INFORMATION-THEORETIC PROBABILITY COMBINATION WITH APPLICATIONS TO RECONCILING STATISTICAL METHODS

David R. Bickel, University of Ottawa

1. MOTIVATION

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 [1], it also increases variance [2] and can substantially increase bias when the data distributions have heavy tails [3]. 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.

Emphasizing the reference class problem, [4] 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 [5], combining inferences from more robust procedures with those from procedures with stronger assumptions, and combining inferences from different alternative distributions [6]. 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, [7], [8], and [9] 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 [7, 9].

Such *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 [10]. The marginalization property forbids certain counterintuitive combinations of distributions, including any combination of distributions that differs in a probability assignment from the unanimous assignment of all distributions combined [11, p. 173]. Combinations violating the marginal-

ization 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., [12, §4.11.1] or [13]. 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.

2. GAME-THEORETIC FRAMEWORK

Like the opinion pools of Section 1, the strategy introduced in [14] 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. (This game is distinct from those of [15, 16], which apply [17] to blending frequentist and Bayesian statistical methods.)

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 combine 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.

[14] 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 special case of combining discrete distributions is summarized in the next section. 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 appear in [14] with an application to the combination of three false discovery rate methods for the analysis of microarray data.

3. SPECIAL CASE: COMBINING DISCRETE DISTRIBUTIONS

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 $\Xi = \{0, 1, ..., |\Xi| - 1\}$ without loss of generality. Then the information divergence of $P \in \mathcal{P}$ with respect to $Q \in \mathcal{P}$ reduces to

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

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

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

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.

Lemma. *Let* $\mathcal{P}^\star$ denote a nonempty, finite subset of $\mathcal{P}$*, and let* $\operatorname{ext}\mathcal{P}^\star$ *denote the set of distributions that are represented by the extreme points of the convex hull of the set of tuples representing the members of* $\mathcal{P}^\star$*. 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* $\mathcal{P}^\star$*.*

Proof. As an immediate consequence of what [18] labels "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 redundancy* problem. □

The **Theorem** in [14] that connects the lemma to the following corollary is based on the *redundancy-capacity theorem*, the celebrated relationship between capacity and minimax redundancy. The redundancy-capacity theorem was presented by R. G. Gallager in 1974 [19, Editor's Note] and published as [20] and [21]; cf. [22]. [23, Theorem 13.1.1], [24, §5.2.1], and [25, Problem 8.1] provide useful introductions. The extension from discrete distributions to general probability measures ([26]; [27]) is exploited in [14].

The combination of a set of probabilities of the same hypothesis 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. *Let* P^+ *denote the combination of the distributions in* $\ddot{\mathcal{P}} \subseteq \mathcal{P}$ *with truth constrained by* $\dot{\mathcal{P}} \subseteq \mathcal{P}$*. Suppose* c *distributions on* $\left(\{0,1\}, 2^{\{0,1\}}\right)$ *are to be combined* $\left(\ddot{\mathcal{P}} = \left\{\ddot{P}_1, ..., \ddot{P}_c\right\}\right)$*, and let* $\dot{\mathfrak{P}}_0 = \left\{\dot{P}\left(\{0\}\right) : \dot{P} \in \dot{\mathcal{P}}\right\}$ *and* $\underline{\ddot{P}}, \overline{\ddot{P}} \in \mathcal{P}$ *such that* $\underline{\ddot{P}}\left(\{0\}\right) = \min \ddot{P}_i\left(\{0\}\right)$ *and* $\overline{\ddot{P}}\left(\{0\}\right) = \max \ddot{P}_i\left(\{0\}\right)$*. If there is at least one* $i \in \{1, ..., c\}$ *for which* $\ddot{P}_i\left(\{0\}\right) \in \dot{\mathfrak{P}}_0$ *holds, then* $P^+ = w^+ \underline{\ddot{P}} + \left(1 - w^+\right) \overline{\ddot{P}}$*, where* $w^+ =$

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

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

4. ACKNOWLEDGMENTS

Most of this extended abstract is derived from [14] with permission from Elsevier.

5. REFERENCES

[1] Bradley Efron, "Size, power and false discovery rates," *Annals of Statistics*, vol. 35, pp. 1351–1377, 2007.

[2] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press, Cambridge, 2010.

[3] D. R. Bickel, "Estimating the null distribution to adjust observed confidence levels for genome-scale screening," *Biometrics*, vol. 67, pp. 363–370, 2011.

[4] Ole E. Barndorff-Nielsen, "Diversity of evidence and Birnbaum's theorem," *Scandinavian Journal of Statistics*, vol. 22, pp. 513–515, 1995.

[5] IJ Good, "A Bayesian interpretation of ancillarity," *Journal of Statistical Computation and Simulation*, vol. 19, no. 4, pp. 302–308, 1984.

[6] I. J. Good, "Significance tests in parallel and in series," *Journal of the American Statistical Association*, vol. 53, pp. 799–813, 1958.

[7] M Toda, "Information-receiving behavior of man," *Psychological Review*, vol. 63, pp. 204–212, 1956.

[8] A. E. Abbas, "A Kullback-Leibler View of Linear and Log-Linear Pools," *Decision Analysis*, vol. 6, pp. 25–37, 2009.

[9] Jan Kracík, "Combining marginal probability distributions via minimization of weighted sum of Kullback-Leibler divergences," *International Journal of Approximate Reasoning*, vol. 52, pp. 659–671, 2011.

[10] K. J. McConway, "Marginalization and linear opinion pools," *Journal of the American Statistical Association*, vol. 76, pp. 410–414, 1981.

[11] Roger M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press, 1991.

[12] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, New York, 1985.

[13] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis," *Risk Analysis*, vol. 19, pp. 187–203, 1999.

[14] D. R. Bickel, "Game-theoretic probability combination with applications to resolving conflicts between statistical methods," *International Journal of Approximate Reasoning*, vol. 53, pp. 880–891, 2012.

[15] D. R. Bickel, "Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes," *Electron. J. Statist.*, vol. 6, pp. 686–709, 2012.

[16] D. R. Bickel, "Blending Bayesian and frequentist methods according to the precision of prior information with applications to hypothesis testing," *Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/23124*, 2012.

[17] Flemming Topsøe, "Information theoretical optimization techniques," *Kybernetika*, vol. 15, no. 1, pp. 8–27, 1979.

[18] K. Nakagawa and F. Kanaya, "A new geometric capacity characterization of a discrete memoryless channel," *IEEE Transactions on Information Theory*, vol. 34, pp. 318–321, 1988.

[19] 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*, vol. 27, pp. 780–781, 1981.

[20] B.Y. Ryabko, "Encoding of a source with unknown but ordered probabilities," *Prob. Pered. Inform.*, vol. 15, pp. 71–77, 1979.

[21] L. Davisson and a. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Transactions on Information Theory*, vol. 26, pp. 166–174, 1980.

[22] Robert G. Gallager, "Source coding with side information and universal coding," *Technical Report LIDS-P-937, Laboratory for Information Decision Systems, MIT*, 1979.

[23] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, 2006.

[24] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer, New York, 2007.

[25] Imre Csiszár and János Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, Cambridge, 2011.

[26] D Haussler, "A general minimax result for relative entropy," *IEEE Transactions on Information Theory*, vol. 43, pp. 1276 – 1280, 1997.

[27] P.D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Annals of Statistics*, vol. 32, pp. 1367–1433, 2004.