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

July 28, 2012

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa; 451 Smyth Road; Ottawa, Ontario, K1H 8M5

Keywords: blended inference; conditional Gamma-minimax; confidence distribution; confidence posterior; empirical Bayes; hybrid inference; maximum entropy; maxmin expected utility; minimum cross entropy; minimum divergence; minimum information for discrimination; minimum relative entropy; multiple hypothesis testing; multiple comparison procedure; observed confidence level; robust Bayesian analysis

Abstract

The following zero-sum game between nature and a statistician blends Bayesian methods with frequentist methods such as p-values and confidence intervals. Nature chooses a posterior distribution consistent with a set of possible priors. At the same time, the statistician selects a parameter distribution for inference with the goal of maximizing the minimum Kullback-Leibler information gained over a confidence distribution or other benchmark distribution. In cases of hypothesis testing, the statistician reports a posterior probability of the hypothesis that is informed by both Bayesian and frequentist methodology, each weighted according how well the prior is known.

As is generally acknowledged, the Bayesian approach is ideal given knowledge of a prior distribution that can be interpreted in terms of relative frequencies. On the other hand, frequentist methods such as confidence intervals and p-values have the advantage that they perform well without knowledge of such a distribution of the parameters. However, neither the Bayesian approach nor the frequentist approach is entirely satisfactory in situations involving partial knowledge of the prior distribution, the proposed procedure reduces to a Bayesian method given complete knowledge of the prior, to a frequentist method given complete ignorance about the prior, and to a blend between the two methods given partial knowledge of the prior. The blended approach resembles the Bayesian method rather than the frequentist method to the precise extent that the prior is known.

The proposed framework offers a simple solution to the enduring problem of testing a point null hypothesis. The blended probability that the null hypothesis is true is equal to the p-value or a lower bound of an unknown Bayesian posterior probability, whichever is greater. Thus, given total ignorance represented by a lower bound of 0, the p-value is used instead of any Bayesian posterior probability. At the opposite extreme of a known prior, the p-value is ignored. In the intermediate case, the possible Bayesian posterior probability that is closest to the p-value is used for inference. Thus, both the Bayesian method and the frequentist method influence the inferences made.

Similarly, blended inference may help resolve ongoing controversies in testing multiple hypotheses. Whereas the adjusted p-value is often considered the multiple comparison procedure (MCP) of choice for small numbers of hypotheses, large numbers of p-values enable accurate estimation of the local false discovery rate, a physical posterior probability of hypothesis truth. Each blended posterior probability reduces to either the adjusted p-value or the LFDR estimate by effectively determining on a hypothesis-by-hypothesis basis whether the LFDR can be estimated with sufficient accuracy. This blended MCP is applied to both a microarray data set and a more conventional biostatistics data set to illustrate its generality.

1 Introduction

1.1 Motivation

Various compromises between Bayesian and frequentist approaches to statistical inference represent first attempts to combining attractive aspects of each approach (Good, 1983). While the hybrid inference approach of Yuan (2009) succeeded in leveraging Bayesian point estimators with maximum likelihood estimates, reducing to the former or the latter in the presence or absence of a reliably estimated prior on all parameters, how to extend the theory beyond point estimation is not yet clear. Further, hybrid inference in its current form does not cover the case of a parameter of interest that has a partially known prior. Since such partial knowledge of a prior occurs in many scientific inference situations, it calls for a theoretical framework for method development that appropriately blends Bayesian and frequentist methods.

Ideally, blended inference would meet 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 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.

Instead of framing the knowledge of a prior in terms of confidence intervals, as in pure empirical Bayes approaches, it will be framed more generally herein in terms of a set of plausible priors, as in interval probability (Weichselberger, 2000; Augustin, 2002, 2004) and robust Bayesian (Berger, 1984) approaches. Whereas the concept of an unknown prior cannot arise in strict Bayesian statistics, it does arise in robust Bayesian statistics when the levels of belief of an intelligent agent have not been fully assessed (Berger, 1984). Unknown priors also occur in many more 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. For example, frequency-based priors are routinely estimated under random effects and empirical Bayes models; see, e.g., Efron (2010a). (Remark 1 comments further on interpretations of probability and relaxes the convenient assumption of a true prior.)

The most well known approaches for this problem are the *minimax Bayes risk* ("Γ-minimax") practice of minimizing the maximum Bayes risk (Robbins, 1951; Good, 1952; Berger, 1985; Vidakovic, 2000) and the *maxmin expected utility* ("conditional Γ-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 that need not be subjective. 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 confidence intervals, p-values, and objective-Bayes posteriors routinely used in bio-

statistics 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 alternative procedures under the framework outlined below, which can be understood in terms of a payoff function that incorporates the alternative procedures to be used as a benchmark for the Bayesian posteriors.

1.2 Heuristic overview

To define a general theory of blended inference that meets a formal statement of the three criteria, Section 2 introduces a variation of a zero-sum game of Topsøe (1979), Harremoës and Topsøe (2001), and Topsøe (2007). (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 confidence interval procedures, frequentist hypothesis tests, or other techniques that are not necessarily Bayesian.

From that theory, Section 3 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 θ_* of interest is equal to the point 0 on the real line $\mathbb{R}$. The observed data vector x is modeled as a realization of a random variable denoted by X. 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 θ_* has an unknown prior distribution according to

which the prior probability of H_0 is π_0. While π_0 is unknown, it is assumed to be no less than some known lower bound denoted by $\underline{\pi}_0$.

Following the methodology of Berger et al. (1994), Sellke et al. (2001) found a generally applicable lower bound on the Bayes factor. As Section 3.1 will explain, that bound immediately leads to

$$\underline{\Pr}\left(H_0|p\left(X\right)=p\left(x\right)\right)=\left(1-\left(\frac{1-\underline{\pi}_0}{\underline{\pi}_0 ep\left(x\right)\log p\left(x\right)}\right)\right)^{-1} \tag{1}$$

as a lower bound on the unknown posterior probability of the null hypothesis for $p\left(x\right)<1/e$ and to $\underline{\pi}_0$ as a lower bound on the probability if $p\left(x\right)\geq 1/e$.

In addition to $\Pr\left(H_0|p\left(X\right)=p\left(x\right)\right)$, 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 $\pi_0\geq\underline{\pi}_0$ and other constraints summarized by $\Pr\left(H_0|p\left(X\right)=p\left(x\right)\right)\geq\underline{\Pr}\left(H_0|p\left(X\right)=p\left(x\right)\right)$. While it is incorrect to interpret the p-value $p\left(x\right)$ as a *Bayesian* posterior probability, it will be seen in Section 3.2 that $p\left(x\right)$ is a *confidence* posterior probability that H_0 is true.

With the confidence posterior as the benchmark, the solution to the optimization problem 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:

$$\Pr\left(H_0;p\left(x\right)\right)=p\left(x\right)\vee\underline{\Pr}\left(H_0|p\left(X\right)=p\left(x\right)\right). \tag{2}$$

By plotting $\Pr\left(H_0;p\left(x\right)\right)$ as a function of $p\left(x\right)$ and $\underline{\pi}_0$, Figures 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 $\underline{\pi}_0 = 1$, in which case

$$\Pr\left(H_0; p\left(x\right)\right) = \underline{\Pr}\left(H_0 | p\left(X\right) = p\left(x\right)\right) = 1$$

for all $p\left(x\right)$. Thus, the p-value is ignored in the presence of a known prior.

2. **Negligible knowledge of the prior.** There is no knowledge of the prior when $\underline{\pi}_0 = 0$ and negligible knowledge when $\underline{\pi}_0$ is so low that $\underline{\Pr}\left(H_0 | p\left(X\right) = p\left(x\right)\right) \leq p\left(x\right)$. In such cases, $\Pr\left(H_0; p\left(x\right)\right) = p\left(x\right)$, and the Bayesian posteriors are ignored.

3. **Continuum between extremes.** When $\underline{\pi}_0$ is of intermediate value in the sense that $\underline{\Pr}\left(H_0 | p\left(X\right) = p\left(x\right)\right)$ is exclusively between $p\left(x\right)$ and 1,

$$\Pr\left(H_0; p\left(x\right)\right) = \underline{\Pr}\left(H_0 | p\left(X\right) = p\left(x\right)\right) < 1.$$

Consequently, $\Pr\left(H_0; p\left(x\right)\right)$ increases gradually from $p\left(x\right)$ to 1 as $\underline{\pi}_0$ increases (Figures 1 and 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 Section 3.3 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 Section 4, 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,

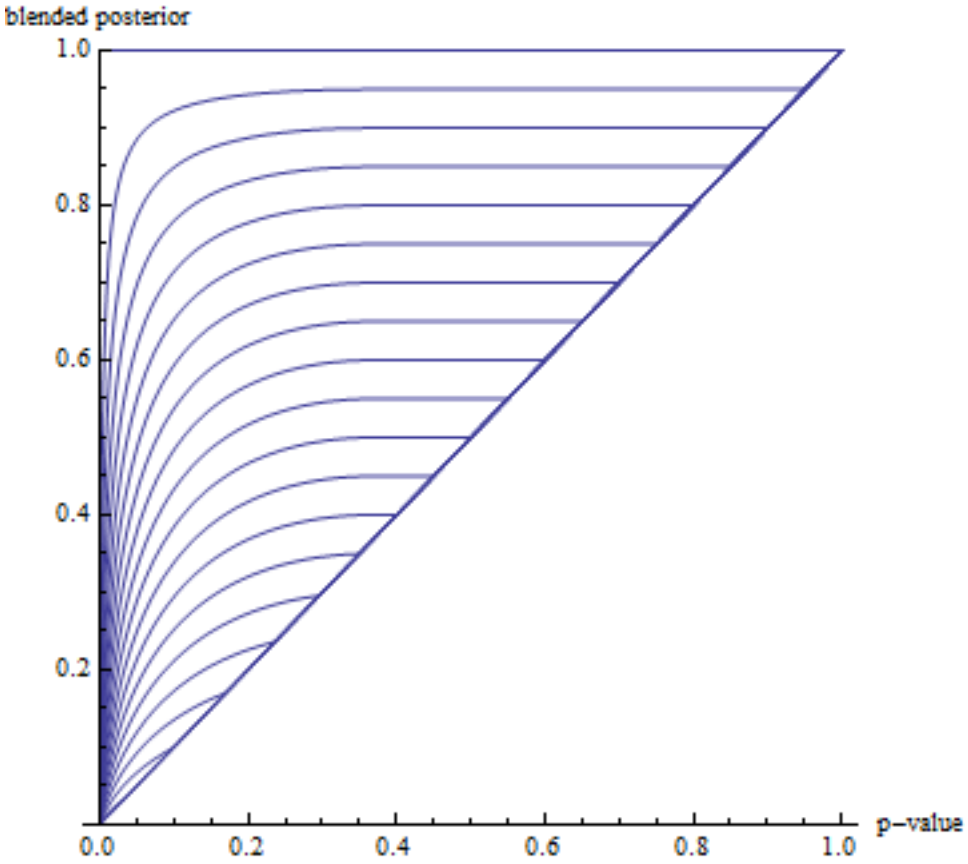

Figure 1: Blended posterior probability that the null hypothesis is true versus the p-value. The curves correspond to lower bounds of prior probabilities ranging in 5% increments from 0% on the bottom to 100% on the top.

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. 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.

Finally, Section 5 contributes additional details and generalizations in a series of remarks.

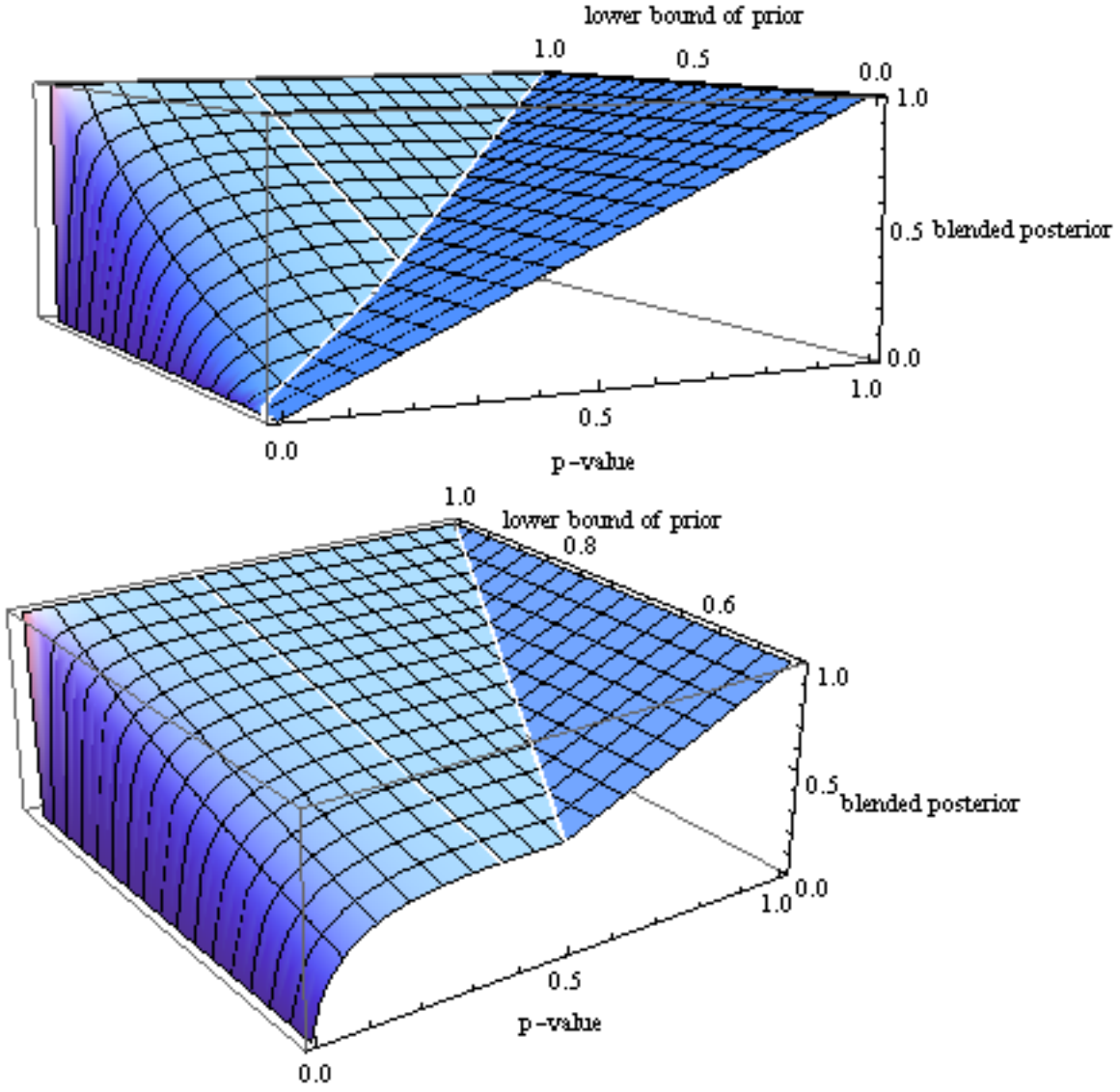

Figure 2: Blended posterior probability that the null hypothesis is true versus the p-value and the lower bound of the prior probability that the null hypothesis is true. The top plot displays the full domain, half of which is shown in the bottom plot.

2 General theory

2.1 Preliminary notation and definitions

Denote the observed data set, typically a vector or matrix of observations, by x, a member of a set $\mathcal{X}$ that is endowed with a σ-algebra $\mathfrak{X}$. 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 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 Section 3.2.1, is a parameter distribution from which confidence intervals and p-values may be extracted. As such, it facilitates blending typical frequentist procedures with Bayesian procedures.)

2.1.1 Bayesian posteriors

With some measurable space $\left(\dot{\Theta}_*, \dot{\mathcal{A}}_*\right)$ for parameter values in $\dot{\Theta}_*$, let $\mathcal{P}_*^{\text{prior}}$ denote a set of probability distributions on $\left(\mathcal{X} \times \dot{\Theta}_*, \mathfrak{X} \otimes \dot{\mathcal{A}}_*\right)$. Any distribution in $\mathcal{P}_*^{\text{prior}}$ is called a *prior (distribution)*, understood in the broad sense of a model that includes the possible likelihood functions as well as the parameter distribution. It encodes the constraints and other information available about the parameter before observing x.

On the other hand, any distribution of a parameter is called a *posterior (distribution)* if it depends on x. For some $P_*^{\text{prior}} \in \mathcal{P}_*^{\text{prior}}$, an example of a posterior distribution on $\left(\dot{\Theta}_*, \dot{\mathcal{A}}_*\right)$ is $\dot{P}_* = P_*^{\text{prior}}(\bullet | X = x)$, where X is a random variable of a distribution on $(\mathcal{X}, \mathfrak{X})$ that is determined by P_*^{prior}. $\dot{P}_*$ 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 $\dot{\mathcal{P}}_* = \left\{P_*^{\text{prior}}(\bullet | X = x) : P_*^{\text{prior}} \in \mathcal{P}_*^{\text{prior}}\right\}$ of Bayesian posteriors on $\left(\dot{\Theta}_*, \dot{\mathcal{A}}_*\right)$ may be considered the "knowledge base." For a set $\dot{\Theta}$, if $\dot{\tau} : \dot{\Theta}_* \to \Theta$ is an $\dot{\mathcal{A}}_*$-measurable map and if $\dot{\theta}_*$ has distribution $\dot{P}_* \in \dot{\mathcal{P}}_*$,

then $\dot{\theta} = \dot{\tau}\left(\dot{\theta}_*\right)$, referred to as an *inferential target of* $\dot{P}_*$, has induced probability space $\left(\Theta, \mathcal{A}, \dot{P}\right)$. The set

$$\dot{\mathcal{P}} = \left\{\dot{P} : \dot{\tau}\left(\dot{\theta}_*\right) \sim \dot{P}, \dot{\theta}_* \sim \dot{P}_* \in \dot{\mathcal{P}}_*\right\}$$

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

Example 1. In the hypothesis test of Section 1.2, $\dot{\theta} = 0$ if the null hypothesis that $\dot{\theta}_* = 0$ is true and $\dot{\theta} = 1$ if the alternative hypothesis that $\dot{\theta}_* \neq 0$ is true, where $\dot{\theta}_*$ and $\dot{\theta}$ are random variables with distributions respectively defined on the Borel space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and the discrete space $\left(\{0,1\}, 2^{\{0,1\}}\right)$, where $2^{\{0,1\}}$ is the power set of $\{0,1\}$. Thus, in this case, $\dot{\tau}$ is the indicator function $1_{(-\infty,0)\cup(0,\infty)} : \mathbb{R} \to \{0,1\}$, yielding $\dot{\theta} = 1_{(-\infty,0)\cup(0,\infty)}\left(\dot{\theta}_*\right)$. Section 3 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, 2012a) of the set. Perhaps the best known inference process for a discrete parameter set Θ is that of the *maximum entropy principle*, which would select a member of $\dot{\mathcal{P}}$ such that it has higher entropy than any other member of the set (see Remark 2).

This paper presents a wide class of inference processes such that each transforms $\dot{\mathcal{P}}$ 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.

2.1.2 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 Section 2.1.1, with the double-dot $\ddot{\bullet}$ replacing the single-dot $\dot{\bullet}$. Let $\ddot{\mathcal{P}}_*$ represent a set of posterior distributions on some measurable space $\left(\ddot{\Theta}_*, \ddot{\mathcal{A}}_*\right)$, and let $\ddot{\mathfrak{P}}_*$ represent a set of such sets. For instance, considering any $\ddot{P}_*$ in $\ddot{\mathcal{P}}_*$, $\ddot{P}_*$ may be a fiducial posterior (Bickel, 2012d), a confidence posterior (a fiducial-like distribution to be defined precisely in Section 3.2), a generalized fiducial posterior of Hannig (2009), or even a Bayesian posterior based on an improper prior. (In the first case, nested confidence intervals 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 confidence intervals (Bickel, 2012a).) Suppose there exists a function $\ddot{\boldsymbol{\tau}} : \ddot{\mathfrak{P}}_* \rightarrow \Theta$ such that $\ddot{P}$, the probability distribution of $\ddot{\boldsymbol{\tau}}\left(\ddot{\mathcal{P}}_*\right)$, is defined on $(\Theta, \mathcal{A})$. $\ddot{P}$ is called the *benchmark posterior (distribution)*, and $\ddot{\theta} = \ddot{\boldsymbol{\tau}}\left(\ddot{\mathcal{P}}_*\right)$ is the *inferential target of* $\ddot{\mathcal{P}}_*$. It follows that $\ddot{P}$ is in $\mathcal{P}$ but not necessarily in $\dot{\mathcal{P}}$.

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

The function $\ddot{\boldsymbol{\tau}}$ allows consideration of a set of possible benchmark posteriors by

transforming it to a single benchmark posterior defined on $(\Theta, \mathcal{A})$, the same measurable space as the above Bayesian posteriors of $\dot{\theta}$. Since that function is unusual, two ways to compose it will now be explained.

Example 3. Consider the inference process $\ddot{\Pi} : \ddot{\mathfrak{P}}_* \to \mathcal{P}_*$, where $\mathcal{P}_*$ is the set of all probability distributions on $\left(\ddot{\Theta}_*, \ddot{\mathcal{A}}_*\right)$. Define the random variable $\ddot{\theta}_*$ to have distribution $\ddot{\Pi}\left(\ddot{\mathcal{P}}_*\right)(\bullet) = \ddot{\Pi}\left(\ddot{\mathcal{P}}_*\right)$. If $\ddot{\tau} : \ddot{\Theta}_* \to \Theta$ is an $\ddot{\mathcal{A}}_*$-measurable function, then $\ddot{\theta} = \ddot{\tau}\left(\ddot{\theta}_*\right)$ is the inferential target of $\ddot{\mathcal{P}}_*$. Further, the distribution $\ddot{P}$ of $\ddot{\theta}$ 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 $\ddot{\tau}$. Let $\ddot{\mathcal{P}}$ denote the subset of $\mathcal{P}$ consisting of all distributions of the parameters transformed by $\ddot{\tau}$:

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

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

2.2 Blended inference

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

$$I\left(P||Q\right) = \int dP \log\left(\frac{dP}{dQ}\right) \tag{3}$$

if Q dominates P $(P \ll Q)$ and $I\left(P||Q\right) = \infty$ otherwise, according to a measure-theoretic definition of relative entropy (Kakihara, 1999, pp. 49-52). $I\left(P||Q\right)$ is also

known as cross entropy, I-divergence, information for discrimination, and Kullback-Leibler divergence. Other measures of information may also be used (Remark 3). For any posteriors $\dot{P} \in \dot{\mathcal{P}}$ and $Q \in \mathcal{P}$, the *inferential gain* $I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right)$ *of* Q *relative to* $\ddot{P}$ *given* $\dot{P}$ is the amount of information gained by making inferences on the basis of Q instead of the benchmark posterior $\ddot{P}$:

$$I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right) = I\left(\dot{P}||\ddot{P}\right) - I\left(\dot{P}||Q\right).$$

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

$$\dot{\mathcal{P}}\left(\ddot{P}\right) = \left\{\dot{P} \in \dot{\mathcal{P}} : I\left(\dot{P}||\ddot{P}\right) < \infty\right\}, \tag{4}$$

which is nonempty by assumption. (The assumption is not necessary under the generalization described in Remark 4.)

The *blended posterior (distribution)* $\widetilde{P}$ is the probability distribution on $(\Theta, \mathcal{A})$ that maximizes the inferential gain relative to the benchmark posterior given the worst-case posterior restricted by the constraints that defined $\dot{\mathcal{P}}$ and $\dot{\mathcal{P}}\left(\ddot{P}\right)$:

$$\inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P}||\ddot{P} \rightsquigarrow \widetilde{P}\right) = \sup_{Q \in \mathcal{P}} \inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P}||\ddot{P} \rightsquigarrow Q\right), \tag{5}$$

where the supremum and infinum over any set including an indeterminate number are ∞ and $-\infty$, respectively (Topsøe, 2007). Inferences based on $\widetilde{P}$ are blended in the sense that they depend on both $\dot{\mathcal{P}}$ and $\ddot{P}$ in the ways to be specified in Section 2.3.

The main result of Theorem 2 of Topsøe (2007) gives a simply stated solution of the optimization problem of equation (5) under broad conditions.

Proposition 1. *If* $I\left(\dot{P}||\ddot{P}\right) < \infty$ *for some* $\dot{P} \in \dot{\mathcal{P}}$ *and if* $\dot{\mathcal{P}}\left(\ddot{P}\right)$ *is convex, then the*

blended posterior $\widetilde{P}$ *is the probability distribution in* $\dot{\mathcal{P}}$ *that minimizes the information divergence with respect to the benchmark posterior:*

$$I\left(\widetilde{P}||\ddot{P}\right) = \inf_{\dot{P}\in\dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P}||\ddot{P}\right). \tag{6}$$

Proof. Topsøe (2007) proved the result from inequalities of information theory given the additional stated condition of his Theorem 2 that $I\left(\dot{P}||\ddot{P}\right) < \infty$ for all $\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)$. (See Remark 4.) The condition that $I\left(\dot{P}||\ddot{P}\right) < \infty$ for some $\dot{P} \in \dot{\mathcal{P}}$ and the above definition of $\dot{\mathcal{P}}\left(\ddot{P}\right)$ ensure that the condition is met. □

Alternatively, the minimization of information divergence may define $\widetilde{P}$ rather than result from its definition in terms of the game (Remark 5).

2.3 Properties of blended inference

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

1. **Complete knowledge of the prior.** If $\dot{\mathcal{P}}$ has a single member $\dot{P}$, then $B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right) = \dot{P}$.

2. **Negligible knowledge of the prior.** If $\ddot{P} \in \dot{\mathcal{P}}$ and if $\dot{\mathcal{P}}$ has at least two members, then $B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right) = \ddot{P}$.

3. **Continuum between extremes.** For any $D \geq 0$ and any $\mathcal{P}^\star \subseteq \mathcal{P}$ such that

$$\sup_{P\in\mathcal{P}^\star, \dot{P}\in\dot{\mathcal{P}}\left(\ddot{P}\right)} \left| I\left(P||\ddot{P}\right) - I\left(\dot{P}||\ddot{P}\right)\right| \leq D \tag{7}$$

and such that $\dot{\mathcal{P}}\left(\ddot{P}\right) \cup \mathcal{P}^{\star}$ is convex,

$$\left| I\left(B\left(\bullet; \dot{\mathcal{P}} \cup \mathcal{P}^{\star}, \ddot{P}\right) || \ddot{P}\right) - I\left(B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right) || \ddot{P}\right) \right| \leq D. \tag{8}$$

Theorem 1. *The blended posterior $\widetilde{P}$ blends the set $\dot{\mathcal{P}}$ of Bayesian posteriors with the benchmark posterior $\ddot{P}$ for inference about the parameter in Θ.*

Proof. Since the criteria are only required under the conditions of Proposition 1, it will suffice to prove that the criteria follow from equation (6). If $\dot{\mathcal{P}}$ has a single member $\dot{P}$, then equation (6) implies that $\widetilde{P} = \dot{P}$, thereby ensuring Criterion 1. Similarly, if $\ddot{P} \in \dot{\mathcal{P}}$, then equation (6) implies that $\widetilde{P} = \ddot{P}$, thus proving that Criterion 2 is met. Assume, contrary to Criterion 3, that there exist a $D \geq 0$ and a $\mathcal{P}^{\star} \subseteq \mathcal{P}$ such that $\dot{\mathcal{P}}\left(\ddot{P}\right) \cup \mathcal{P}^{\star}$ is convex, equation (7) is true, and equation (8) is false with $B\left(\bullet; \dot{\mathcal{P}} \cup \mathcal{P}^{\star}, \ddot{P}\right)$ and $B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right)$ equal to the blended posteriors respectively using $\dot{\mathcal{P}} \cup \mathcal{P}^{\star}$ and $\dot{\mathcal{P}}$ as the sets of Bayesian posteriors. Then equation (6) can be written as

$$I\left(B\left(\bullet; \dot{\mathcal{P}} \cup \mathcal{P}^{\star}, \ddot{P}\right) || \ddot{P}\right) = \inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right) \cup \mathcal{P}^{\star}} I\left(\dot{P} || \ddot{P}\right);$$

$$I\left(B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right) || \ddot{P}\right) = \inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P} || \ddot{P}\right).$$

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

$$\left| I\left(B\left(\bullet; \dot{\mathcal{P}} \cup \mathcal{P}^{\star}, \ddot{P}\right) || \ddot{P}\right) - I\left(B\left(\bullet; \dot{\mathcal{P}}, \ddot{P}\right) || \ddot{P}\right) \right| =$$

$$\inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P} || \ddot{P}\right) - \inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P} || \ddot{P}\right) \wedge \inf_{P \in \mathcal{P}^{\star}} I\left(P || \ddot{P}\right),$$

which cannot exceed $\inf_{\dot{P} \in \dot{\mathcal{P}}\left(\ddot{P}\right)} I\left(\dot{P} || \ddot{P}\right) - \inf_{P \in \mathcal{P}^{\star}} I\left(P || \ddot{P}\right)$ and thus, according to equation (7), cannot exceed D. Therefore, the above assumption that equation (8) is false is contradicted, thereby establishing satisfaction of Criterion 3. □

Example 5. Suppose the set of possible priors consists of a single frequency-matching prior, i.e., a prior that leads to 95% posterior credible intervals that are equal to 95% confidence intervals, etc. If the benchmark posterior is the confidence posterior that yields the same confidence intervals, then it is the Bayesian posterior distribution corresponding to the prior. In that case, the blended distribution is equal to that Bayesian/confidence posterior. Thus, the first condition of blended inference applies. The second condition would instead apply if the set of possible priors contained at least one other prior in addition to the frequency-matching prior.

Criterion 3 is much stronger than the heuristic idea of continuity introduced in Section 1.1. Its use of information divergence can be generalized to other measures of divergence (Remark 3).

3 Testing a two-sided null hypothesis

A fertile field of application for the theory of Section 2 is that of testing hypotheses, as outlined in Section 1.2. Building on Example 1, this section provides methodology for a wide class of models used in hypothesis testing.

3.1 A bound on the Bayesian posterior

Defining that class in terms of the concepts of Section 2.1.1 requires additional notation. For a continuous sample space $\mathcal{X}$ and a function $p : \mathcal{X} \to [0,1]$ such that $p(X) \sim \mathrm{U}(0,1)$ 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 $\left(\dot{\theta} = 0\right)$ and under the alternative hypothesis $\left(\dot{\theta} = 1\right)$, respectively. For the observed x, the likelihood ratio $f_0(p(x)) / f_1(p(x))$ is

called the *Bayes factor* since, for a prior distribution P_*^{prior}, Bayes's theorem gives

$$\frac{\psi\left(p\left(x\right)\right)}{1-\psi\left(p\left(x\right)\right)}=\frac{P_*^{\text{prior}}\left(\dot{\theta}=0\right)}{P_*^{\text{prior}}\left(\dot{\theta}=1\right)}\frac{f_0\left(p\left(x\right)\right)}{f_1\left(p\left(x\right)\right)}, \tag{9}$$

where $\psi\left(p\left(x\right)\right)=P_*^{\text{prior}}\left(\dot{\theta}=0|p\left(X\right)=p\left(x\right)\right)$. In a parametric setting, $f_1\left(p\left(x\right)\right)$ would be the likelihood integrated over the prior distribution conditional on the alternative hypothesis.

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

Sellke et al. (2001) obtained the following lower bound $\underline{b}\left(p\left(x\right)\right)$ of the Bayes factor $b\left(x\right)$.

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

$$b\left(p\left(x\right)\right)=\frac{f_0\left(p\left(x\right)\right)}{f_1\left(p\left(x\right)\right)}\geq\underline{b}\left(p\left(x\right)\right)=\begin{cases}-ep\left(x\right)\log p\left(x\right) & \text{if } p\left(x\right)<1/e;\\ 1 & \text{if } p\left(x\right)\geq 1/e.\end{cases} \tag{10}$$

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 $\pi_0=P_*^{\text{prior}}\left(\dot{\theta}=0\right)$, the prior probability that the null hypothesis is true, in addition to the hazard rate condition. Specifically, π_0 is known

to have $\underline{\pi}_0 \in [0,1]$ as a lower bound. Thus, rearranging equation (9) as

$$\psi\left(p\left(x\right)\right) = \left(1 + \left(\frac{1-\pi_0}{\pi_0 b_0\left(p\left(x\right)\right)}\right)\right)^{-1},$$

a lower bound on $\psi\left(p\left(x\right)\right)$ is

$$\underline{\Pr}\left(H_0 | p\left(X\right) = p\left(x\right)\right) = \underline{\psi}\left(p\left(x\right)\right) = \left(1 + \left(\frac{1-\underline{\pi}_0}{\underline{\pi}_0 \underline{b}\left(p\left(x\right)\right)}\right)\right)^{-1},$$

leading to equation (1).

Let $\mathcal{P}$ consist of all probability distributions on $(\Theta, \mathcal{A}) = \left(\{0,1\}, 2^{\{0,1\}}\right)$. The subset $\dot{\mathcal{P}}$ consists of all $\dot{P} \in \mathcal{P}$ such that $\dot{P}\left(\dot{\theta} = 0\right) \geq \underline{\psi}\left(p\left(x\right)\right)$.

3.2 A confidence benchmark posterior

3.2.1 Confidence posterior theory

The following parametric framework facilitates the application of Section 2.1.2 to hypothesis testing. The observation x is an outcome of the random variable $\ddot{X}$ of probability space $(\mathcal{X}, \mathfrak{X}, P_{\theta_*,\lambda_*})$, where the interest parameter $\theta_* \in \ddot{\Theta}_*$ and a nuisance parameter λ_* (in some set $\ddot{\Lambda}_*$) are unknown. Let $S : \ddot{\Theta}_* \times \mathcal{X} \to [0,1]$ and $t : \mathcal{X} \times \ddot{\Theta}_* \to \mathbb{R}$ denote functions such that $S\left(\bullet; x\right)$ is a distribution function, $S\left(\theta_*; X\right) \sim \mathrm{U}\left(0,1\right)$, and

$$S\left(\theta_*; x\right) = P_{\theta_*,\lambda_*}\left(t\left(\ddot{X}; \theta_*\right) \geq t\left(x; \theta_*\right)\right)$$

for all $x \in \mathcal{X}$, $\theta_* \in \ddot{\Theta}_*$, and $\lambda_* \in \ddot{\Lambda}_*$. S is known as a *significance function*, and t as a *pivot* or test statistic. It follows that $p\left(x\right) = S\left(0; x\right)$ is a p-value for testing the hypothesis that $\theta_* = 0$ and that $\left[S^{-1}\left(\alpha; X\right), S^{-1}\left(\beta; X\right)\right]$ is a $\left(\beta - \alpha\right)100\%$ confidence interval for θ_* given any $\alpha \in [0,1]$ and $\beta \in [\alpha, 1]$. Thus, whether a significance function is found from p-values over a set of simple null hypotheses or instead from a set of nested confidence intervals, it contains the information needed to derive either

(Schweder and Hjort, 2002; Singh et al., 2007; Bickel, 2012a, 2011a).

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

$$\begin{aligned} \beta - \alpha &= P_{\theta_*,\lambda_*}\left(\theta_* \in \left[S^{-1}(\alpha; X), S^{-1}(\beta; X)\right]\right) \\ &= \ddot{P}_*\left(\ddot{\theta}_* \in \left[S^{-1}(\alpha; x), S^{-1}(\beta; x)\right]\right) \end{aligned}$$

for all $x \in \mathcal{X}$, $\theta_* \in \ddot{\Theta}_*$, and $\lambda_* \in \ddot{\Lambda}_*$. The term "confidence posterior" (Bickel, 2012a, 2011a) is preferred here over the usual term "confidence distribution" (Schweder and Hjort, 2002) to emphasize its use as an alternative to Bayesian posterior distributions. Polansky (2007), Singh et al. (2007), and Bickel (2012a) provide generalizations to vector parameters of interest. Extensions based on multiple comparison procedures are provided in Section 4.

3.2.2 A confidence posterior for testing

For the application to two-sided testing of a simple null hypothesis, let $\theta_* = |\theta_{**}|$, the absolute value of a real parameter θ_{**} of interest, leading to $\ddot{\Theta}_* = [0, \infty)$. Then $p(x) = S(0; x)$ is equivalent to a two-tailed p-value for testing the hypothesis that $\theta_{**} = 0$. Since $\ddot{P}_*\left(\ddot{\theta}_* \leq 0\right) = S(0; x)$ and since $\ddot{P}_*\left(\ddot{\theta}_* \leq 0\right) = \ddot{P}_*\left(\ddot{\theta}_* = 0\right)$, it follows that $p(x) = \ddot{P}_*\left(\ddot{\theta}_* = 0\right)$, 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 θ_{**} (Bickel, 2012d).

If $\ddot{P}_*$ is the only confidence posterior under consideration, then $\ddot{\mathcal{P}}_* = \left\{\ddot{P}_*\right\}$, and there is no need for an inference process. Following the terminology of Example 3,

$\ddot{\tau} : \ddot{\Theta}_* \to \Theta$ is defined by $\ddot{\tau}\left(\ddot{\theta}_*\right) = 1_{(0,\infty)}\left(\ddot{\theta}_*\right)$. By implication, $\ddot{\theta} = 0$ if $\ddot{\theta}_* = 0$ and $\ddot{\theta} = 1$ if $\ddot{\theta}_* > 0$. Thus, $p(x) = \ddot{P}_*\left(\ddot{\theta}_* = 0\right)$ ensures that $\ddot{P}\left(\ddot{\theta} = 0\right) = p(x)$, which in turn implies $\ddot{P}\left(\ddot{\theta} = 1\right) = 1 - p(x)$.

Example 6. In the various t-tests, θ_* 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 $\ddot{P}\left(\ddot{\theta} = 0\right)$ and $p(x)$. Special cases of this $\ddot{P}$ have been presented as fiducial distributions (van Berkum et al. (1996);Bickel, 2011c).

3.3 A blended posterior for testing

This subsection blends the above set $\dot{\mathcal{P}}$ of Bayesian posteriors with the above confidence posterior $\ddot{P}$ as prescribed by Section 2.2. Gathering the results of Sections 3.1 and 3.2,

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

$$\ddot{P}\left(\ddot{\theta} = 0\right) = p(x) = 1 - \ddot{P}\left(\ddot{\theta} = 1\right).$$

Equation (4) then implies that

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

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

$$\widetilde{P}(\theta = 0) = \begin{cases} \underline{\psi}(p(x)) & \text{if } p(x) < \underline{\psi}(p(x)) \\ p(x) & \text{if } p(x) \geq \underline{\psi}(p(x)) \end{cases}, \tag{11}$$

where θ is the random variable of distribution $\widetilde{P}$. With the identities $\underline{\psi}(p(x)) = \underline{\Pr}(H_0|p(X) = p(x))$ and $\widetilde{P}(\theta = 0) = \Pr(H_0; p(x))$ and with the establishment of equation (1) by Section 3.1, equation (11) verifies the claim of equation (2) made in Section 1.2.

4 Testing multiple hypotheses

4.1 Blending multiple comparison procedures

The single-test notation of Section 3 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 (§3.2) 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): $\ddot{P}\left(\ddot{\theta}_j = 0\right) = \pi(x_j)$ 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 jth of N null hypotheses conditional on the p-value:

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

where $\psi(\bullet)$ is the function that satisfies equation (9). In literature on multiple testing influenced by Benjamini and Hochberg (1995), ψ_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 in a model of the physical system, not an epistemological probability in a model of an intelligent agent. (The letter ψ stands for $\psi\varepsilon\upsilon\delta\acute{\eta}\varsigma$ (*pseudēs*), the Greek word translated as "false.")

Let $\widehat{\psi}_j$ denote an estimator of ψ_j that is conservative enough that the interval $\left[0, \widehat{\psi}_j\right]$ can be considered the set of plausible posterior probabilities of the jth null hypothesis. Conservatism is here understood as positive bias of some type, e.g., positive median bias (Bickel, 2011b). The reasoning of Section 3.3, except with the lower bound replaced by the upper bound, results in

$$\dot{\mathcal{P}}_j = \left\{ \dot{P} \in \mathcal{P} : \dot{P}\left(\dot{\theta}_j = 0\right) \leq \widehat{\psi}_j \right\}; \tag{12}$$

$$\dot{\mathcal{P}}_j\left(\ddot{P}\right) = \left\{ \dot{P} \in \mathcal{P} : 0 < \dot{P}\left(\dot{\theta}_j = 0\right) \leq \widehat{\psi}_j \right\}. \tag{13}$$

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

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* $\widehat{\psi}_1 = 1$ *whenever* $N = 1$ *and that equations* (12)-(13) *hold. It follows that*

$$N = 1 \implies \widetilde{P}(\theta_1 = 0) = p(x_1).$$

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 Section 4.2.2.

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.

4.2 Biostatistics applications

In both of this section's applications of the strategy of blending multiple comparison procedures (§4.1), 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 (2011b) 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.

4.2.1 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 (§4.1) was applied to

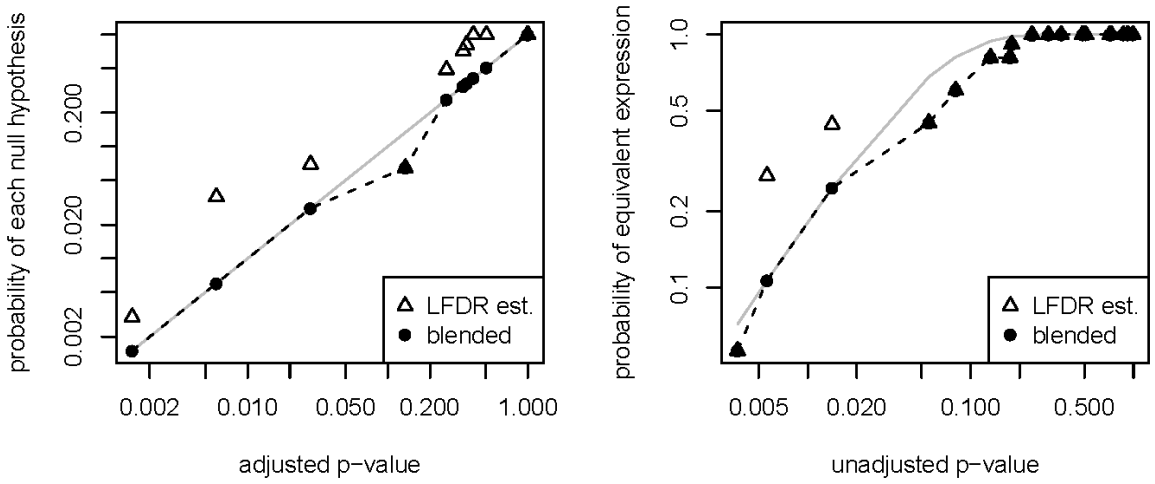

Figure 3: Blending of empirical Bayes estimators with adjusted p-values of two biostatistics applications. The dashed line represents the blended posterior probabilities of the null hypothesis, and the gray line represents p-values adjusted by the independent-statistic method of Sidak (1967). The plot on the left is based on the 15 p-values related to cardiac disease (§4.1), and that on the right is based on 20 randomly sampled microarray p-values (§4.2.2). Each horizontal axis is an adjusted or unadjusted p-value, depending on which is visually more informative.

those p-values. The left-hand side of Figure 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.

4.2.2 Application to gene expression

The multiple comparison procedure of Section 4.1 is also illustrated with a gene expression data set (Alba et al., 2005) in which x_j is a tuple of 6 logarithms of the measured ratios of mutant tomato expression to wild-type tomato expression of the jth gene.

The plot on the right-hand side of Figure 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 6 biological replicates at three 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).

5 Remarks

Remark 1. As mentioned in Section 1.1, the use of Bayes's theorem with proper priors need not involve subjective interpretations of probability. The set of posteriors may be determined by interval constraints on the corresponding priors without any requirement that they model levels of belief (Weichselberger, 2000; Augustin, 2002, 2004). However, subjective applications of blended inference are also possible. While the framework was developed 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. By allowing the partial order of agent preferences, imprecise probability theories need not assume the existence of any true prior (Walley, 1991; Coletti and Scozzafava, 2002). As often happens, the same mathematical framework is subject to very different philosophical interpretations.

Remark 2. Technically, the principle of maximum entropy (Paris, 1994; Paris and Vencovská, 1997) mentioned in Section 2.1.1 could be used if Θ is finite or countable infinite. However, unlike the proposed methodology, that practice is equivalent to making the benchmark posterior $\ddot{P}$ depend on the function $\dot{\tau}$ that maps a parameter space to Θ 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 Θ. For example, averaging with respect to the Lebesgue measure, as Bickel (2012a) did with confidence posteriors,

leads to $\left(1+\underline{\psi}\left(p\left(x\right)\right)\right)/2$ as the posterior probability of the null hypothesis under the assumptions of Section 3.1. Remark 5 discusses a more tenable version of the maximum entropy principle for blended inference.

Remark 3. 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 4. A generalization of Section 2 in a different direction from that of Remark 3 replaces each "$\inf_{\dot{P}\in\dot{\mathcal{P}}\left(\ddot{P}\right)}$" of equation (5) with "$\inf_{\dot{P}\in\dot{\mathcal{P}}}$." For that optimization problem, Theorem 2 of Topsøe (2007) has the condition that $\dot{P}\in\dot{\mathcal{P}} \implies I\left(\dot{P}||\ddot{P}\right) < \infty$ in addition to the convexity of $\dot{\mathcal{P}}$ that Proposition 1 of the present paper requires. Thus, in that formulation, the blended posterior $\widetilde{P}$ need not satisfy equation (6) even if $\dot{\mathcal{P}}$ is convex.

Remark 5. A posterior distribution $\widetilde{P}$ that *is defined by*

$$I\left(\widetilde{P}||\ddot{P}\right) = \inf_{\dot{P}\in\dot{\mathcal{P}}} I\left(\dot{P}||\ddot{P}\right) \tag{14}$$

satisfies the desiderata of Section 2.3 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 2). For example, an agent that makes inferences on the basis of a confidence posterior $\ddot{P}$ in the absence of parameter constraints would, upon learning such constraints in the form of $\dot{\mathcal{P}}$, update that posterior to $\widetilde{P}$ by maximizing entropy relative to that posterior, at least according to the reasonable axioms of Paris and Vencovská (1997). However, the optimization problem

of equation (5) seems even more compelling in this context and defines $\widetilde{P}$ even when no distribution satisfying equation (14) exists.

Acknowledgments

I thank Xuemei Tang for sending me 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, G. B., Tanksley, S. D., Giovannoni, J. J., 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. Statistical Papers 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 fur Angewandte Mathematik und Mechanik 84 (10-11), 678–687.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

Berger, J., Brown, L., Wolpert, R., 1994. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis-testing. Annals of Statistics 22 (4), 1787–1807.

Berger, J. O., 1984. Robustness of Bayesian analyses. Studies in Bayesian econometrics. North-Holland.

Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Berger, J. O., Sellke, T., 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association 82, 112–122.

Bickel, D. R., 2011a. Estimating the null distribution to adjust observed confidence levels for genome-scale screening. Biometrics 67, 363–370.

Bickel, D. R., 2011b. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Technical Report, Ottawa Institute of Systems Biology, arXiv:1106.4490.

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

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

Bickel, D. R., 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, D. R., 2012c. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. International Journal of Approximate Reasoning 53, 880–891.

Bickel, D. R., 2012d. A prior-free framework of coherent inference and its derivation of simple shrinkage estimators. Technical Report, University of Ottawa, available from http://hdl.handle.net/10393/23093.

Coletti, C., Scozzafava, R., 2002. Probabilistic Logic in a Coherent Setting. Kluwer, Amsterdam.

Cox, D. R., 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. Statistics & Decisions 7, 333–361.

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". Journal of the American Statistical Association 105, 1067–1069.

Fraser, D. A. S., 2004. Ancillaries and conditional inference. Statistical Science 19, 333–351.

Fraser, D. A. S., Reid, N., 1990. Discussion: An ancillarity paradox which appears in multiple linear regression. The Annals of Statistics 18, 503–507.

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

Gilboa, I., Schmeidler, D., 1989. Maxmin expected utility with non-unique prior. Journal of Mathematical Economics 18 (2), 141–153.

Good, I., 1983. Good Thinking: the Foundations of Probability and Its Applications. G - Reference, Information and Interdisciplinary Subjects Series. University of Minnesota Press.

Good, I. J., 1952. Rational decisions. Journal of the Royal Statistical Society B 14, 107–114.

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

Hannig, J., 2009. On generalized fiducial inference. Statistica Sinica 19, 491–544.

Harremoës, P., Topsøe, F., 2001. Maximum entropy fundamentals. Entropy 3 (3), 191–226.

Kakihara, Y., 1999. Abstract Methods in Information Theory (Series on Multivariate Analysis, Volume 4). World Scientific Pub Co Inc, Singapore.

Lindley, D. V., 1957. A statistical paradox. Biometrika 44, pp. 187–192.

Neuhaus, K. L., von Essen, R., Tebbe, U., Vogt, A., Roth, M., Riess, M., Niederer, W., Forycki, F., Wirtzfeld, A., Maeurer, W., 1992. Improved thrombolysis in acute myocardial infarction with front-loaded administration of alteplase: results of the rt-PA-APSAC patency study (TAPS). Journal of the American College of Cardiology 19, 885–91.

Padilla, M., Bickel, D. R., 2012. Empirical Bayes methods corrected for small numbers of tests. Technical Report, Ottawa Institute of Systems Biology, arXiv:1009.5981.

Paris, J., Vencovská, A., 1997. In defense of the maximum entropy inference process. International Journal of Approximate Reasoning 17 (1), 77 – 103.

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

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

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

Robbins, H., 1951. Asymptotically subminimax solutions of compound statistical decision problems. Proc. Second Berkeley Symp. Math. Statist. Probab. 1, 131–148.

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

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

Sidak, Z., 1967. Rectangular confidence regions for means of multivariate normal distributions. Journal of the American Statistical Association 62 (318), 626–633.

Singh, K., Xie, M., Strawderman, W. E., 2007. Confidence distribution (CD) –

distribution estimator of a parameter. IMS Lecture Notes Monograph Series 2007 54, 132–150.

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

Topsøe, F., 2004. Entropy and equilibrium via games of complexity. Physica A 340 (1-3), 11–31.

Topsøe, F., 2007. Information theory at the service of science. In: Csiszár, I., Katona, G. O. H., Tardos, G., Wiener, G. (Eds.), Entropy, Search, Complexity. Bolyai Society Mathematical Studies. Springer Berlin Heidelberg, pp. 179–207.

van Berkum, E., Linssen, H., Overdijk, D., 1996. Inference rules and inferential distributions. Journal of Statistical Planning and Inference 49, 305–317.

Vidakovic, B., 2000. Gamma-minimax: A paradigm for conservative robust Bayesians. Robust Bayesian Analysis. Springer, New York, pp. 241–260.

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

Weichselberger, K., 2000. The theory of interval-probability as a unifying concept for uncertainty. International Journal of Approximate Reasoning 24 (2-3), 149–170.

Weichselberger, K., 2001. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept. Physica-Verlag, Heidelberg.

Westfall, P. H., 2010. Comment on B. Efron, "Correlated z-values and the accuracy of large-scale statistical estimates". Journal of the American Statistical Association 105, 1063–1066.

Westfall, P. H., Johnson, W. O., Utts, J. M., 1997. A Bayesian perspective on the Bonferroni adjustment. Biometrika 84, 419–427.

Yuan, B., 2009. Bayesian frequentist hybrid inference. Annals of Statistics 37, 2458–2501.