allowed Bayesian
posteriors but is on the boundary of that set that is the closest to the $p$ value. Thus, both the $p$ value and the Bayesian posteriors influence the blended posterior and thus the inferences made on its basis.

The plotted parameter distribution will be presented in Sect. 4.3 as a widely applicable blended posterior.

While the assumptions leading to the above lower bound are often reasonable for two-sided testing, they are less reasonable for one-sided testing. They are relaxed in Sect. 5, which derives a new class of multiple comparison procedures from the framework of blended inference. The resulting blended posterior probabilities of the null hypotheses tend to be equal to estimates of local false discovery rates to the extent that there are enough hypotheses to make such estimates reliable. On the other hand, to the degree that the estimates are unreliable, the blended posterior probabilities are equal to $p$ values adjusted by non-Bayesian multiple comparison procedures, which is consistent with the commonly held position (e.g., Westfall 2010; Efron 2010b) that such procedures are suitable when the number of $p$ values is insufficient for accurate estimation of the local false discovery rate (contrast Bickel 2011b). In the most extreme case in that direction, there is only a single null hypothesis, and its blended posterior probability is equal to the unadjusted $p$ value.

Next, Sect. 6 contributes additional details and generalizations in a series of remarks. Section 7 discusses aspects of probability interpretation with an emphasis on decision theory and concludes the paper with a brief summary.

2 Preliminary concepts

2.1 Basic notation

Denote the observed data set, typically a vector or matrix of observations, by $x$, a member of a set $X$ that is endowed with a $\sigma$-algebra $\mathcal{X}$. The value of $x$ determines two sets of posterior distributions that can be blended for inference about the value of a target parameter. Much of the following notation is needed to specify those sets and to transform general Bayesian posteriors and confidence posteriors or other benchmark posteriors such that they are defined on the same measurable space, that of the target parameter. (A confidence posterior, to be defined in Sect. 4.2.1, is a parameter distribution from which CIs and $p$ values may be extracted. As such, it facilitates blending typical frequentist procedures with Bayesian procedures).

Due to the simultaneous use of probability distributions on the parameter space both from robust Bayesian statistics and from a development of fiducial inference, additional notation distinguishes between Bayesian and non-Bayesian posterior distributions and between the random variables and probability spaces of those distributions. This is simplified by using the same letters and fonts for the same types of objects and decorations on those letters to distinguish Bayesian versions from non-Bayesian versions according to the following conventions. An advantage is that the decorations may be ignored on a first reading of Sects. 2.2 and 2.3 and largely omitted in classroom expositions.

While regular $P$s and $Q$s denote probability distributions, calligraphic $\mathcal{P}$s denote sets of probability distributions. This keeps with the literature on information theory.
2.2 Bayesian posteriors

With some measurable space \((\hat{\Theta}_*, \hat{A}_*)\) for parameter values in \(\hat{\Theta}_*\), let \(\mathcal{P}_*^{\text{plaus}}\) denote a set of probability distributions on \((\mathcal{X} \times \hat{\Theta}_*, \mathcal{X} \otimes \hat{A}_*)\). Any distribution in \(\mathcal{P}_*^{\text{plaus}}\) is called a plausible distribution. Each plausible distribution is understood as a joint distribution of the data and the parameter that is plausible in the sense that it complies with any constraints and other information available about the parameter before observing \(x\).

Each plausible distribution of the parameter, marginalized over the data, is called a plausible prior (distribution). By the definition of a plausible distribution, this prior distribution is proper in the sense that it, unlike many default priors, is a probability measure.

On the other hand, any distribution of a parameter is called a posterior (distribution) if it depends on \(x\), the observed value. For some \(\hat{P}_*^{\text{plaus}} \in \mathcal{P}_*^{\text{plaus}}\), an example of a posterior distribution on \((\hat{\Theta}_*, \hat{A}_*)\) is \(\hat{P}_* = \hat{P}_*^{\text{plaus}}(\cdot|X = x)\), where \(X\) is a random variable of a distribution on \((\mathcal{X}, \mathcal{X})\) that is determined by \(\hat{P}_*^{\text{plaus}}\). \(\hat{P}_*\) is called a Bayesian posterior (distribution) since it is equal to a conditional distribution of the parameter given \(X = x\). Adapting an apt term from Topsøe (2007), the set \(\hat{\mathcal{P}}_* = \left\{ \hat{P}_*^{\text{plaus}}(\cdot|X = x) : \hat{P}_*^{\text{plaus}} \in \mathcal{P}_*^{\text{plaus}} \right\}\) of Bayesian posteriors on \((\hat{\Theta}_*, \hat{A}_*)\) may be considered the “knowledge base.” For a set \(\Theta\), if \(\hat{t} : \hat{\Theta}_* \to \Theta\) is an \(\hat{A}_*\)-measurable map and if \(\hat{\theta}_*\) has distribution \(\hat{P}_* \in \hat{\mathcal{P}}_*\), then \(\hat{\theta} = \hat{t}(\hat{\theta}_*)\), referred to as an inferential target of \(\hat{P}_*\), has induced probability space \((\Theta, \mathcal{A}, \hat{P})\). The parameter \(\hat{\theta}\), rather than \(\hat{\theta}_*\), is called the “inferential target” because \(\hat{\theta}\) is estimated or because any hypotheses to be tested may be more simply stated in terms of \(\hat{\theta}\). The set

\[
\hat{\mathcal{P}} = \left\{ \hat{P} : \hat{t}(\hat{\theta}_*) \sim \hat{P}, \hat{\theta}_* \sim \hat{P}_* \in \hat{\mathcal{P}}_* \right\}
\]

of all distributions thereby induced and the set \(\mathcal{P}\) of all probability distributions on \((\Theta, \mathcal{A})\) are related by \(\hat{\mathcal{P}} \subseteq \mathcal{P}\).

Example 1 In the hypothesis test of Sect. 1.4, \(\hat{\theta} = 0\) if the null hypothesis that \(\hat{\theta}_* = 0\) is true and \(\hat{\theta} = 1\) if the alternative hypothesis that \(\hat{\theta}_* \neq 0\) is true, where \(\hat{\theta}_*\) and \(\hat{\theta}\) are random variables with distributions respectively defined on the Borel space \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\) and the discrete space \(\{0, 1\}\), where \(2^{\{0, 1\}}\) is the power set of \(\{0, 1\}\). Thus, in this case, \(\hat{t}\) is the indicator function \(1_{(-\infty, 0) \cup (0, \infty)} : \mathbb{R} \to \{0, 1\}\), yielding \(\hat{\theta}_* = 1_{(-\infty, 0) \cup (0, \infty)}(\hat{\theta}_*)\). Section 4 considers this example in more detail.

A function that transforms a set of parameter distributions to a single parameter distribution on the same measurable space has been called an inference process (Paris 1994; Paris and Vencovská 1997) and a “representation” (Weichselberger 2001, p. 258; Augustin 2002). The resulting distribution is known as a “reduction” (Bickel 2012b).
of the set. Perhaps the best known inference process for a discrete parameter set $\Theta$ is that of the \textit{maximum entropy principle}, which would select a member of $\hat{\mathcal{P}}$ such that it has higher entropy than any other member of the set (see Remark 1).

This paper presents a wide class of inference processes such that each transforms $\hat{\mathcal{P}}$ to a member of $\mathcal{P}$ on the basis the following concept of a benchmark distribution on $(\Theta, \mathcal{A})$. Unlike the related approach introduced in Bickel (2012b), the approach studied herein does not require specification of an operational posterior distribution, a posterior that depends on a single prior. See Bickel (2012b) for other comparisons between the approaches.

2.3 Benchmark posteriors

For the convenience of the reader, the same Latin and Greek letters will be used for the set of posteriors that will represent a gold standard or benchmark method of inference as for the Bayesian posteriors of Sect. 2.2, with the double-dot $\ddot{\cdot}$ replacing the single-dot $\dot{\cdot}$. Let $\ddot{\mathcal{P}}_*$ represent a set of one or more posterior distributions on some measurable space $(\ddot{\Theta}_*, \ddot{\mathcal{A}}_*)$, and let $\hat{\mathcal{P}}_*$ represent a set of such sets. For instance, considering any $\hat{\mathcal{P}}_*$ in $\ddot{\mathcal{P}}_*$, $\hat{\mathcal{P}}_*$ may be a confidence posterior (a fiducial-like distribution to be defined precisely in Sect. 4.2), a confidence structure (Balch 2012), a fiducial posterior (Bickel and Padilla 2014), a generalized fiducial posterior (Hannig 2009), or even a Bayesian posterior based on a default prior such as an improper prior, a measure that is not a probability distribution. For example, many commonly used reference priors (Bernardo 1979; Berger et al. 2009) are improper. Each posterior distribution represents a single method of data analysis. [In the first case, nested CIs with inexact coverage rates generate a set $\ddot{\mathcal{P}}_*$ of multiple confidence posteriors rather than the single confidence posterior that is generated by exact CIs (Bickel 2012a)].

Suppose there exists a subparameter function $\ddot{\tau} : \ddot{\mathcal{P}}_* \rightarrow \Theta$ such that $\hat{\mathcal{P}}_*$, the probability distribution of the subparameter $\ddot{\tau}(\ddot{\mathcal{P}}_*)$, is defined on $(\Theta, \mathcal{A})$. $\hat{\mathcal{P}}_*$ is called the benchmark posterior (distribution), and $\ddot{\theta} = \ddot{\tau}(\ddot{\mathcal{P}}_*)$ is the inferential target of $\ddot{\mathcal{P}}_*$. It follows that $\hat{\mathcal{P}}_*$ is in $\mathcal{P}$ but not necessarily in $\hat{\mathcal{P}}$. In the simplest case, $(\Theta, \mathcal{A}) = (\ddot{\Theta}_*, \ddot{\mathcal{A}}_*)$, and $\hat{\mathcal{P}}_* = \{ \hat{\mathcal{P}} \}$ is the set of a single posterior $\hat{\mathcal{P}}$, the benchmark posterior.

\textbf{Example 2} Consider a model in which the full parameter $\hat{\Theta}_* \in \hat{\Theta}_*$ consists of an interest parameter $\hat{\theta}$ and a nuisance parameter $\hat{\lambda}$. The measurable space of $\hat{\Theta}_* = \{ \hat{\theta}, \hat{\lambda} \}$ is denoted by $(\hat{\Theta}_*, \hat{\mathcal{A}}_*)$, and that of $\hat{\theta}$ by $(\Theta, \mathcal{A})$. Suppose that a set of Bayesian posteriors is available for $\hat{\Theta}_*$ but that nested CIs are only available for an unknown parameter $\theta \in \Theta$. Thus, a confidence posterior $\hat{\mathcal{P}}$ is available on $(\Theta, \mathcal{A})$ but not on $(\hat{\Theta}_*, \hat{\mathcal{A}}_*)$. Then the framework of this section can be applied by using the function $\ddot{\tau}$ such that $\theta = \ddot{\tau}(\hat{\theta}_*)$ in order to project the Bayesian posteriors onto $(\Theta, \mathcal{A})$, the measurable space on which $\hat{\mathcal{P}}$ is defined. In this case, since there is only one possible benchmark posterior, the function $\ddot{\tau}$ need not be explicitly constructed.

The function $\ddot{\tau}$ allows consideration of a set of possible benchmark posteriors by transforming it to a random variable distributed as a single benchmark posterior defined on $(\Theta, \mathcal{A})$, the same measurable space as the above Bayesian posteriors of $\hat{\theta}$. Since that function is unusual, two ways to compose it will now be explained.
Example 3 Consider the inference process $\tilde{\Pi} : \tilde{\mathcal{P}}_* \rightarrow \mathcal{P}_*$, where $\mathcal{P}_*$ is the set of all probability distributions on $\left( \Theta_*, \mathcal{A}_* \right)$. Define the random variable $\tilde{\theta}_*$ to have distribution $\tilde{\Pi}(\tilde{\mathcal{P}}_*)(\bullet) = \tilde{\Pi}(\tilde{\mathcal{P}}_*)$. If $\tilde{\tau} : \tilde{\Theta}_* \rightarrow \Theta$ is an $\mathcal{A}_*$-measurable function, then $\tilde{\theta} = \tilde{\tau}(\tilde{\theta}_*)$ is the inferential target of $\tilde{\mathcal{P}}_*$. Further, the distribution $\tilde{P}$ of $\tilde{\theta}$ is the benchmark posterior.

Example 4 Whereas Example 3 applied an inference process before a parameter transformation, this example reverses the order by first applying $\tilde{\tau}$. Let $\tilde{\mathcal{P}}$ denote the subset of $\mathcal{P}$ consisting of all distributions of the parameters transformed by $\tilde{\tau}$:

$$\tilde{\mathcal{P}} = \{ P : \tilde{\tau}(\tilde{\theta}_*) \sim P, \tilde{\theta}_* \sim \tilde{\mathcal{P}}_* \in \mathcal{P}_* \}.$$ 

Then an inference process transforms $\tilde{\mathcal{P}}$ to the benchmark posterior $\tilde{\mathcal{P}}$, which in turn is the distribution of $\tilde{\theta}$, the inferential target of $\mathcal{P}_*$.

2.4 Tools from information theory

The basic concepts of convexity and information divergence are fundamental to the proposed method of blending Bayesian and benchmark posterior distributions.

A set $Q$ of probability distributions on $(\Theta, \mathcal{A})$ is called convex if $\eta P + (1 - \eta) Q \in Q$ for all $P, Q \in Q$ and $\eta \in [0, 1]$ (Cover and Thomas 2006).

In terms of Radon–Nikodym differentiation, the information divergence of $P$ with respect to $Q$ on $(\Theta, \mathcal{A})$ is

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

if $Q$ dominates $P(P \ll Q)$ and $I(P||Q) = \infty$ otherwise, according to a measure-theoretic definition of relative entropy (Kakihara 1999, pp. 49–52). $I(P||Q)$ is also known as cross entropy, $I$-divergence, information for discrimination, and Kullback–Leibler divergence. Other measures of information may also be used (Remark 2).

For any posteriors $\hat{P} \in \hat{\mathcal{P}}$ and $Q \in \mathcal{P}$, the inferential gain $I(\hat{P}||\hat{P} \sim Q)$ of $Q$ relative to $\hat{P}$ given $\hat{P}$ is the amount of information gained by making inferences on the basis of $Q$ instead of the benchmark posterior $\hat{P}$:

$$I(\hat{P}||\hat{P} \sim Q) = I(\hat{P}||\hat{P}) - I(\hat{P}||Q).$$

(4)

Working from a coding perspective, Pfaffelhuber (1977) called this quantity “information gain,” and Topsøe (2007) used the symbol $\sim$.

Let $\hat{\mathcal{P}}(\hat{P})$ denote the largest subset of $\hat{\mathcal{P}}$ such that the information divergence of any of its members with respect to $\hat{P}$ is finite. That is,

$$\hat{\mathcal{P}}(\hat{P}) = \{ \hat{P} \in \hat{\mathcal{P}} : I(\hat{P}||\hat{P}) < \infty \},$$

which is nonempty by assumption. (The assumption is not necessary under the generalization described in Remark 3).
The information-theoretic origin of the other mathematical tools used is recorded by citation in Sect. 3. While knowledge of the literature cited is not necessary for understanding the proposed framework, it has value in proving the foundational results from first principles and in drawing connections to other research in information theory.

Much of that background is tailored to the statistics community in terms of maximum entropy in Grünwald and Dawid (2004). The appendix of Bickel (2011b) introduces concepts from coding theory in terms familiar to statisticians.

3 General theory

3.1 Blended inference

Recall the concept of inferential gain defined in Eq. (4). The blended posterior (distribution) $\tilde{P}$ is the probability distribution on $(\Theta, A)$ that maximizes the inferential gain relative to the benchmark posterior given the worst-case posterior restricted by the constraints that defined $\hat{P}$ and $\hat{P}(\hat{P})$:

$$\inf_{\hat{P} \in \hat{P}(\hat{P})} I(\hat{P} || \hat{P}) = \sup_{Q \in \mathcal{P}} \inf_{\hat{P} \in \hat{P}(\hat{P})} I(\hat{P} || \hat{P} \Rightarrow Q),$$

(6)

where the supremum and infimum over any set including an indeterminate number are $\infty$ and $-\infty$, respectively (Topsøe 2007). Inferences based on $\tilde{P}$ are blended in the sense that they depend on both $\hat{P}$ and $\hat{P}$ in the ways to be specified in Sect. 3.2.

The main result of Theorem 2 of Topsøe (2007) gives a simply stated solution of the optimization problem of Eq. (6) under a broad condition involving convexity as defined in Sect. 2.4.

**Proposition 1** If $I(\hat{P} || \hat{P}) < \infty$ for some $\hat{P} \in \hat{P}$ and if $\hat{P}(\hat{P})$ is convex, then the blended posterior $\tilde{P}$ is the probability distribution in $\hat{P}$ that minimizes the information divergence with respect to the benchmark posterior:

$$I(\tilde{P} || \tilde{P}) = \inf_{\hat{P} \in \hat{P}(\hat{P})} I(\hat{P} || \hat{P}).$$

(7)

**Proof** Topsøe (2007) proved the result from inequalities of information theory given the additional stated condition of his Theorem 2 that $I(\hat{P} || \hat{P}) < \infty$ for all $\hat{P} \in \hat{P}(\hat{P})$. (See Remark 3). The condition that $I(\hat{P} || \hat{P}) < \infty$ for some $\hat{P} \in \hat{P}$ and the above definition of $\hat{P}(\hat{P})$ ensure that the condition is met.

Alternatively, the minimization of information divergence may define $\tilde{P}$ rather than result from its definition in terms of the game (Remark 4).

3.2 Properties of blended inference

The desiderata of Sect. 1 for blended inference can now be formalized. A posterior distribution $B(\bullet; \hat{P}, \hat{P})$ on $(\Theta, A)$ is said to blend the set $\hat{P}$ of Bayesian posteriors
with the benchmark posterior \( \hat{P} \) for inference about the parameter in \( \Theta \) provided that \( B(\bullet; \hat{P}, \check{P}) \) satisfies the following criteria under the conditions of Proposition 1:

1. **Complete knowledge of the prior** If \( \hat{P} \) has a single member \( \hat{P} \), then \( B(\bullet; \hat{P}, \check{P}) = \hat{P} \).
2. **Negligible knowledge of the prior** If \( \check{P} \in \hat{P} \) and if \( \hat{P} \) has at least two members, then \( B(\bullet; \hat{P}, \check{P}) = \check{P} \).
3. **Continuum between extremes** For any \( D \geq 0 \) and any \( \mathcal{P}^* \subseteq \mathcal{P} \) such that

\[
\sup_{\hat{P} \in \mathcal{P}^*} |I(P||\hat{P}) - I(\hat{P}||\hat{P})| \leq D \tag{8}
\]

and such that \( \hat{P}(\check{P}) \cup \mathcal{P}^* \) is convex,

\[
|I(B(\bullet; \hat{P} \cup \mathcal{P}^*, \check{P})||\check{P}) - I(B(\bullet; \hat{P}, \check{P})||\check{P})| \leq D. \tag{9}
\]

**Theorem 1** The blended posterior \( \tilde{P} \) blends the set \( \hat{P} \) of Bayesian posteriors with the benchmark posterior \( \check{P} \) for inference about the parameter in \( \Theta \).

**Proof** Since the criteria are only required under the conditions of Proposition 1, it will suffice to prove that the criteria follow from Eq. (7). If \( \hat{P} \) has a single member \( \hat{P} \), then Eq. (7) implies that \( \check{P} = \hat{P} \), thereby ensuring Criterion 1. Similarly, if \( \check{P} \in \hat{P} \), then Eq. (7) implies that \( \check{P} = \hat{P} \), thus proving that Criterion 2 is met. Assume, contrary to Criterion 3, that there exist a \( D \geq 0 \) and a \( \mathcal{P}^* \subseteq \mathcal{P} \) such that \( \hat{P}(\check{P}) \cup \mathcal{P}^* \) is convex, Eq. (8) is true, and Eq. (9) is false with \( B(\bullet; \hat{P} \cup \mathcal{P}^*, \check{P}) \) and \( B(\bullet; \hat{P}, \check{P}) \) equal to the blended posteriors respectively using \( \hat{P} \cup \mathcal{P}^* \) and \( \check{P} \) as the sets of Bayesian posteriors. Then Eq. (7) can be written as

\[
I(B(\bullet; \hat{P} \cup \mathcal{P}^*, \check{P})||\check{P}) = \inf_{\hat{P} \in \hat{P}(\check{P}) \cup \mathcal{P}^*} I(\hat{P}||\check{P});
\]

\[
I(B(\bullet; \hat{P}, \check{P})||\check{P}) = \inf_{\hat{P} \in \hat{P}(\check{P})} I(\hat{P}||\check{P}).
\]

Hence, with \( a \wedge b \) signifying the minimum of \( a \) and \( b \),

\[
|I(B(\bullet; \hat{P} \cup \mathcal{P}^*, \check{P})||\check{P}) - I(B(\bullet; \hat{P}, \check{P})||\check{P})| = \inf_{\hat{P} \in \hat{P}(\check{P})} I(\hat{P}||\check{P}) - \inf_{\hat{P} \in \hat{P}(\check{P})} I(\hat{P}||\check{P}) \wedge \inf_{\hat{P} \in \mathcal{P}^*} I(P||\check{P}),
\]

which cannot exceed \( \inf_{\hat{P} \in \hat{P}(\check{P})} I(\hat{P}||\check{P}) - \inf_{P \in \mathcal{P}^*} I(P||\check{P}) \) and thus, according to Eq. (8), cannot exceed \( D \). Therefore, the above assumption that Eq. (9) is false is contradicted, thereby establishing satisfaction of Criterion 3.

Criterion 3 is much stronger than the heuristic idea of continuity introduced in Sects. 1.2 and 1.4. Its use of information divergence can be generalized to other measures of divergence (Remark 2).
4 Testing a two-sided null hypothesis

A fertile field of application for the theory of Sect. 3 is that of testing hypotheses, as outlined in Sect. 1.4. Building on Example 1, this section provides methodology for a wide class of models used in hypothesis testing.

4.1 A bound on the Bayesian posterior

Defining that class in terms of the concepts of Sect. 2.2 requires additional notation. For a continuous sample space \( \mathcal{X} \) and a function \( p : \mathcal{X} \to [0, 1] \) such that \( p(X) \sim U(0, 1) \) under a null hypothesis, each \( p(x) \) for any \( x \in \mathcal{X} \) will be called a \( p \) value. Using some dominating measure, let \( f_0 \) and \( f_1 \) denote probability density functions of \( p(X) \) under the null hypothesis (\( \hat{\theta} = 0 \)) and under the alternative hypothesis (\( \hat{\theta} = 1 \)), respectively. For the observed \( x \), the likelihood ratio \( f_0(p(x))/f_1(p(x)) \) is called the Bayes factor since, for a plausible distribution \( P_\text{plaus} \), Bayes’s theorem gives

\[
\frac{\psi(p(x))}{1 - \psi(p(x))} = \frac{P_\text{plaus}(\hat{\theta} = 0)}{P_\text{plaus}(\hat{\theta} = 1)} \frac{f_0(p(x))}{f_1(p(x))},
\]

where \( \psi(p(x)) = P_\text{plaus}(\hat{\theta} = 0)p(X) = p(x) \). In a parametric setting, \( f_1(p(x)) \) would be the likelihood integrated over the prior distribution conditional on the alternative hypothesis.

Let \( \kappa : \mathcal{X} \to \mathbb{R} \) denote the function defined by the transformation \( \kappa(x) = -\log p(x) \) for all \( x \in \mathcal{X} \). Then a probability density of \( \kappa(x) \) under the null hypothesis is the standard exponential density \( g_0(\kappa(x)) = e^{-\kappa(x)} \). Assume that, under the alternative hypothesis (\( \hat{\theta} = 1 \)), \( \kappa(X) \) admits a density function \( g_1 \) with respect to the same dominating measure as \( g_0 \). Consequently, \( g_0(\kappa(x))/g_1(\kappa(x)) = f_0(p(x))/f_1(p(x)) \). The hazard rate \( h_1(\kappa(x)) \) under the alternative is defined by \( h_1(\kappa(x)) = g_1(\kappa(x))/\int_{\kappa(x)}^{\infty} g_1(k)dk \) for all \( x \in \mathcal{X} \), and \( h_1 : (0, \infty) \to [0, \infty) \) is called the hazard rate function.

Sellke et al. (2001) obtained the following lower bound \( b(p(x)) \) of the Bayes factor \( b(x) \).

**Lemma 1** If \( h_1 \) is nonincreasing, then, for all \( x \in \mathcal{X} \),

\[
b(p(x)) = \frac{f_0(p(x))}{f_1(p(x))} \geq b(p(x)) = \begin{cases} -ep(x) \log p(x) & \text{if } p(x) < 1/e; \\ 1 & \text{if } p(x) \geq 1/e. \end{cases}
\]  

The condition on the hazard rate defines a wide class of models that is useful for testing simple, two-sided null hypotheses. A broad subclass will now be defined by imposing constraints on \( \pi_0 = P_\text{plaus}(\hat{\theta} = 0) \), the prior probability that the null hypothesis is true, in addition to the hazard rate condition. Specifically, \( \pi_0 \) is known to have \( \pi_0 \in [0, 1] \) as a lower bound. Thus, rearranging Eq. (10) as
The following parametric framework facilitates the application of Sect. 2.3 to hypothesis testing. The observation \( x \) is an outcome of the random variable \( \tilde{X} \) of probability space \((\mathcal{X}, \mathcal{F}, P_{\theta_*, \lambda_*})\), where the interest parameter \( \theta_* \in \tilde{\Theta}_* \) and a nuisance parameter \( \lambda_* \) (in some set \( \tilde{\Lambda}_* \)) are unknown. Let \( S : \tilde{\Theta}_* \times \mathcal{X} \to [0, 1] \) and \( t : \mathcal{X} \times \tilde{\Theta}_* \to \mathbb{R} \) denote functions such that \( S(\bullet; x) \) is a distribution function, \( S(\theta_*; X) \sim U(0, 1) \), and

\[
S(\theta_*; x) = P_{\theta_*, \lambda_*}(t(\tilde{X}; \theta_*) \geq t(x; \theta_*))
\]

for all \( x \in \mathcal{X}, \theta_* \in \tilde{\Theta}_*, \) and \( \lambda_* \in \tilde{\Lambda}_* \). \( S \) is known as a significance function, and \( t \) as a pivot or test statistic. It follows that \( p(x) = S(0; x) \) is a \( p \) value for testing the hypothesis that \( \theta_* = 0 \) and that \([S^{-1}(\alpha; X), S^{-1}(\beta; X)]\) is a (\( \beta - \alpha \))100% CI for \( \theta_* \) given any \( \alpha \in [0, 1] \) and \( \beta \in [\alpha, 1] \), where \( S^{-1}(\bullet; x) : [0, 1] \to \tilde{\Theta}_* \) is the inverse of \( S(\bullet; x) \) for each \( x \in \mathcal{X} \). Thus, whether a significance function is found from \( p \) values over a set of simple null hypotheses or instead from a set of nested CIs, it contains the information needed to derive either (Schweder and Hjort 2002; Singh et al. 2007; Bickel 2012a).

Let \( \hat{\theta}_* \) denote the random variable of the probability measure \( \hat{P}_* \) that has \( S(\bullet; x) \) as its distribution function. In other words, \( \hat{P}_*(\theta_* \leq \hat{\theta}_*) = S(\theta_*; x) \) for all \( \theta_* \in \tilde{\Theta}_* \). \( \hat{P}_* \) is called a confidence posterior (distribution) since it equates the frequentist coverage rate of a CI with the probability that the parameter lies in the fixed, observed CI:

\[
\beta - \alpha = P_{\theta_*, \lambda_*}(\theta_* \in [S^{-1}(\alpha; X), S^{-1}(\beta; X)])
= \hat{P}_*(\hat{\theta}_* \in [S^{-1}(\alpha; x), S^{-1}(\beta; x)])
\]

for all \( x \in \mathcal{X}, \theta_* \in \tilde{\Theta}_*, \) and \( \lambda_* \in \tilde{\Lambda}_* \). The term “confidence posterior” (Bickel 2011a, 2012a) is preferred here over the usual term “confidence distribution” (Schweder and Hjort 2002; Nadarajah et al. 2015) to emphasize its use not only as a way to express a procedure of CIs and \( p \) values but also as an alternative to Bayesian posterior distributions for inference and decision (Sect. 7.1.2). Polansky (2007), Singh et al.
(2007), Bickel (2012a) and Bickel and Padilla (2014) provide generalizations to vector parameters of interest. The confidence posterior $\hat{P}_\ast$ is essentially equivalent to the basic fiducial distribution (Bickel and Padilla 2014). The framework of this paper may be generalized by replacing it with other coherent fiducial distributions. Extensions based on multiple comparison procedures are provided in Sect. 5.

4.2.2 A confidence posterior for testing

For the application to two-sided testing of a simple null hypothesis, let $\theta_\ast = |\theta_{\ast\ast}|$, the absolute value of a real parameter $\theta_{\ast\ast}$ of interest, leading to $\Theta_1 = (0, \infty)$. Then $p(x) = S(0; x)$ is equivalent to a two-tailed $p$ value for testing the hypothesis that $\theta_{\ast\ast} = 0$. Since $\hat{P}_\ast(\theta_\ast \leq 0) = S(0; x)$ and since $\hat{P}_\ast(\theta_\ast \leq 0) = \hat{P}_\ast(\hat{\theta}_\ast = 0)$, it follows that $p(x) = \hat{P}_\ast(\hat{\theta}_\ast = 0)$, i.e., the $p$ value is equal to the probability that the null hypothesis is true. This little-known equality cannot be derived from Bayes’s theorem, and it generates distinctive point and interval estimators of $\theta_{\ast\ast}$ (Bickel and Padilla 2014).

If $\hat{P}_\ast$ is the only confidence posterior under consideration, then $\hat{P}_\ast = \{\hat{P}_\ast\}$, and there is no need for an inference process. Following the terminology of Example 3, $\check{\tau} : \Theta_1 \to \Theta$ is defined by $\check{\tau}(\hat{\theta}_\ast) = 1_{(0, \infty)}(\hat{\theta}_\ast)$. By implication, $\check{\theta} = 0$ if $\hat{\theta}_\ast = 0$ and $\check{\theta} = 1$ if $\hat{\theta}_\ast > 0$. Thus, $p(x) = \hat{P}_\ast(\hat{\theta}_\ast = 0)$ ensures that $\hat{P}(\check{\theta} = 0) = p(x)$, which in turn implies $\hat{P}(\check{\theta} = 1) = 1 - p(x)$.

Example 5 In the various $t$ tests, $\theta_\ast$ is the mean of $X$ or a difference in means, and the statistic $t(X; 0)$ is the absolute value of a statistic with a Student $t$ distribution of known degrees of freedom. The above formalism then gives the usual two-sided $p$ value from a $t$ test as $\hat{P}(\check{\theta} = 0)$ and $p(x)$. Special cases of this $\hat{P}$ have been presented as fiducial distributions (van Berkum et al. 1996; Bickel 2014c). Bickel and Padilla (2014) defends this type of interpretation of a two-sided $p$ value as a non-Bayesian posterior probability of the null hypothesis.

4.3 A blended posterior for testing

This subsection blends the above set $\hat{P}$ of Bayesian posteriors with the above confidence posterior $\check{P}$ as prescribed by Sect. 3.1. Gathering the results of Sects. 4.1 and 4.2,

$$\hat{P} = \left\{ \hat{P} \in \mathcal{P} : \hat{P}(\check{\theta} = 0) \geq \psi(p(x)) \right\} ;$$

$$\hat{P}(\check{\theta} = 0) = p(x) = 1 - \hat{P}(\check{\theta} = 1).$$

Equation (5) then implies that

$$\hat{P}(\check{P}) = \left\{ \hat{P} \in \mathcal{P} : \psi(p(x)) \leq \hat{P}(\check{\theta} = 0) < 1 \right\} .$$
in which the first inequality is strict if and only if $\psi(p(x)) = 0$ and the second inequality is strict unless $p(x) = 1$. Since $\dot{P}(\ddot{\theta})$ is convex, Proposition 1 yields

$$\tilde{P}(\theta = 0) = \begin{cases} \psi(p(x)) & \text{if } p(x) < \psi(p(x)) \\ p(x) & \text{if } p(x) \geq \psi(p(x)) \end{cases} ,$$

(12)

where $\theta$ is the random variable of distribution $\tilde{P}$. With the identities $\psi(p(x)) = \Pr(H_0|p(X) = p(x))$ and $\tilde{P}(\theta = 0) = \Pr(H_0; p(x))$ and with the establishment of Eq. (1) by Sect. 4.1, Eq. (12) verifies the claim of Eq. (2) made in Sect. 1.4.

5 Testing multiple hypotheses

5.1 Blending multiple comparison procedures

The single-test notation of Sect. 4 is extended to the problem of testing $N$ null hypotheses based on $N$ data vectors by adding the subscript $j \in \{1, \ldots, N\}$ to $x$, $\dot{\theta}$, and $\ddot{\theta}$. The $p$ values $p(x_1), \ldots, p(x_N)$ may be one-sided or two-sided and do not necessarily meet the condition of Lemma 1.

The hypothesis posterior probabilities that are equal to $p$ values (Sect. 4.2) would only be appropriate if it is suspected that most of the null hypotheses are false. On the other hand, information indicating that the vast majority of null hypotheses are true for all practical purposes is what can justify adjusting $p$ values to control family-wise error rates (Westfall et al. 1997; Cox 2006, p. 88). Thus, given such information, the benchmark posterior probabilities of the hypotheses may be equated with the adjusted $p$ values $\pi(x_1), \ldots, \pi(x_N)$ instead of the original, unadjusted $p$ values (Bickel 2012c):

$$\ddot{P}(\ddot{\theta}_j = 0) = \pi(x_j) \quad \text{for } j = 1, \ldots, N.$$ 

Since a joint distribution of $\ddot{\theta}_1, \ldots, \ddot{\theta}_N$ is not defined, each marginal distribution $\ddot{P}$ depends on $j$, which is suppressed to simplify the notation is much as possible.

A more modern class of multiple comparison procedures is that consisting of “estimators” (technically predictors) of the Bayesian posterior probability of the $j$th of $N$ null hypotheses conditional on the $p$ value:

$$\psi_j = \psi(p(x_j)) = \ddot{P}(\ddot{\theta}_j = 0),$$

where $\psi(\bullet)$ is the function that satisfies Eq. (10). In literature on multiple testing influenced by Benjamini and Hochberg (1995), $\psi_j$ is called the local false discovery rate (LFDR) since it is the probability that rejecting the null hypothesis results in a Type I error. This probability is physical rather than epistemological in the sense that it is a limiting relative frequency under a model of the physical system, not a level of belief under a model of an intelligent agent. (The letter $\psi$ stands for $\psi \varepsilon \nu \delta \hat{n} \varepsilon$ (pseudēs), the Greek word translated as “false”).

Let $\hat{\psi}_j$ denote an estimator of $\psi_j$ that is conservative enough that the interval $[0, \hat{\psi}_j]$ can be considered the set of plausible posterior probabilities of the $j$th null hypothesis. Conservatism is here understood as positive bias of some type, e.g., positive median
bias (Bickel 2013). The reasoning of Sect. 4.3, except with the lower bound replaced by the upper bound, results in

\[
\hat{P}_j = \left\{ \hat{P} \in \mathcal{P} : \hat{P} (\hat{\theta}_j = 0) \leq \hat{\psi}_j \right\};
\]

\[
\hat{P}_j (\tilde{P}) = \left\{ \hat{P} \in \mathcal{P} : 0 < \hat{P} (\hat{\theta}_j = 0) \leq \hat{\psi}_j \right\}.
\]

By Proposition 1, the blended posterior probability of the \( j \)th null hypothesis is the conservative LFDR estimate or the adjusted \( p \) value, whichever is lower: \( \hat{P}(\theta_j = 0) = \hat{\psi}_j \wedge \pi(x_j) \).

From that result, the next proposition is trivially proved but is worth stating to support the intuition of Efron (2010a) and others that the \( p \) value is appropriate when there is only a single test.

**Proposition 2** Suppose \( \pi(x_1) = p(x_1) \) and \( \hat{\psi}_1 = 1 \) whenever \( N = 1 \) and that Eqs. (13)–(14) hold. It follows that

\[
N = 1 \implies \hat{P}(\theta_1 = 0) = p(x_1).
\]

At the other extreme, LFDR estimates are deemed more appropriate when the number of tests is sufficiently high (Efron 2010a). That is reflected in the smoothly increasing tendency of the adjusted \( p \) values to exceed the LFDR estimates as the number of tests diverges, as is evident in the application of Sect. 5.2.2; cf. Bickel (2014a, b).

It is when the number of tests is between the smallest and largest scales that intuition cannot determine whether an adjusted \( p \) value or LFDR estimate is more suitable. In that situation, blended inference objectively determines whether a given data set calls for adjusted \( p \) values, LFDR estimates, or the combination of both that is seen in the following medium-scale applications.

### 5.2 Biostatistics applications

In both of this section’s applications of the strategy of blending multiple comparison procedures (Sect. 5.1), the following methods are used to generate each plausible set of Bayesian posterior probabilities and each confidence posterior used as the benchmark. As the benchmark probabilities, the confidence posterior probabilities are equated with the \( p \) values adjusted by the independent-statistic method of Sidak (1967).

The LFDR is conservatively estimated by the “MLE” method of Bickel (2013) with monotonicity enforced as described therein. That method of LFDR estimation performs well when the number of hypotheses is relatively small (Padilla and Bickel 2012), and its LFDR estimate is equal to 1 when there is only a single null hypothesis. The latter property is a condition of Proposition 2.
Fig. 3  Blending of empirical Bayes estimators with adjusted \( p \) values of two biostatistics applications. The dashed line represents the blended posterior probabilities of the null hypothesis, and the gray line represents \( p \) values adjusted by the independent-statistic method of Sidak (1967). The plot on the left is based on the 15 \( p \) values related to cardiac disease (Sect. 5.1), and that on the right is based on 20 randomly sampled microarray \( p \) values (Sect. 5.2.2). Each horizontal axis is an adjusted or unadjusted \( p \) value, depending on which is visually more informative.

5.2.1 Application to biomedical research

Neuhaus et al. (1992) tested \( N = 15 \) hypotheses about medical outcomes of a thrombolytic treatment, and Benjamini and Hochberg (1995) listed the corresponding 15 \( p \) values to illustrate their method of controlling the false discovery rate.

The method of blending multiple comparison procedures (Sect. 5.1) was applied to those \( p \) values. The left-hand side of Fig. 3 displays the adjusted \( p \) values, the conservative estimates of the LFDR, and the resulting blended posterior probabilities that each of the null hypotheses is true.

5.2.2 Application to gene expression

The multiple comparison procedure of Sect. 5.1 is also illustrated with a gene expression data set (Alba et al. 2005) in which \( x_j \) is a tuple of six logarithms of the measured ratios of mutant tomato expression to wild-type tomato expression of the \( j \)th gene.

The plot on the right-hand side of Fig. 3 displays the blended probabilities that the mutation did not affect the expression of \( N = 20 \) genes randomly sampled without replacement from the 6103 genes that have expression measurement for all six biological replicates at 3 days after the breaker stage of ripening. As the number of genes increases, the \( p \) value adjustment becomes increasingly severe, with the result that all of the blended probabilities are equal to the conservative LFDR estimates when the adjustment is applied to 6103 tests. The \( p \) values were obtained from the one-sample \( t \) test, as in Bickel (2012c).

6 Technical remarks

Remark 1  Technically, the principle of maximum entropy (Paris 1994; Paris and Venkovská 1997) mentioned in Sect. 2.2 could be used if \( \Theta \) is finite or countable infinite. However, unlike the proposed methodology, that practice is equivalent to making the
benchmark posterior $\tilde{P}$ depend on the function $\tilde{\tau}$ that maps a parameter space to $\Theta$ rather than on a method of data analysis that is coherent in the sense that its posterior depends on the data rather than on the hypothesis. If blending with such a method is not desired, one may average the Bayesian posteriors with respect to some measure that is not a function of $\Theta$. For example, averaging with respect to the Lebesgue measure, as Bickel (2012a) did with confidence posteriors, leads to $(1 + \psi(p(x)))/2$ as the posterior probability of the null hypothesis under the assumptions of Sect. 4.1. Remark 4 discusses a more tenable version of the maximum entropy principle for blended inference.

Remark 2 Using definitions of divergence that include information divergence (3) as a special case, Grünwald and Dawid (2004) and Topsøe (2004) generalized variations of Proposition 1. The theory of blended inference extends accordingly.

Remark 3 A generalization of Sect. 3 in a different direction from that of Remark 2 replaces each “$\inf_{\hat{\tau} \in \hat{\tau}(\tilde{\tau})}$” of Eq. (6) with “$\inf_{\hat{\tau} \in \hat{\tau}}$.” For that optimization problem, Theorem 2 of Topsøe (2007) has the condition that $\hat{\tau} \in \hat{\tau} \implies I(\hat{\tau}||\tilde{\tau}) < \infty$ in addition to the convexity of $\hat{\tau}$ that Proposition 1 of the present paper requires. Thus, in that formulation, the blended posterior $\tilde{P}$ need not satisfy Eq. (7) even if $\hat{\tau}$ is convex.

Remark 4 A posterior distribution $\tilde{P}$ that is defined by

$$I(\tilde{P}||\hat{P}) = \inf_{\hat{P} \in \hat{P}} I(\hat{P}||\tilde{P})$$

(15)

satisfies the desiderata of Sect. 3.2 whether or not the conditions of Proposition 1 hold. Certain axiomatic systems (e.g., Csiszár 1991) lead to this generalization of the principle of maximum entropy (Remark 1). For example, an agent that makes inferences on the basis of a confidence posterior $\hat{P}$ in the absence of parameter constraints would, upon learning such constraints in the form of $\hat{P}$, update that posterior to $\tilde{P}$ by maximizing entropy relative to that posterior, at least according to the reasonable axioms of Paris and Vencovská (1997). However, the optimization problem of Eq. (6) seems even more compelling in this context and defines $\tilde{P}$ even when no distribution satisfying Eq. (15) exists.

7 Discussion

7.1 Interpreting Bayes and non-Bayes posteriors

7.1.1 Objective and subjective sets of priors

As often happens, the same mathematical framework is subject to very different philosophical interpretations. Both objective and subjective applications of the sets of priors used in blended inference are possible, as seen briefly in Sect. 1.1.

First, the idea of an unknown prior within some set fits naturally in the use of Bayes’s theorem with proper priors that do not involve subjective interpretations of probability. The set of posteriors corresponding to the set of priors may be determined
by interval constraints on the corresponding priors without any requirement that they model levels of belief (Weichselberger 2000; Augustin 2002, 2004). In fact, unknown priors occur in many objective contexts involving purely frequentist interpretations of probability in terms of variability in the observable world rather than the uncertainty in the mind of an agent (Robinson 1991). For example, frequency-based priors are routinely estimated under random effects and empirical Bayes models; see, for example, Efron (2010a).

On the other hand, although the concept of an unknown prior cannot arise in Bayesian statistics under variations of the axioms of Savage (1954), it does arise in robust Bayesian statistics when the levels of belief of an intelligent agent have not been fully assessed (Berger 1984). In addition, while the proposed framework was motivated in Sect. 1.1 with an unknown prior in mind, the concept of imprecise or indeterminate probability (Walley 1991) could take the place of the set in which an unknown prior lies.

For instance, by relaxing the complete order of agent preferences required by Savage (1954) to allow a partial order of preferences, imprecise probability theories need not assume the existence of any true prior (Walley 1991; Coletti and Scozzafava 2002). When the partial order of preferences does not determine a unique decision, a rational agent, perhaps artificial, is free to base its action on additional desiderata; such multiple-stage decisions are called lexicographic [Levi (1986a, §§5.7, 6.9); Levi (1986b); see Remark 3 of Bickel (2012c)]. (In the case that each plausible prior corresponds to a statistician or other expert (Sect. 1.1), the agent may be a group of people acting corporately). A possible desideratum for the first stage is represented by the blended posterior defined in Sect. 3.1 to achieve the goals of Sects. 1.4 and 3.2. The blended posterior is then used to minimize expected loss as the second stage of the lexicographic procedure (Sect. 7.1.3).

### 7.1.2 Interpreting confidence posteriors

In the most intuitive interpretation of CIs, one has greater belief that the parameter value lies in an observed CI of a higher frequentist coverage rate than that it lies in one of a lower coverage rate. Under that interpretation and a subjective decision theory of maximizing expected utility (e.g., Savage 1954) or minimizing expected loss, the confidence posterior defined in Sect. 4.2.1 is the only posterior probability distribution that can determine the expected loss for generating estimates and other actions without violating coherence (Bickel and Padilla 2014). In that sense, intuitive frequentist inference is encapsulated in a confidence posterior in exactly the same way as Bayesian inference is encapsulated in a Bayesian posterior (Bickel 2012a). For other reasons to use confidence posteriors and their imprecise probability generalizations for inference rather than merely as means for encoding \( p \) values and CIs, see Bickel (2012a) and Balch (2012). The latter also argues in favor of confidence structures, which are essentially fiducial distributions that have conservative frequentist coverage.

The decision-theoretic interpretation of the confidence distribution is consistent with the following use of the blended posterior.
7.1.3 Decisions based on blended posteriors

Let $D$ denote an action space and $L$ a loss function on $\Theta \times D$. Regardless of whether the set of priors is objective or subjective (Sect. 7.1.1), the blended posterior $\tilde{P}$ yields the optimal action

$$\arg \inf_{\delta \in D} \int L(\theta_0, \delta) d\tilde{P}(\theta_0),$$

where $L(\theta_0, \delta)$ is the loss that would be incurred by taking decision $\delta$ if $\theta_0 \in \Theta_1$ were the true value of the parameter of interest. Since $\tilde{P} \in \hat{P}$, that action is determined by a Bayes rule, ensuring highly desirable properties (Seidenfeld 1988, 2004). It is the same decision taken by the benchmark posterior $\ddot{P}$ when $\ddot{P} \in \hat{P}$, reducing to the decision specified in Sect. 7.1.2 when $\ddot{P}$ is a confidence posterior. Simple examples of this minimization of expected loss with respect to a blended posterior are available in Bickel (2012b).

7.2 Concluding summary

The premise of this paper is that the relative reliabilities of a proper-prior Bayesian method, a benchmark (frequentist or default-prior Bayesian) method, and a blend of both depend on the extent to which prior information is available in the form of a probability distribution on parameter space. Strict Bayesian inference is used when the true prior distribution is known, benchmark inference is used when nothing is known about the prior, and both types of inference are combined according to the “game” outlined at the end of this section when the prior is known to be a member of some set that represents the available pre-data information.

The set of plausible priors generates a plausible set of Bayesian posterior distributions. If the benchmark posterior that corresponds to a method of frequentist inference or default-prior Bayesian inference lies within the plausible set, then the benchmark posterior is used for inference and decisions such as point estimates in the same way as a fully Bayesian posterior would be used were the true prior known. Otherwise, the posterior within the plausible set that is closest to the benchmark posterior is used for inference and decisions.

The game blends minimaxity and expected utility maximization into a single decision-theoretic method:

1. Analyze the data using a $p$ value method, a confidence interval method, a fiducial probability method, or a default-prior Bayesian method, and encode the results as a benchmark posterior distribution.
2. Analyze the data using Bayes’s theorem for each of the plausible prior distributions, and encode the results as the corresponding set of plausible posterior distributions.
3. The statistician player chooses a posterior distribution to maximize the amount of information gained by replacing the above benchmark posterior with the chosen posterior. At the same time, the nature opponent chooses the true posterior distrib-
ution from the plausible posteriors to minimize that amount of information. From the statistician player’s perspective, the best move is the worst-case posterior.

4. That worst-case posterior is used for inference and decisions, for example, to report the results of the data analyses as probabilities that null hypotheses are true and as the effect-size estimates that maximize expected payoff or, equivalently, minimize expected loss.

Special cases of the blending framework solve commonly arising problems of testing single and multiple hypotheses. The resulting algorithms are simple enough to implement in a line of code each:

- The blended posterior probability of a single null hypothesis is its two-sided \( p \) value or a lower bound of a Bayesian posterior probability, whichever is higher.
- The blended posterior probability of a null hypothesis in a family of other null hypotheses is its multiplicity-adjusted \( p \) value or a conservative estimate of its local false discovery rate, whichever is lower.

Acknowledgments The comments of the two anonymous reviewers and of the editor-in-chief are gratefully acknowledged for improving the clarity of presentation. In addition, I thank Xuemei Tang for providing the fruit-development microarray data. 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

Alba R, Payton P, Fei Z, McQuinn R, Debbie P, Martin GB, Tanksley SD, Giovannoni JJ (2005) Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. Plant Cell 17:2954–2965

Augustin T (2002) Expected utility within a generalized concept of probability—a comprehensive framework for decision making under ambiguity. Stat Pap 43(1):5–22

Augustin T (2004) Optimal decisions under complex uncertainty—basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. Zeitschrift für Angewandte Mathematik und Mechanik 84(10–11):678–687

Balch MS (2012) Mathematical foundations for a theory of confidence structures. Int J Approx Reason 53(7):1003–1019

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Berger JO (1984) Robustness of Bayesian analyses. Studies in Bayesian econometrics. North-Holland, Ch. The robust Bayesian viewpoint, pp 63–124

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, New York

Berger JO (1990) Robust Bayesian analysis: sensitivity to the prior. J Stat Plan Inference 25:303–328

Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of \( p \) values and evidence. J Am Stat Assoc 82:112–122

Berger JO, Brown L, Wolpert R (1994) A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis-testing. Ann Stat 22(4):1787–1807

Berger JO, Bernardo J, Sun D (2009) The formal definition of reference priors. Ann Stat 37(2):905–938

Bernardo JM (1979) Reference posterior distributions for Bayesian inference. J R Stat Soc B 41:113–147

Bickel DR (2011a) Estimating the null distribution to adjust observed confidence levels for genome-scale screening. Biometrics 67:363–370

Bickel DR (2011b) A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Can J Stat 39:610–631

Bickel DR (2012a) Coherent frequentism: a decision theory based on confidence sets. Commun Stat Theory Methods 41:1478–1496
Bickel DR (2012b) Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes. Electron J Statist 6:686–709
Bickel DR (2012c) Game-theoretic probability combination with applications to resolving conflicts between statistical methods. Int J Approx Reason 53:880–891
Bickel DR (2013) Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Stat Appl Genet Mol Biol 12:529–543
Bickel DR (2014a) A fiducial continuum from confidence sets to empirical Bayes set estimates as the number of comparisons increases. Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/31898
Bickel DR (2014b) Model fusion and multiple testing in the likelihood paradigm: shrinkage and evidence supporting a null hypothesis. Working paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/31897
Bickel DR (2014c) Small-scale inference: empirical Bayes and confidence methods for as few as a single comparison. Int Stat Rev 82:457–476
Bickel DR, Padilla M (2014) A prior-free framework of coherent inference and its derivation of simple shrinkage estimators. J Stat Plan Inference 145:204–221
Carlin BP, Louis TA (2009) Bayesian methods for data analysis, 3rd edn. Chapman and Hall/CRC, New York
Coletti C, Scozzafava R (2002) Probabilistic logic in a coherent setting. Kluwer, Amsterdam
Cover T, Thomas J (2006) Elements of information theory. Wiley, New York
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Csiszár I (1991) Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. Ann Stat 19:2032–2066
DasGupta A, Studden W (1989) Frequentist behavior of robust Bayes estimates of normal means. Stat Decis 7:333–361
Dawid AP, Stone M (1982) The functional-model basis of fiducial inference (with discussion). Ann Stat 10:1054–1074
Efron B (2010a) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge
Efron B (2010b) Rejoinder to comments on B. Efron, “Correlated z-values and the accuracy of large-scale statistical estimates”. J Am Stat Assoc 105:1067–1069
Fisher RA (1973) Statistical methods and scientific inference. Hafner Press, New York
Fraser DAS (2004) Ancillaries and conditional inference. Stat Sci 19:333–351
Fraser DAS, Reid N (1990) Discussion: an ancillarity paradox which appears in multiple linear regression. Ann Stat 18:503–507
Gärdenfors P, Sahlin N-E (1982) Unreliable probabilities, risk taking, and decision making. Synthese 53:361–386
Genest C, Zidek JV (1986) Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1:114–135
Gilboa I, Schmeidler D (1989) Maxmin expected utility with non-unique prior. J Math Econ 18(2):141–153
Good II (1952) Rational decisions. J R Stat Soc B 14:107–114
Good II (1983) Good thinking: the foundations of probability and its applications. G—reference, information and interdisciplinary subjects series. University of Minnesota Press, USA
Grunwald P, Dawid AP (2004) Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Ann Stat 32:1367–1433
Hannig J (2009) On generalized fiducial inference. Stat Sinica 19:491–544
Harremoës P, Topsøe F (2001) Maximum entropy fundamentals. Entropy 3(3):191–226
Jaynes E (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge
Jozani M, Marchand É, Parsian A (2012) Bayesian and robust bayesian analysis under a general class of balanced loss function. Stat Pap 53(1):51–60
Kakihara Y (1999) Abstract methods in information theory (series on multivariate analysis, volume 4). World Scientific, Singapore
Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91:1343–1370
Kračň J (2011) Combining marginal probability distributions via minimization of weighted sum of Kullback–Leibler divergences. Int J Approx Reason 52:659–671
Lavine M (1991) Sensitivity in Bayesian statistics: the prior and the likelihood. J Am Stat Assoc 86:396–399
