International Journal of Approximate Reasoning 53 (2012) 880–891

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Game-theoretic probability combination with applications to resolving conflicts between statistical methods

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, Canada K1H 8M5

ARTICLE INFO

Article history:
Received 25 January 2012
Received in revised form 11 April 2012
Accepted 13 April 2012
Available online 25 April 2012

Keywords:
Combining probabilities
Combining probability distributions
Confidence distribution
Linear opinion pool
Minimax redundancy
Redundancy-capacity theorem

ABSTRACT

In the typical analysis of a data set, a single method is selected for statistical reporting even when equally applicable methods yield very different results. Examples of equally applicable methods can correspond to those of different ancillary statistics in frequentist inference and of different prior distributions in Bayesian inference. More broadly, choices are made between parametric and nonparametric methods and between frequentist and Bayesian methods. Rather than choosing a single method, it can be safer, in a game-theoretic sense, to combine those that are equally appropriate in light of the available information. Since methods of combining subjectively assessed probability distributions are not objective enough for that purpose, this paper introduces a method of distribution combination that does not require any assignment of distribution weights. It does so by formalizing a hedging strategy in terms of a game between three players: nature, a statistician combining distributions, and a statistician refusing to combine distributions. The optimal move of the first statistician reduces to the solution of a simpler problem of selecting an estimating distribution that minimizes the Kullback–Leibler loss maximized over the plausible distributions to be combined. The resulting combined distribution is a linear combination of the most extreme of the distributions to be combined that are scientifically plausible. The optimal weights are close enough to each other that no extreme distribution dominates the others. The new methodology is illustrated by combining conflicting empirical Bayes methods in the context of gene expression data analysis.

1. Introduction

The analysis of biological data often requires choices between methods that seem equally applicable and yet that can yield very different results. This occurs not only with the notorious problems in frequentist statistics of conditioning on one of multiple ancillary statistics and in Bayesian statistics of selecting one of many appropriate priors, but also in choices between frequentist and Bayesian methods, in whether to use a potentially powerful parametric test to analyze a small sample of unknown distribution, in whether and how to adjust for multiple testing, and in whether to use a frequentist model averaging procedure. Today, statisticians simultaneously testing thousands of hypotheses must often decide whether to apply a multiple comparisons procedure using the assumption that the *p*-value is uniform under the null hypothesis (theoretical null distribution) or a null distribution estimated from the data (empirical null distribution). While the empirical null reduces estimation bias in many situations [21], it also increases variance [22] and can substantially increase bias when the data distributions have heavy tails [8]. Without any strong indication of which method can be expected to perform better for a particular data set, combining their estimated false discovery rates or adjusted *p*-values may be the safest approach.

E-mail address: dbickel@uOttawa.ca

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2012.04.002

Emphasizing the reference class problem, Barndorff-Nielsen [3] pointed out the need for ways to assess the evidence in the diversity of statistical inferences that can be drawn from the same data. Previous applications of p-value combination have included combining inferences from different ancillary statistics [29], combining inferences from more robust procedures with those from procedures with stronger assumptions, and combining inferences from different alternative distributions [30]. However, those combination procedures are only justified by a heuristic Bayesian argument and have not been widely adopted. To offer a viable alternative, the problem of combining conflicting methods is framed herein in terms of probability combination.

Most existing methods of automatically combining probability distributions have been designed for the integration of expert opinions. For example, [1,36,53] proposed combining distributions to minimize a weighted sum of Kullback–Leibler divergences from the distributions being combined, with the weights determined subjectively, e.g., by the elicitation of the opinions of the experts who provided the distributions or by the extent to which each expert is considered credible. Under broad conditions, that approach leads to the linear combination of the distributions that is defined by those weights [53,36].

"Linear opinion pools" also result from this *marginalization property*: any linearly combined marginal distribution is the same whether marginalization or combination is carried out first [39]. The marginalization property forbids certain counter-intuitive combinations of distributions, including any combination of distributions that differs in a probability assignment from the unanimous assignment of all distributions combined [16, p. 173]. Combinations violating the marginalization property can be expected to perform poorly as estimators regardless of their appeal as distributions of belief. On the other hand, invariance to reversing the order of Bayesian updating and distribution combination instead requires a "logarithmic opinion pool," which uses a geometric mean in place the arithmetic mean of the linear opinion pool; see, e.g., [5, Section 4.11.1] or [15]. While that property is preferable to the marginalization property from the point of view of a Bayesian agent making decisions on the basis of independent reports of other Bayesian agents, it is less suitable for combining distributions that are highly dependent or that are distribution estimates rather than actual distributions of belief. Genest and Zidek [27] and Cooke [16, Ch. 11] review arguments for and against linear and logarithmic combinations of distributions.

Like those methods, the strategy introduced in this paper is intended for combining distributions based on the same data or information as opposed to combining distributions based on independent data sets. However, to relax the requirement that the distributions be provided by experts, the weights are optimized rather than specified. While the new strategy leads to a linear combination of distributions, the combination hedges by including only the most extreme distributions rather than all of the distributions. In addition, the game leading to the hedging takes into account any known constraints on the true distribution. See Remark 1 on the pivotal role of game theory in laying the foundations of statistics.

The game that generates the hedging strategy is played between three players: the mechanism that generates the true distribution ("Nature"), a statistician who never combines distributions ("Chooser"), and a statistician who is willing to combining distributions ("Combiner"). Nature must select a distribution that complies with constraints known to the statisticians, who want to choose distributions as close as possible to the distribution chosen by Nature. Other things being equal, each statistician would also like to select a distribution that is as much better than that of the other statistician as possible. Thus, each statistician seeks primarily to come close to the truth and secondarily to improve upon the distribution selected by the other statistician. Combiner has the advantage over Chooser that the former may select any distribution, whereas the latter must select one from a given set of the distributions that estimate the true distribution or that encode expert opinion. On the other hand, Combiner is disadvantaged in that the game rules specify that Nature seeks to maximize the gain of Chooser albeit without concern for the gain of Combiner. Since Nature favors Chooser without opposing Combiner, the optimal strategy of Combiner is one of hedging but is less cautious than the minimax strategies that are often optimal for typical two-player zero-sum games against Nature. The distribution chosen according to the strategy of Combiner will be considered the combination of the distributions available to Chooser. The combination distribution is a function not only of the combining distributions but also of the constraints on the true distribution.

Section 2 encodes the game and strategy described above in terms of Kullback–Leibler loss and presents its optimal solution as a general method of combining distributions. The important special case of combining probabilities is then worked out. A framework for using the proposed combination method to resolve method conflicts in point and interval estimation, hypothesis testing, and other aspects of statistical data analysis will be presented in Section 3. The framework is illustrated by applying it to the combination of three false discovery rate methods for the analysis of microarray data in Section 4. Finally, Section 5 and Appendix A collect miscellaneous remarks and proofs, respectively.

2. Framework for combining distributions

Following the usual notation, tuples (ordered sets) are enclosed with parentheses. The open interval between a and b is denoted by $]a, b[$ to avoid confusion with the ordered pair (2-tuple) (a, b), which represents a 2-dimensional vector if a and b are numbers.

2.1. Information-theoretic background

Let $\mathcal{P}$ denote the set of probability distributions on a Borel space $(\Xi, \mathcal{B}(\Xi))$, where $\mathcal{B}(\Xi)$ is the set of all Borel subsets of Ξ. The *information divergence of $P \in \mathcal{P}$ with respect to $Q \in \mathcal{P}$* is defined as

$$D\left(P\|Q\right) = \int dP\left(\xi\right) \log \frac{dP\left(\xi\right)}{dQ\left(\xi\right)}, \tag{1}$$

where dP and dQ are probability density functions of P and Q in the sense of Radon–Nikodym differentiation with respect to the same dominating measure [32]. The integrand follows the $0 \log(0) = 0$ and $0 \log(0/0) = 0$ conventions. $D\left(P\|Q\right)$ is also known as "information for discrimination", "Kullback–Leibler information", "Kullback–Leibler divergence", and "cross entropy". Calling $D\left(P\|Q\right)$ "information divergence" emphasizes its interpretation as the amount of information that would be gained by replacing any distribution Q with the true distribution P. That interpretation accords with calling

$$D\left(P'\|P'' \rightsquigarrow Q\right) = D\left(P'\|P''\right) - D\left(P'\|Q\right) \tag{2}$$

the *information gain* [42], the amount of information gained by using Q rather than $P'' \in \mathcal{P}$ when the true distribution is $P' \in \mathcal{P}$ [cf. 54].

Each member of Φ, some set of real parameter values, corresponds to a probability distribution in the family $\mathcal{P}^\star = \{P_\phi : \phi \in \Phi\}$ such that $\mathcal{P}^\star \subseteq \mathcal{P}$. The distribution

$$\operatorname{cent} \mathcal{P}^\star = \arg \inf_{Q \in \mathcal{P}} \sup_{P^\star \in \mathcal{P}^\star} D\left(P^\star\|Q\right) \tag{3}$$

is called the *centroid* of $\mathcal{P}^\star$ [19, p. 131]. Let $\mathcal{W}$ denote the set of all probability measures on the Borel space $(\Phi, \mathcal{B}(\Phi))$. Reserving the term *prior* for Section 3, members of $\mathcal{W}$ will be called *weighting distributions*. Then $P^W = \int P_\phi dW(\phi)$ defines the *mixture distribution of* $\mathcal{P}^\star$ *with respect to some* $W \in \mathcal{W}$, and

$$W_{\mathcal{P}^\star} = \arg \sup_{W \in \mathcal{W}} \int D\left(P_\phi\|P^W\right) dW(\phi) \tag{4}$$

defines the *weighting distribution induced by* $\mathcal{P}^\star$.

Example 1. In the case of a family of ν distributions, the parameter set can be written as $\Phi = \{\phi_1, \ldots, \phi_\nu\}$ and the weighting distribution induced by $\mathcal{P}^\star$ as the probability measure $W_{\mathcal{P}^\star}$ such that

$$\left(W_{\mathcal{P}^\star}(\phi_1), \ldots, W_{\mathcal{P}^\star}(\phi_\nu)\right) = \arg \sup_{(w_1,\ldots,w_\nu)\in[0,1]^\nu : \sum_{i=1}^{\nu} w_i = 1} \sum_{i=1,\ldots,\nu} w_i D\left(P_{\phi_i}\| \sum_{j=1,\ldots,\nu} w_j P_{\phi_j}\right).$$

Shulman and Feder [51] proved that, for all $i = 1, \ldots, \nu$,

$$W_{\mathcal{P}^\star}(\phi_i) \leq 1 - e^{-1} \doteq 63\%, \tag{5}$$

where e is the base of the natural logarithm.

The next known result will be used to determine the optimal move in the game of combining distributions that was mentioned in Section 1.

Lemma 1. *The centroid of any nonempty* $\mathcal{P}^\star \subseteq \mathcal{P}$ *is the mixture distribution of* $\mathcal{P}^\star$ *with respect to the weighting distribution induced by* $\mathcal{P}^\star$:

$$\operatorname{cent} \mathcal{P}^\star = P^{W_{\mathcal{P}^\star}}.$$

Proof. Two different proofs appear in Haussler [32] and in Grünwald and Dawid [31]. For some history of this result, see Remark 2. □

2.2. Distribution–combination game

The game sketched in Section 1 will now be specified in the above notation. Two sets constrain moves in the game: the *plausible set* $\dot{\mathcal{P}}$ is the subset of $\mathcal{P}$ consisting of given *plausible distributions*, and $\ddot{\mathcal{P}} \subseteq \mathcal{P}$ consists of given *combining distributions*. The move of Nature is a distribution $\dot{P} \in \dot{\mathcal{P}}$; the move of Chooser is a distribution $\ddot{P} \in \ddot{\mathcal{P}}$; the move of Combiner is a distribution $P^+ \in \mathcal{P}$. Chooser and Combiner are called *statisticians*. If P_1 is the move of one statistician and P_2 is that of the other, then the latter seeks primarily to minimize the disagreement between P_2 and $\dot{P}$ and secondarily to maximize the improvement of P_2 in performance over P_1. In terms of the language of the quantities defined in equations (1) and (2), that player wishes primarily to minimize $D\left(\dot{P}\|P_2\right)$ (maximize $-D\left(\dot{P}\|P_2\right)$) and secondarily to maximize $D\left(\dot{P}\|P_1 \rightsquigarrow P_2\right)$.

More precisely, the amount of utility paid to the statistician making move P_2 is

$$U\left(\dot{P}, P_1, P_2\right) = \left(-D\left(\dot{P}\|P_2\right), D\left(\dot{P}\|P_1 \rightsquigarrow P_2\right)\right), \tag{6}$$

understood in terms of preferring $v = (v_1, v_2) \in [0, \infty[\times [0, \infty[$ over $u = (u_1, u_2) \in [0, \infty[\times [0, \infty[$ if and only if $u \preceq v$. Here, $u \preceq v$ means that either $u_1 < v_1$ or both $u_1 = v_1$ and $u_2 \leq v_2$. Such preferences are said to have *lexicographic ordering* (Remark 3).

Thus, the utility paid to Combiner will be $U\left(\dot{P}, \ddot{P}, P^+\right)$ and that paid to Chooser will be $U\left(\dot{P}, P^+, \ddot{P}\right)$. The utility paid to Nature will also be $U\left(\dot{P}, P^+, \ddot{P}\right)$, with the implication that it is to the advantage of Nature and Chooser to act as a *coalition* with move $\left(\dot{P}, \ddot{P}\right)$ [56, Ch. 5]. Although that reduces the three-player game to a two-player game of the coalition versus Combiner, the game is not necessarily of zero sum.

The *combination of the distributions in* $\ddot{\mathcal{P}}$ *with truth constrained by* $\dot{\mathcal{P}}$ is defined as Combiner's optimal move in the game. Combiner's best move may be written as

$$P^+ = \arg \overset{\preceq}{\sup_{Q \in \mathcal{P}}} U\left(\dot{P}_Q, \ddot{P}_Q, Q\right), \tag{7}$$

where $\sup^{\preceq}$ is the least upper bound according to $\preceq$, and $\dot{P}_Q$ and $\ddot{P}_Q$ would be the best moves of Nature and Chooser if Q were the move of Combiner. Thus, $\left(\dot{P}_Q, \ddot{P}_Q\right)$ is the optimal pair of distributions that constitutes the optimal move for the Nature-Chooser coalition given move $Q \in \mathcal{P}$ by Combiner:

$$\left(\dot{P}_Q, \ddot{P}_Q\right) = \arg \overset{\preceq}{\sup_{(P', P'') \in \dot{\mathcal{P}} \times \ddot{\mathcal{P}}}} U\left(P', Q, P''\right). \tag{8}$$

While P^+ is not necessarily a plausible distribution, it is typically at the center of the plausible set:

Theorem 1. *Let P^+ denote the combination of the distributions in $\ddot{\mathcal{P}}$ with truth constrained by $\dot{\mathcal{P}}$. If $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}} \neq \emptyset$, then*

$$P^+ = \operatorname{cent} \dot{\mathcal{P}} \cap \ddot{\mathcal{P}} = P^{W_{\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}}}, \tag{9}$$

where $\operatorname{cent} \dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$ *is the centroid of $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$, and $W_{\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}}$ is the weighting distribution induced by $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$, as defined by Eq. (4).*

Let $\mathcal{A}$ denote an action space. A decision made by taking the action $a \in \mathcal{A}$ that minimizes the expectation value of a loss function $L : \Xi \times \mathcal{A} \to \mathbb{R}$ with respect to P^+ is optimal in the game when the utility function of Eq. (6) is replaced with

$$\left(-D\left(\dot{P} \| P_2\right), D\left(\dot{P} \| P_1 \rightsquigarrow P_2\right), -\int L(\xi, a)\, dP_2(\xi)\right).$$

The latter utility function is understood in terms of the lexicographic ordering relation $\preccurlyeq$, which is defined such that $(u_1, u_2, u_3) \preccurlyeq (v_1, v_2, v_3)$ if and only if one of the following is true: (i) $u_1 < v_1$; (ii) $u_1 = v_1$ and $u_2 < v_2$; (iii) $u_1 = v_1$, $u_2 = v_2$, and $u_3 \leq v_3$. On related orderings in the literature, see Remark 3.

2.3. Combining discrete distributions

Now let $\mathcal{P}$ denote the set of probability distributions on $\left(\Xi, 2^{\Xi}\right)$, where Ξ is a finite set. It is written as $\Xi = \{0, 1, \ldots, |\Xi| - 1\}$ without loss of generality. Then the information divergence of P with respect to Q (1) reduces to

$$D(P \| Q) = \sum_{i \in \Xi} P(\{i\}) \log \frac{P(\{i\})}{Q(\{i\})}.$$

For any $P \in \mathcal{P}$ and the random variable ξ of distribution P, the $|\Xi|$-tuple

$$T(P) = (P(\xi = 0), P(\xi = 1), \ldots, P(\xi = |\Xi| - 1))$$

will be called the *tuple representing P*.

Consider $\mathcal{P}^\star = \{P_\phi : \phi \in \Phi\}$, a nonempty subset of $\mathcal{P}$. Every $\phi \in \Phi$ corresponds to a different random variable and thus to a different $|\Xi|$-tuple. Let $\mathcal{T}\left(\mathcal{P}^\star\right)$ denote the set of tuples representing the members of $\mathcal{P}^\star$, i.e., $\mathcal{T}\left(\mathcal{P}^\star\right) = \left\{T\left(P^\star\right) : P^\star \in \mathcal{P}^\star\right\}$. Likewise, noting that the map $\mathcal{T}$ is invertible (bijective), the *extreme subset of* $\mathcal{P}^\star$ is defined as $\operatorname{ext} \mathcal{P}^\star = \mathcal{T}^{-1}\left(\operatorname{ext} \operatorname{co} \mathcal{T}\left(\mathcal{P}^\star\right)\right)$, where $\operatorname{ext} \operatorname{co} \mathcal{T}\left(\mathcal{P}^\star\right)$ is the set of extreme points of $\operatorname{co} \mathcal{T}\left(\mathcal{P}^\star\right)$, the convex hull of $\mathcal{T}\left(\mathcal{P}^\star\right)$. The extreme subset simplifies the problem of locating a centroid:

Lemma 2. *Let $\mathcal{P}^\star$* denote a nonempty, finite subset of $\mathcal{P}$. *If there are a $Q \in \mathcal{P}$ and a $C > 0$ such that $D\left(P^\star \| Q\right) = C$ for all $P^\star \in \operatorname{ext} \mathcal{P}^\star$, then Q is the centroid of $\mathcal{P}^\star$.*

More simplification is possible if at least one of the combining distributions is plausible:

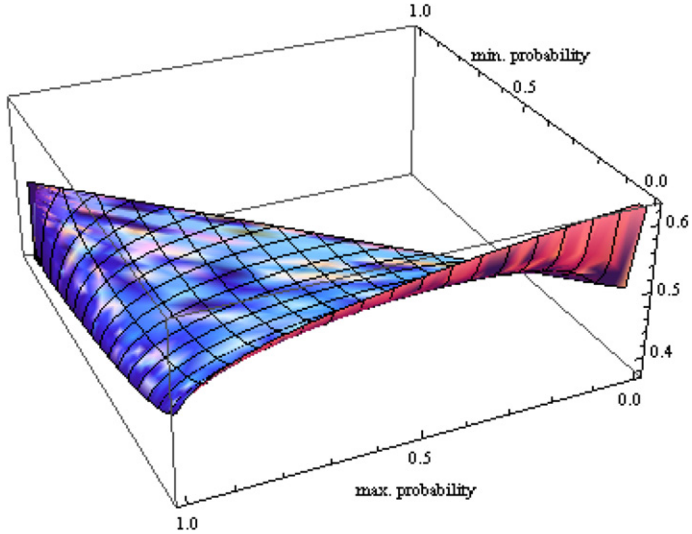

Fig. 1. Optimal weight w^+ versus $\underline{\ddot{P}}(\{0\})$ and $\overline{\ddot{P}}(\{0\})$, the lowest and highest of the plausible probabilities to be combined (10), labeled here as "min. probability" and "max. probability," respectively. The combination probability is $P^+(0) = w^+\underline{\ddot{P}}(0) + (1-w^+)\overline{\ddot{P}}(0)$ according to Corollary 1.

Theorem 2. *Let P^+ denote the combination of the distributions in $\ddot{\mathcal{P}}$ with truth constrained by $\dot{\mathcal{P}}$. If $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$ is nonempty and finite, then $P^+ = P^{W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}}$, where $W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}$ is the weighting distribution induced (4) by $\mathrm{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$, the extreme subset of $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$.*

The combination of a set of probabilities of the same hypothesis (Section 3) or event is simply the linear combination or mixture of the highest and lowest of the plausible probabilities in the set such that the mixing proportion is optimal:

Corollary 1. *Let P^+ denote the combination of the distributions in $\ddot{\mathcal{P}}$ with truth constrained by $\dot{\mathcal{P}}$. Suppose c distributions on $\left(\{0,1\}, 2^{\{0,1\}}\right)$ are to be combined $\left(\ddot{\mathcal{P}} = \left\{\ddot{P}_1, \dots, \ddot{P}_c\right\}\right)$, and let $\dot{\mathfrak{P}}_0 = \left\{\dot{P}(\{0\}) : \dot{P} \in \dot{\mathcal{P}}\right\}$ and $\underline{\ddot{P}}, \overline{\ddot{P}} \in \mathcal{P}$ such that*

$$\underline{\ddot{P}}(\{0\}) = \min_{i=1,\dots,c:\ddot{P}_i(\{0\})\in\dot{\mathfrak{P}}_0} \ddot{P}_i(\{0\})\,;\ \overline{\ddot{P}}(\{0\}) = \max_{i=1,\dots,c:\ddot{P}_i(\{0\})\in\dot{\mathfrak{P}}_0} \ddot{P}_i(\{0\})\,.$$

If there is at least one $i \in \{1, \dots, c\}$ for which $\ddot{P}_i(\{0\}) \in \dot{\mathfrak{P}}_0$ holds, then $P^+ = w^+\underline{\ddot{P}} + (1-w^+)\overline{\ddot{P}}$, where

$$w^+ = \arg\sup_{w\in[0,1]} \left(w\Delta\left(\underline{\ddot{P}}\|w\right) + (1-w)\,\Delta\left(\overline{\ddot{P}}\|w\right)\right); \tag{10}$$

$$\Delta(\bullet\|w) = D\left(\bullet\|w\underline{\ddot{P}} + (1-w)\,\overline{\ddot{P}}\right).$$

Proof. This follows immediately from Theorem 2 and the definition of an extreme subset. □

By Eq. (5), $37\% \mathrel{\dot{\lesssim}} w^+ \mathrel{\dot{\lesssim}} 63\%$, implying that $P^+(\{0\})$ is close to the arithmetic mean $\left[\underline{\ddot{P}}(\{0\}) + \overline{\ddot{P}}(\{0\})\right]\big/2$, as Shulman and Feder [51] observed in a coding context. Fig. 1 plots w^+ versus $\underline{\ddot{P}}(\{0\})$ and $\overline{\ddot{P}}(\{0\})$, and Fig. 2 compares the resulting $P^+(\{0\})$ to the arithmetic mean, the geometric mean, and the harmonic mean of $\underline{\ddot{P}}(\{0\})$ and $\overline{\ddot{P}}(\{0\})$.

The next result is important for multiple hypothesis testing (Sections 3 and 4) and, more generally, for combining probabilities of independent events rather than entire distributions.

Corollary 2. *Let $\xi = (\xi_1, \dots, \xi_N)$, where ξ_j is a Bernoulli random variable and ξ_j is independent of ξ_J for all $j, J = 1, \dots N$. (The Bernoulli distributions need not be identical: in general, each has a different probability $\ddot{P}(\xi_i = 0)$ of failure. Every $\ddot{P} \in \ddot{\mathcal{P}}$ has a one-to-one correspondence to a tuple $\left(\ddot{P}(\xi_1 = 0), \dots, \ddot{P}(\xi_N = 0)\right)$.) Assuming $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$ is nonempty and finite, let $\ddot{P}_i$ denote the ith of the ν members of $\mathrm{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$; thus,*

$$\mathrm{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}) = \left\{\ddot{P}_1, \dots, \ddot{P}_\nu\right\}.$$

If the constraints are in the form of lower and upper probabilities $\underline{P}_{0,1}, \dots, \underline{P}_{0,N}$ and $\overline{P}_{0,1}, \dots, \overline{P}_{0,N}$ such that

$$\dot{\mathcal{P}} = \left\{\dot{P} \in \mathcal{P} : \underline{P}_{0j} \leq \dot{P}(\xi_j = 0) \leq \overline{P}_{0j},\quad j \in \{1, \dots, N\}\right\}, \tag{11}$$

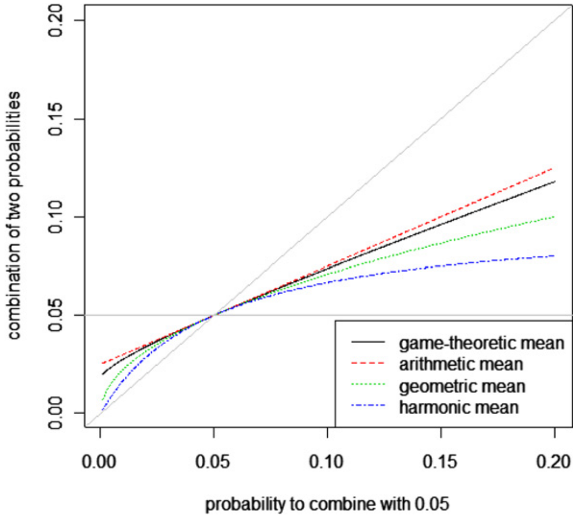

Fig. 2. Three equal-weight averages of probabilities and the game-theoretic combination of two probabilities based on Corollary 1. The two probabilities that are combined are 0.05 and each value of the x-axis; the former is the horizontal line at 0.05, and the latter is the line of equality.

then the set of combining distributions that satisfy the constraints is

$$\dot{\mathcal{P}} \cap \ddot{\mathcal{P}} = \left\{ \ddot{P} \in \ddot{\mathcal{P}} : \underline{P}_{0j} \leq \ddot{P}\left(\xi_j = 0\right) \leq \overline{P}_{0j}, \quad j \in \{1, \ldots, N\} \right\}. \tag{12}$$

Further, $P^{+} = P^{\boldsymbol{w}_{\text{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})}}$ *is the combination of the distributions in* $\ddot{\mathcal{P}}$ *with truth constrained by* $\dot{\mathcal{P}}$*, where*

$$\boldsymbol{w}_{\text{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})} = \arg \sup_{(w_1, \ldots, w_\nu) \in \mathfrak{W}} \sum_{i=1}^{\nu} w_i \sum_{j=1}^{N} \sum_{k=0}^{1} \ddot{P}_i\left(\xi_j = k\right) \log \frac{\ddot{P}_i\left(\xi_j = k\right)}{P^{(w_1, \ldots, w_\nu)}\left(\xi_j = k\right)} \tag{13}$$

with the supremum over $\mathfrak{W} = \left\{ \left(w_1', \ldots, w_\nu'\right) \in [0, 1]^\nu : \sum_{i=1}^{\nu} w_i' = 1 \right\}$.

Proof. Eq. (12) is obvious from Eq. (11). By the independence condition, the chain rule for information divergence [see, e.g., 17, Theorem 2.5.3] reduces finding the weighting distribution according to Theorem 2 to finding

$$\boldsymbol{w}_{\text{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})} = \arg \sup_{(w_1, \ldots, w_\nu) \in \mathfrak{W}} \sum_{i=1}^{\nu} w_i \sum_{j=1}^{N} D\left(\ddot{P}_i\left(\xi_j = \bullet\right) \| P^{(w_1, \ldots, w_\nu)}\left(\xi_j = \bullet\right) \right),$$

which, with Eq. (1), yields Eq. (13). □

3. Distribution combination for statistical inference

Whereas much of the literature focuses on combining priors from experts, Section 3.1 instead focuses on combining posteriors. The posterior-inference setting enables the combination not only of Bayesian posterior distributions from improper priors as well as proper priors but also the combination of confidence intervals (or other regions) and p-values encoded as frequentist posterior distributions, as will be explained in Section 3.2.

3.1. Combining posterior distributions and probabilities

In the context of posterior statistical inference, ξ represents a random parameter. Further, all distributions in $\dot{\mathcal{P}}$, the set of plausible distributions of ξ, and $\ddot{\mathcal{P}}$, the set of combining distributions of ξ, are posterior with respect to the same data set x. All distributions in $\dot{\mathcal{P}}$ and all Bayesian posteriors in $\ddot{\mathcal{P}}$ are conditional on $X = x$, where, for any such posterior, the distribution of X depends on the random value of the parameter drawn from some prior. $\ddot{\mathcal{P}}$ may also contain non-Bayesian posteriors such as a confidence posterior or a distribution derived from a confidence posterior (Section 3.2).

Accordingly, the information divergence $D\left(P \| Q\right)$ becomes the amount of information that would be gained by replacing any posterior Q with the true posterior P. That interpretation leads to viewing $D\left(P' \| P'' \rightsquigarrow Q\right)$ as the amount of information gained for statistical inference by using some posterior $Q \in \mathcal{P}$ rather than a given posterior $P'' \in \ddot{\mathcal{P}}$ when the plausible posterior is $P' \in \dot{\mathcal{P}}$. Thus, $D\left(P' \| P'' \rightsquigarrow Q\right)$ defines the *inferential gain of* Q *relative to* P'' *given* P' [7,12].

The posterior distributions are combined according to Section 2.2, using Theorem 1 whenever possible. If $\dot{\mathcal{P}}$ represents the uncertainty around a Bayesian posterior $\dot{P} \in \dot{\mathcal{P}}$, as in Gajdos et al. [25] and Bickel [12], then $\dot{P}$ is included in $\ddot{\mathcal{P}}$ as one of the distributions to combine. The resulting combination P^{+} is then used to minimize expected loss in order to optimize actions such as point, interval, and function estimators and predictors.

In model selection and hypothesis testing, ξ has 0 or 1 as its realized value, with $\xi = 0$ if a reduced model or null hypothesis is true or $\xi = 1$ if a full model or alternative hypothesis is true. Corollary 1 applies to this problem with $\dot{\mathfrak{P}}_0$ as the set of feasible null hypothesis posterior probabilities and $\ddot{\mathfrak{P}}_0 = \left\{\ddot{P}_1(\{0\}), \ldots, \ddot{P}_c(\{0\})\right\}$ as the set of null hypothesis posterior probabilities to be combined, where $\ddot{P}_i(\{0\}) = \ddot{P}_i(\xi = 0)$ is the ith posterior probability that the null hypothesis is true. Thus, the combination posterior probability that the null hypothesis is true is

$$P^{+}(\xi = 0) = w^{+}\underline{\ddot{P}}(\xi = 0) + \left(1 - w^{+}\right)\overline{\ddot{P}}(\xi = 0). \tag{14}$$

Here, $\underline{\ddot{P}}(\xi = 0) = \underline{\ddot{P}}(0)$ and $\overline{\ddot{P}}(\xi = 0) = \overline{\ddot{P}}(0)$ are respectively the lowest and highest null hypothesis posterior probabilities that are in $\dot{\mathfrak{P}}_0 \cap \ddot{\mathfrak{P}}_0$, presently assumed to have at least one member, and w^{+} is determined by Eq. (10). The same idea applies to multiple hypothesis testing, as will be seen in Section 4.

3.2. Combining frequentist posteriors

3.2.1. Confidence posteriors

As mentioned in the beginning of Section 3, the set $\ddot{\mathcal{P}}$ of posterior combining distributions can include those representing confidence intervals, more general confidence sets, and p-values. To emphasize their comparability to Bayesian posterior distributions, these frequentist distributions are called "confidence posterior distributions" [11], also known as "confidence distributions" [see, e.g., 48].

Briefly, a *confidence posterior distribution* that corresponds to a system of nested confidence intervals or other confidence sets evaluated for the observed data is defined as the probability distribution according to which the posterior probability that the interest parameter lies within a confidence set is equal to the confidence level of the set. For example, if a 95% confidence interval for a real parameter is $[-2.2, 1.7]$, then there is a 95% posterior probability that the parameter is between -2.2 and 1.7 according to the confidence posterior. The same confidence posterior for the data also assigns posterior probability to parameter intervals of interest according to the confidence levels of the matching confidence intervals, e.g., Bickel [8] considered a one-sided p-value as the posterior probability that the population mean is in $]-\infty, 0[$ rather than $[0, \infty[$. Efron and Tibshirani [23] and Polansky [43] considered exact confidence posterior probabilities of intervals or other regions specified before observing the data as ideal cases of "attained confidence levels" and "observed confidence levels," respectively.

Bickel [8,11] proposed taking actions that minimize expected loss with respect to a confidence posterior distribution. Since that distribution is a Kolmogorov probability distribution of the parameter of interest, such actions comply with most axiomatic systems usually considered Bayesian, e.g., the systems of von Neumann and Morgenstern [56] and Savage [47]. A human or artificial intelligent agent that bets and makes other decisions in accordance with minimizing expected loss with respect to a confidence posterior corresponds to equating the confidence level of a confidence interval with the agent's level of belief that the parameter value lies in the interval [11,6].

The decision-theoretic framework makes confidence posteriors suitable as members of $\ddot{\mathcal{P}}$, the set of combining distributions, according to the methodology of Section 2. They can be combined to not only with each other, but also with other parameter distributions such as Bayesian posteriors based on proper or improper priors. The same applies to a probability distribution of a function of a parameter drawn from a confidence posterior. Such posteriors have been used to equate posterior probabilities of simple null hypotheses with two-sided p-values [55,10,7,12]. For terminological economy, these posteriors will now be called "confidence posteriors."

To approximate Bayesian model averaging, Good [30] recommended a weighted harmonic mean of p-values computed from the same data, provided that they range from 10^{-3} to 0.2, the limits used in Fig. 2. Since the one-sided or two-sided p-values are posterior probabilities of the null hypothesis derived from different confidence posteriors, Eqs. (10) and (14) can be applied with $\underline{\ddot{P}}(\xi = 0)$ and $\overline{\ddot{P}}(\xi = 0)$ as the lowest and highest p-values that are plausible as null hypothesis probabilities, i.e., that are in $\dot{\mathfrak{P}}_0$. The resulting combination p-value differs from that of Good [30] in two respects: the mean is arithmetic (14) rather than harmonic and, even more important, the weights are optimal for the game (10) rather than subjective. The use of optimal weights leads to preparing for the worst case by averaging only the two most extreme p-values rather than all of them. It also applies much more generally since strictly Bayesian model averaging tends to fail in the presence of improper priors [9,13, and references].

Example 2. Given a small sample of data drawn from a distribution that might be approximately normal, let $p^{(1)}$ and $p^{(2)}$ denote the two-sided p-values according to the t-test and the Wilcoxon signed-rank test, respectively; $p^{(1)} < p^{(2)}$. Under conditions often applicable to simple (point) hypothesis testing with a diffuse alternative hypothesis [50], the plausible set of posterior probabilities of the null hypothesis is the closed interval $\dot{\mathfrak{P}}_0 = \left[\underline{\dot{P}}_0, 1\right]$ with lower bound

$$\underline{\dot{P}}_0 = \left(1 + \left(\frac{1 - \underline{\dot{P}}_0^{\text{prior}}}{\underline{\dot{P}}_0^{\text{prior}} e p^{(2)}(x) \ln\left[1/p^{(2)}(x)\right]}\right)\right)^{-1} \wedge \underline{\dot{P}}_0^{\text{prior}},$$

where $e = \ln^{-1}(1) \doteq 2.72$, $\wedge$ is the minimum operator, and $\underline{\dot{P}}_0^{\text{prior}}$ is the lowest plausible prior probability that the null hypothesis is true [12]. Then the combined p-value $P^+(\{0\})$ is $\underline{\dot{P}}_0$ if $p^{(2)} < \underline{\dot{P}}_0$, $p^{(2)}$ if $p^{(1)} < \underline{\dot{P}}_0 \leq p^{(2)}$, and, according to Corollary 1, the weighted arithmetic mean $w^+p^{(1)} + (1 - w^+)\, p^{(2)}$ if $p^{(1)} \geq \underline{\dot{P}}_0$ with the weights w^+ and $(1 - w^+)$ fixed by Eq. (10). Of the three cases, the third yields a combined p-value that differs from the blended posterior probability suggested in Bickel [7].

When $\ddot{\mathcal{P}}$ consists of a single confidence posterior $\ddot{P}$, the resulting P^+, degenerate as a "combination" of a single distribution, is better viewed as a solution to the problem of blending frequentist inference with constraints encoded as the Bayesian posteriors that constitute $\dot{\mathcal{P}}$. That solution in general differs from the minimax-type solutions considered [7,12]. Under $\ddot{P} \in \dot{\mathcal{P}}$ and the convexity of $\dot{\mathcal{P}}$, they lead to the P that minimizes the information divergence $D\left(P\|\ddot{P}\right)$, which is dual to the Q that minimizes $D\left(\ddot{P}\|Q\right)$, the information divergence that is minimized (Appendix A, Eq. (15)) when maximizing Eq. (6) according to the game introduced in Section 2.2.

3.2.2. Multiple comparison procedures

The distribution–combination theory is now applied to adjustments for multiple comparisons by formalizing the observation that p-values are often adjusted to the extent of prior belief in the null hypothesis. That is, multiple comparison procedures (MCPs) designed to control error rates are "most likely to be used, if at all, when most of the individual null hypotheses are essentially correct" [18, p. 88]. A first-order formalization would take the p-value that is adjusted according to an MCP as the posterior probability of the null hypothesis. To the extent that the knowledge or opinion of the agent is such that its decisions would be made to minimize the expected loss with respect to that posterior distribution, the use of the MCP is warranted. In this interpretation, combining p-values across different MCPs is equivalent to combining the posterior distributions that represent the corresponding opinions.

Example 3. The Bonferroni procedure controls the family-wise error rate, the probability that one or more true null hypotheses will be rejected, at any level $\alpha \in [0, 1]$. That is accomplished on the basis of p-values $p_1, \ldots, p_N$ by rejecting the ith of N null hypotheses if the adjusted p-value $Np_i \wedge 1$ is less than α. Thus, the posterior probabilities generated by the Bonferroni procedure are appropriate only when the prior probability of each null hypothesis is inversely proportional to the number of tests. As Westfall et al. [59] pointed out, the "Bonferroni method is based upon the implicit presumption of a moderate degree of belief in the event" that all null hypotheses under consideration are true and that the prior truth values of the hypotheses are approximately independent.

Accordingly, the Bonferroni method is widely used to analyze genome-wide association data, largely because only an extremely small fraction π_1 of the hundreds of thousands of markers tested is thought to be associated with the trait of interest. [58] guessed $10^{-6} \leq \pi_1 \leq 10^{-4}$, interpreting π_1 as the prior probability of association between a given marker and the trait. The corresponding range of posterior probabilities and Bayes factors such as those of [58] would define $\dot{\mathcal{P}}$ for ruling out MCPs that yield implausible results (Theorem 2). (On the other hand, some evidence that $\pi_1 \geq 10^{-4}$ is now available in preliminary estimates [60] and in indications that thousands of small-effect single-nucleotide polymorphisms (SNPs) may be associated with any particular disease [28,41].)

By assuming adjusted p-values are equal to independent posterior probabilities of the null hypotheses, the methodology of Corollary 2 can combine the results of various MCPs.

4. Large-scale case study

Using microarray technology, Alba et al. [2] measured the levels of tomato gene expression for 13,440 genes at three days after the breaker stage of ripening, but one or more measurements were missing for 7337 genes. The data available across all $n = 6$ biological replicates for $N = 6103$ of the genes illustrate the methodology of Sections 2 and 3.

For $j = 1, \ldots, N$, the logarithms of the measured ratios of mutant expression to wild-type expression in the jth gene were modeled as realizations of a normal variate and are denoted by the n-tuple x_j. Because the mean and variance are unknown, the one-sample t-test was used to test the null hypothesis $(\xi_j = 0)$ that the population mean is 0 against the two-sided alternative hypothesis $(\xi_j = 1)$ that there is differential expression of the jth gene between the mutant and the wild type, i.e., that the mutation affects the expression of gene j.

The posterior probability of a null hypothesis conditional on the p-value is called its *local false discovery rate* (LFDR) [24]. Three very different methods $(i = 1, 2, 3)$ of estimating the LFDR were considered. The first two methods are based on fitting a histogram of transformed p-values that is described by Efron [21]. They differ in that whereas the first method assumes the p-value has a uniform distribution under the null hypothesis $(i = 1)$, the second method estimates the p-value

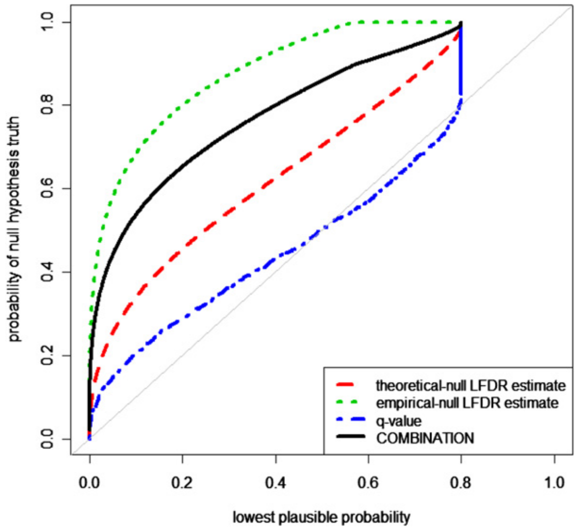

Fig. 3. Combination of estimates of local false discovery rates, which are empirical Bayes posterior probabilities that the null hypotheses of equivalent gene expression are true.

null distribution by maximizing a truncated likelihood function ($i = 2$). Each method has its own advantages (Section 1). The distributions are called the *theoretical null* and the *empirical null*, respectively. The third method for combination is the q-value [52], here defined according to the algorithm of Benjamini and Hochberg [4] as the lowest false discovery rate at which a null hypothesis will be rejected ($i = 3$). While the q-value was not originally intended as an estimator of the LFDR, it is included here since its negative bias as such an estimator [33] may have a corrective effect on the positive bias (conservatism) of the first two LFDR estimators.

For this application, $\mathcal{P}$ is the set of all probability distributions on $\left(\{0, 1\}^N, 2^{\{0,1\}^N}\right)$. Corresponding to those three methods, let $\ddot{P}_1, \ddot{P}_2$, and $\ddot{P}_3$ denote the members of $\mathcal{P}$ such that the ith estimate of the LFDR of the jth gene is $\ddot{P}_i(\xi_j = 0)$. To combine the three methods, $\ddot{\mathcal{P}} = \left\{\ddot{P}_1, \ddot{P}_2, \ddot{P}_3\right\}$ is taken as the set of $\nu = 3$ combining distributions.

For the jth gene, $f(t(x_j); \theta_j)$ will represent the probability density of the Student t statistic $t(x_j)$, where θ_j is the reciprocal of the coefficient of variation and, for any $\theta \in \mathbb{R}$, $f(\bullet; \theta)$ is the probability density function of $|T|$ when T has the noncentral t distribution of $n - 1$ degrees of freedom and noncentrality parameter $\sqrt{n}\theta$. The set $\dot{\mathcal{P}}$ of plausible distributions will be determined on the basis of $\{L_j(\bullet) = f(t(x_j); \bullet) : j = 1, \ldots, N\}$, the set of likelihood functions. The plausible distributions are also based on $\underline{\pi}_0 = 80\%$, an assumed lower bound on the proportion of genes that are not differentially expressed. By Bayes's theorem, the posterior odds of the jth null hypothesis is the product of the prior odds, which is the least $\underline{\pi}_0 / (1 - \underline{\pi}_0)$, and the Bayes factor, which must be at least $L_j(0) / \max_{\theta \neq 0} L_j(\theta)$. Thus, for gene j, a lower bound $\underline{\Omega}_j$ of the posterior odds is the product of the last two quantities, and a lower bound of the LFDR is $\underline{P}(\xi_j = 0) = \underline{\Omega}_j / \left(1 + \underline{\Omega}_j\right)$. In the notation of Corollary 2, $\underline{P}_{0,j} = \underline{P}(\xi_j = 0)$ and, trivially, $\overline{P}_{0,j} = 1$ for all $j = 1, \ldots, N$. Thus, the plausible set specified by Eq. (11) consists of the posterior distributions satisfying the lower bound derived from the likelihood functions and $\underline{\pi}_0 = 80\%$.

The horizontal axis and straight line in Fig. 3 represent $\underline{P}$, and the intermediate, highest, and lowest dashed curves represent $\ddot{P}_1, \ddot{P}_2$, and $\ddot{P}_3$, respectively. Since some of the q-values are less than the lower bound $\left(\exists j : \ddot{P}_3(\xi_j = 0) < \underline{P}(\xi_j = 0)\right)$ but all of the other LFDR estimates satisfy the bound $\left(i = 1, 2; \forall j : \ddot{P}_i(\xi_j = 0) \geq \underline{P}(\xi_j = 0)\right)$, the former are excluded when computing the combined estimates according to Eq. (12), in which $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}} = \left\{\ddot{P}_1, \ddot{P}_2\right\}$. Since there are only two distributions, each corresponds to an extreme point, leading to $\text{ext}(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}) = \left\{\ddot{P}_1, \ddot{P}_2\right\}$ in Eq. (13). The combined distribution is numerically found to be the linear combination $P^+ = w_1\ddot{P}_1 + w_2\ddot{P}_2$ with $w_1 = 0.43$ and $w_2 = 0.57$. By implication, the game-optimal LFDR estimate for the jth gene is $P^+(\xi_j = 0) = w_1\ddot{P}_1(\xi_j = 0) + w_2\ddot{P}_2(\xi_j = 0)$. Those combined estimates are plotted as the solid curve in Fig. 3.

5. Remarks

Remark 1 (Section 1)**.** Since formulating the distribution combination problem in terms of a game is unconventional, it is worth noting that game theory laid the foundations of the two dominant schools of statistical decision theory. The maximum-expected-payoff solution of a one-player game [56, Ch. I, Section 3.6–3.7] led to axiomatic systems that support Bayesian statistics [e.g., 47]. Likewise, the worst-case (minimax) solutions of certain two-player zero-sum games [56, Ch. III] led to frequentist decision theory [57].

Remark 2 (Section 2.1). The discrete-distribution version of Lemma 1, the main result of the "redundancy-capacity theorem," was presented by R. G. Gallager in 1974 [46, Editor's Note] and published by Ryabko [45] and Davisson and Leon-Garcia [20]; cf. Gallager [26]. Cover and Thomas [17, Theorem 13.1.1], Rissanen [44, Section 5.2.1], and Csiszár and Körner [19, Problem 8.1] provide useful introductions.

Remark 3 (Section 2.2). Previous instances of lexicographically maximizing expected utility with respect to an optimal probability distribution include the use of the least informative prior [49] and the use of the posterior P_2 used to maximize $D\left(\dot{P}\|P_1 \rightsquigarrow P_2\right)$ in a two-player zero-sum game [12]. On lexicographic decision making in other contexts, see Levi [37, Sections 5.7 and 6.9], Levi [38], and Keeney and Raiffa [34, Section 3.3.1]. Ciesielski [14, Ch. 4] and Koshy [35, Ch. 7] provide more formal set-theoretic expositions of lexicographic ordering.

Acknowledgments

Two anonymous reviewers made several suggestions that substantially improved the notation and clarity of the paper. I also 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.

Appendix A. Additional proofs

Proof of Theorem 1

For any $Q \in \mathcal{P}$, Eqs. (2) and (6) yield

$$\begin{aligned}\overset{\preceq}{\sup_{(P',P'')\in\dot{\mathcal{P}}\times\ddot{\mathcal{P}}}} U\left(P', Q, P''\right) &= \overset{\preceq}{\sup_{(P',P'')\in\dot{\mathcal{P}}\times\ddot{\mathcal{P}}}} \left(-D\left(P'\|P''\right), D\left(P'\|Q \rightsquigarrow P''\right)\right)\\ &= \overset{\preceq}{\sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}}} \left(-D\left(P'\|P'\right), D\left(P'\|Q \rightsquigarrow P'\right)\right)\\ &= \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} \left[D\left(P'\|Q\right) - D\left(P'\|P'\right)\right] = \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} D\left(P'\|Q\right).\end{aligned}$$

Thus, by Eqs. (2), (6), (7), and (8),

$$\begin{aligned}P^{+} &= \arg \overset{\preceq}{\sup_{Q\in\mathcal{P}}} U\left(\arg \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} D\left(P'\|Q\right), \arg \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} D\left(P'\|Q\right), Q\right)\\ &= \arg \sup_{Q\in\mathcal{P}} \left(-D\left(\arg \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} D\left(P'\|Q\right) \|Q\right)\right)\\ &= \arg \inf_{Q\in\mathcal{P}} \sup_{P'\in\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}} D\left(P'\|Q\right). \qquad (15)\end{aligned}$$

Hence, according to Eq. (3), P^{+} is cent $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$, the centroid of $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$. Since $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}} \neq \emptyset$ by assumption, the conditions of Lemma 1 are satisfied.

Proof of Lemma 2

As an immediate consequence of what [40] label "Theorem (Csiszár)" and "Theorem 1,"

$$\min_{P''\in\mathcal{P}} \max_{P'\in\mathcal{P}^{\star}} D\left(P'\|P''\right) = C.$$

By definition, the centroid is the solution of that minimax problem (3).

Proof of Theorem 2

Lemma 1 implies that cent ext $(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$, the centroid of ext $(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$, is $P^{W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}}$. Since, according to the definition of an extreme point and the definition of a centroid (3), it is not possible that there exist a $P' \in$ ext $(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$ and a $P'' \in$ ext $(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$

such that $D\left(P' \,\|\, \mathrm{cent\,ext}\,(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})\right) < D\left(P'' \,\|\, \mathrm{cent\,ext}\,(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})\right)$, it follows that

$$D\left(P^{\star} \,\|\, P^{W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}}\right) = \max_{P' \in \mathrm{ext}\left(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}}\right)} D\left(P' \,\|\, P^{W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}}\right)$$

for all $P^{\star} \in \mathrm{ext}\,(\dot{\mathcal{P}} \cap \ddot{\mathcal{P}})$. According to Lemma 2, $P^{W_{\mathrm{ext}(\dot{\mathcal{P}}\cap\ddot{\mathcal{P}})}}$ is the centroid of $\dot{\mathcal{P}} \cap \ddot{\mathcal{P}}$. That centroid is P^{+} by Theorem 1.

References

[1] A.E. Abbas, A Kullback–Leibler view of linear and log-linear pools, Decision Analysis 6 (2009) 25–37.
[2] R. Alba, P. Payton, Z. Fei, R. McQuinn, P. Debbie, G.B. Martin, S.D. Tanksley, J.J. Giovannoni, Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development, Plant Cell 17 (2005) 2954–2965.
[3] O.E. Barndorff-Nielsen, Diversity of evidence and Birnbaum's theorem, Scandinavian Journal of Statistics 22 (1995) 513–515.
[4] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society B 57 (1995) 289–300.
[5] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, New York, 1985.
[6] D.R. Bickel, A frequentist framework of inductive reasoning, Technical Report, Ottawa Institute of Systems Biology, 2009, arXiv:math.ST/0602377.
[7] D.R. Bickel, Blending Bayesian and frequentist methods according to the precision of prior information with an application to hypothesis testing. Technical Report, Ottawa Institute of Systems Biology, 2011, arXiv:1107.2353.
[8] D.R. Bickel, Estimating the null distribution to adjust observed confidence levels for genome-scale screening, Biometrics 67 (2011) 363–370.
[9] D.R. Bickel, A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons, Canadian Journal of Statistics 39 (2011) 610–631.
[10] D.R. Bickel, Small-scale inference: Empirical Bayes and confidence methods for as few as a single comparison, Technical Report, Ottawa Institute of Systems Biology, 2011, arXiv:1104.0341.
[11] D.R. Bickel, Coherent frequentism: A decision theory based on confidence sets, Communications in Statistics – Theory and Methods 41 (2012) 1478–1496.
[12] D.R. Bickel, Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes, Electronic Journal of Statistics 6 (2012) 686–709.
[13] D.R. Bickel, Empirical Bayes interval estimates that are conditionally equal to unadjusted confidence intervals or to default prior credibility intervals, Statistical Applications in Genetics and Molecular Biology 11 (3) (2012), art. 7.
[14] K. Ciesielski, Set Theory for the Working Mathematician, Cambridge University Press, Cambridge, 1997.
[15] R.T. Clemen, R.L. Winkler, Combining probability distributions from experts in risk analysis, Risk Analysis 19 (1999) 187–203.
[16] R.M. Cooke, Experts in Uncertainty: Opinion and Subjective Probability in Science, Oxford University Press, 1991.
[17] T. Cover, J. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 2006.
[18] D.R. Cox, Principles of Statistical Inference, Cambridge University Press, Cambridge, 2006.
[19] I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Cambridge University Press, Cambridge, 2011.
[20] L. Davisson, A. Leon-Garcia, A source matching approach to finding minimax codes, IEEE Transactions on Information Theory 26 (1980) 166–174.
[21] B. Efron, Size, power and false discovery rates, Annals of Statistics 35 (2007) 1351–1377.
[22] B. Efron, Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge University Press, Cambridge, 2010.
[23] B. Efron, R. Tibshirani, The problem of regions, Annals of Statistics 26 (1998) 1687–1718.
[24] B. Efron, R. Tibshirani, J.D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment, Journal of the American Statistical Association 96 (2001) 1151–1160.
[25] T. Gajdos, J.M. Tallon, J.C. Vergnaud, Decision making with imprecise probabilistic information, Journal of Mathematical Economics 40 (2004) 647–681.
[26] R.G. Gallager, Source coding with side information and universal coding, Technical Report LIDS-P-937, Laboratory for Information Decision Systems, MIT, 1979.
[27] C. Genest, J.V. Zidek, Combining probability distributions: a critique and an annotated bibliography, Statistical Science 1 (1986) 114–135.
[28] G. Gibson, Hints of hidden heritability in GWAS, Nature Genetics 42 (2010) 558–560.
[29] I. Good, A Bayesian interpretation of ancillarity, Journal of Statistical Computation and Simulation 19 (4) (1984) 302–308.
[30] I.J. Good, Significance tests in parallel and in series, Journal of the American Statistical Association 53 (1958) 799–813.
[31] P. Grünwald, A.P. Dawid, Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, Annals of Statistics 32 (2004) 1367–1433.
[32] D. Haussler, A general minimax result for relative entropy, IEEE Transactions on Information Theory 43 (1997) 1276–1280.
[33] W.-J. Hong, R. Tibshirani, G. Chu, Local false discovery rate facilitates comparison of different microarray experiments, Nucleic Acids Research 37 (22) (2009) 7483–7497.
[34] R.L. Keeney, H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs, Cambridge University Press, Cambridge, 1993.
[35] T. Koshy, Discrete mathematics with applications, Academic Press, 2004.
[36] J. Kracík, Combining marginal probability distributions via minimization of weighted sum of Kullback–Leibler divergences, International Journal of Approximate Reasoning 52 (2011) 659–671.
[37] I. Levi, Hard Choices: Decision Making under Unresolved Conflict, Cambridge University Press, Cambridge, 1986.
[38] I. Levi, The paradoxes of Allais and Ellsberg, Economics and Philosophy 2 (1986) 23–53.
[39] K.J. McConway, Marginalization and linear opinion pools, Journal of the American Statistical Association 76 (1981) 410–414.
[40] K. Nakagawa, F. Kanaya, A new geometric capacity characterization of a discrete memoryless channel, IEEE Transactions on Information Theory 34 (1988) 318–321.
[41] J.-H. Park, S. Wacholder, M.H. Gail, U. Peters, K.B. Jacobs, S.J. Chanock, N. Chatterjee, Estimation of effect size distribution from genome-wide association studies and implications for future discoveries, Nature Genetics 42 (2010) 570–575.
[42] E. Pfaffelhuber, Minimax Information Gain and Minimum Discrimination Principle, in: I. Csiszár, P. Elias (Eds.), Topics in Information Theory. Colloquia Mathematica Societatis János Bolyai, Vol. 16, János Bolyai Mathematical Society and North-Holland, 1977, pp. 493–519.
[43] A.M. Polansky, Observed Confidence Levels: Theory and Application, Chapman and Hall, New York, 2007.
[44] J. Rissanen, Information and Complexity in Statistical Modeling, Springer, New York, 2007.
[45] B. Ryabko, Encoding of a source with unknown but ordered probabilities, Problems of Information Transmission 15 (1979) 71–77.
[46] B. Ryabko, Comments on 'A source matching approach to finding minimax codes' by Davisson, L. D. and Leon-Garcia, A, IEEE Transactions on Information Theory 27 (1981) 780–781.
[47] L.J. Savage, The Foundations of Statistics, John Wiley and Sons, New York, 1954.
[48] T. Schweder, N.L. Hjort, Confidence and likelihood, Scandinavian Journal of Statistics 29 (2002) 309–332.
[49] T. Seidenfeld, A contrast between two decision rules for use with (convex) sets of probabilities: Γ-maximin versus E-admissibility, Synthese 140 (2004) 69–88.
[50] T. Sellke, M.J. Bayarri, J.O. Berger, Calibration of p values for testing precise null hypotheses, American Statistician 55 (2001) 62–71.
[51] N. Shulman, M. Feder, The uniform distribution as a universal prior, IEEE Transactions on Information Theory 50 (2004) 581–586.

[52] J.D. Storey, A direct approach to false discovery rates, Journal of the Royal Statistical Society B 64 (2002) 479–498.
[53] M. Toda, Information-receiving behavior of man, Psychological Review 63 (1956) 204–212.
[54] F. Topsøe, Information theory at the service of science, in: I. Csiszár, G.O.H. Katona, G. Tardos, G. Wiener (Eds.), Entropy, Complexity. Bolyai Society Mathematical Studies, Springer, Berlin Heidelberg, 2007, pp. 179–207.
[55] E. van Berkum, H. Linssen, D. Overdijk, Inference rules and inferential distributions, Journal of Statistical Planning and Inference 49 (1996) 305–317.
[56] J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, 1953.
[57] A. Wald, Statistical Decision Functions, John Wiley and Sons, New York, 1961.
[58] Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 447 (2007) 661–678.
[59] P.H. Westfall, W.O. Johnson, J.M. Utts, A Bayesian perspective on the Bonferroni adjustment, Biometrika 84 (1997) 419–427.
[60] Y. Yang, D.R. Bickel, Minimum description length and empirical Bayes methods of identifying SNPs associated with disease. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 74, 2010, biostats.bepress.com/cobra/ps/art74.