Combining the Dichotomous Opinions of Several Exchangeable Experts*

by

Stephen M. Samuels
Purdue University

Technical Report#86-34

Department of Statistics
Purdue University

August 1986
(Revised October 31, 1986)

* AMS 1980 subject classifications. 60E15, 62C99
Key words and phrases. Consensus, expert opinions, beta prior, Bayesian inference.

Combining the Dichotomous Opinions of Several Exchangeable Experts *

by
Stephen M. Samuels
Purdue University

Revised October 31, 1986

Abstract

Each of m "experts" announces whether or not a particular event, A, will occur. These opinions are assumed to be drawn from an infinite exchangeable sequence. That assumption provides an effective way to model all the dependence relations solely in terms of prior distributions, and to deal with the question: "How much does dependence among so-called *experts* hurt when combining their opinions?" in a way which should appeal to both Bayesians and frequentists.

1 Introduction

A vexing problem in the design of so-called *expert systems* is how to process *uncertain input*: expert *opinions* rather than facts. A major difficulty is that experts tend to be highly dependent due to such things as common knowledge and shared environment. Dependency can have drastic effects. Consider the following standard example reported in Berger (1985, page 307):

> Experts $1,\ldots,m$ report estimates $X_1,\ldots,X_m$ for θ. Suppose $\mathbf{X} = (X_1,\ldots,X_m)'$ is $\mathcal{N}_m(\theta\mathbf{1}, \mathfrak{T})$, where $\mathbf{1} = (1,\ldots,1)'$ and $\mathfrak{T}$ has diagonal elements 1 and known off–diagonal elements $\rho > 0$. (Thus ρ reflects the fact that there is a dependence among experts.) Suppose θ is given the noninformative prior $\pi(\theta) = 1$. Rewriting the posterior variance as $\rho + (1-\rho)m^{-1}$ shows that an infinite number of dependent experts, with $\rho = 0.2$, can convey no more accurate information than five independent experts ($\rho = 0$).

Modeling dependent uncertainty has seemed so complicated that, in practice, various grossly over-simplified, and patently illogical devices are used. Their creators are often well aware of these deficiencies; for example, in the famous expert system, *Mycin*, Shortliffe (1976) uses what he calls a "model of inexact reasoning". They argue that they have no choice because models incorporating uncertainty appear to be quite arbitrary and to involve numerous hard-to-specify parameters. This paper is an attempt to explore the possibilities of, to some extent, overcoming those objections. It also provides further dramatic evidence of the drastic impact of dependence.

*AMS 1980 subject classifications. 60E15, 62C99
Key words and phrases. Consensus, expert opinions, beta prior, Bayesian inference, detection, prediction.

Specifically, we shall consider the case of a single event, say A, which may represent, for example, the occurrence of rain tomorrow, or a fire or an earthquake. We have also m dichotomous expert opinions, $X_1, \ldots, X_m$, with

$$X_i = \begin{cases} 1 & \text{"A has occurred (or will occur)"} \\ 0 & \text{otherwise.} \end{cases}$$

We shall assume infinite exchangeability of the opinions, given A, and likewise, given its complement, $\tilde{A}$, the consequences of which will be made precise in Section 2. This may be regarded as an assumption about the experts, or an assumption about the events (e.g. the fires), or both.

2 The General Model

Of primary interest is the calculation of $P(A \mid X_1, \ldots, X_m)$. This can be written, using Bayes' Theorem in odds ratio form, as

$$\frac{P(A \mid X_1, \ldots, X_m)}{P(\tilde{A} \mid X_1, \ldots, X_m)} = \frac{P(A)}{P(\tilde{A})} \cdot \frac{P(X_1, \ldots, X_m \mid A)}{P(X_1, \ldots, X_m \mid \tilde{A})}. \tag{1}$$

The interpretation here is that a Bayesian Decision Maker, as in, e.g., Lindley (1985), multiplies his or her prior odds by the Likelihood Ratio of A to $\tilde{A}$ provided by the expert opinions, considered as data, to obtain the posterior odds. So the problem is in modeling $P(X_1, \ldots, X_m \mid A)$, and $P(X_1, \ldots, X_m \mid \tilde{A})$.

When the expert opinions are *IID*, with $P(X_i = 1 \mid A) = \pi_A$, we have

$$P(X_1, \ldots, X_m \mid A) = (\pi_A)^{\sum X_i}(1 - \pi_A)^{\sum(1-X_i)}. \tag{2}$$

If we assume infinite exchangeability of $X_1, \ldots, X_m, \ldots$, then, by the well-known de Finetti Theorem, see, e. g., Feller (1971, page 228), there is the representation

$$P(X_1, \ldots, X_m \mid A) = \mathbf{E}(\pi_A)^{\sum X_i}(1 - \pi_A)^{\sum(1-X_i)}, \tag{3}$$

and a similar expression for $P(X_1, \ldots, X_m \mid \tilde{A})$, where π_A and $\pi_{\tilde{A}}$ are now random variables. Their distributions may be regarded as the decision maker's prior opinion of how the infinite pool of experts will respond if A does or does not occur, respectively. In choosing these two priors, the decision-maker has completely specified the dependence relations among the expert opinions.

For binomial likelihoods, the most appealing choice is the conjugate prior family of beta distributions:

$$f_\pi(x;\, \alpha, \beta) = \frac{\Gamma(\alpha + \beta + 2)}{\Gamma(\alpha + 1)\Gamma(\beta + 1)} \cdot x^\alpha (1 - x)^\beta \qquad 0 \le x \le 1,$$

with parameters $\alpha > -1$ and $\beta > -1$, for which,

$$\mathbf{E}\pi^j(1 - \pi)^k = \frac{\Gamma(\alpha + \beta + 2)}{\Gamma(\alpha + 1)\Gamma(\beta + 1)} \frac{\Gamma(\alpha + 1 + j)\Gamma(\beta + 1 + k)}{\Gamma(\alpha + \beta + 2 + j + k)}. \tag{4}$$

Substituting $\sum X_i$ and $\sum(1 - X_i)$ for j and k, and choosing values, $\alpha_A, \beta_A, \alpha_{\tilde{A}}$, and $\beta_{\tilde{A}}$, evaluates (3), as well as the corresponding expression for $P(X_1, \ldots, X_m \mid \tilde{A})$. These choices might be made either directly or from the first two moments:

$$\mathbf{E}\pi = \frac{\alpha + 1}{\alpha + \beta + 2} = \mu, \qquad \mathbf{E}\pi^2 = \frac{(\alpha + 2)(\alpha + 1)}{(\alpha + \beta + 3)(\alpha + \beta + 2)} = \nu.$$

Alternatively, the four parameters can be obtained from the means and correlations. Under exchangeability, it is always the case that specifying the first two moments of the mixing distribution determines the correlation, as the following simple argument shows:

Let X and Y denote any two such dichotomous opinions, where π has mean μ and standard deviation σ. Then the correlation, $\rho_{X,Y}$, satisfies:

$$\begin{aligned}
\rho_{X,Y} &= \frac{\mathbf{E}XY - (\mathbf{E}X)(\mathbf{E}Y)}{\sigma_X \sigma_Y} \qquad \text{(in general)} \\
&= \frac{P(X=Y=1) - P(X=1)P(Y=1)}{\sqrt{[P(X=1) - P^2(X=1)][P(Y=1) - P^2(Y=1)]}} \qquad \text{(for } 0-1 \text{ r.v.'s)} \\
&= \frac{\mathbf{E}\pi^2 - (\mathbf{E}\pi)^2}{\mathbf{E}\pi - (\mathbf{E}\pi)^2} \qquad \text{(under exchangeability)} \\
&= \frac{\mathbf{var}(\pi)}{(\mathbf{E}\pi)(1-\mathbf{E}\pi)} = \frac{\sigma^2}{\mu(1-\mu)}. \square
\end{aligned} \tag{5}$$

For beta priors, this becomes

$$\rho = \frac{\nu - \mu^2}{\mu(1-\mu)} = \frac{1}{\alpha + \beta + 3}; \tag{6}$$

hence

$$\alpha + 1 = \mu(\rho^{-1} - 1), \qquad \beta + 1 = (1-\mu)(\rho^{-1} - 1). \tag{7}$$

2.1 Modeling Prediction

The choice of parameter values will generally be influenced by whether the situation being modeled is one of *detection* or of *prediction*. In the language of Morris (1986), the former is his

> *Model 3.* An event probability is viewed as a single piece of data whose likelihood depends primarily on the occurrence or nonoccurrence of the event in question.

while the latter is

> *Model 4.* The expert's probability is viewed as an estimate of the frequency of a sequence of exchangeable events...

Fire alarms and seismic detectors are *Model 3* examples, while the meteorologist announcing whether or not it will rain tomorrow, is a classic *Model 4* example. In the latter case, we might follow Winkler (1986, page 300) by choosing

$$\alpha_A = \alpha_{\tilde{A}} + 1 \qquad \beta_{\tilde{A}} = \beta_A + 1,$$

for which the posterior odds becomes

$$\frac{P(A \mid X_1, \ldots, X_m)}{P(\tilde{A} \mid X_1, \ldots, X_m)} = \frac{\alpha_{\tilde{A}} + 1 + \sum X_i}{\beta_A + 1 + \sum(1 - X_i)}. \tag{8}$$

When the proportion of experts predicting rain, $\sum X_i/m$, equals the decision maker's prior probability, $(\alpha_{\tilde{A}} + 1)/(\alpha_{\tilde{A}} + \beta_A + 2)$, the above posterior odds are the same as the prior odds. This is analogous to Winkler's result.

3 Effect of Dependence

To further examine the effect of dependence under exchangeability, let us consider an interesting special case, motivated by examples like the "fire alarm". Of particular interest in such cases is $P(A \mid 0,\ldots,0)$, the probability of all m "alarms" failing to report a "fire"; and especially its behavior as $m \to \infty$. As (1) shows, this depends essentially on (3)—which now becomes

$$P(\text{All } m \text{ experts wrong} \mid A) = \mathbf{E}(1-\pi_A)^m \tag{9}$$

—and the corresponding expression, given $\tilde{A}$. This is a *frequentist* type of measure, which should be of considerable interest to non-Bayesians as well as to Bayesians.

3.1 Beta Priors

When π_A and $\pi_{\tilde{A}}$ have beta distributions (9) is of the form:

$$\begin{aligned} \mathbf{E}(1-\pi)^m &= \frac{(\beta+1)(\beta+2)\cdots(\beta+m)}{(\alpha+\beta+2)(\alpha+\beta+3)\cdots(\alpha+\beta+m+1)} \\ &= O\left(m^{-(\alpha+1)}\right) \qquad \text{as } m \to \infty \end{aligned} \tag{10}$$

(Note that β does not appear, asymptotically.) Hence, the Likelihood Ratio in (1) is

$$\frac{P(0,\ldots,0 \mid A)}{P(0,\ldots,0 \mid \tilde{A})} = O\left(m^{-(\alpha_A - \alpha_{\tilde{A}})}\right). \tag{11}$$

Let us look more closely at the implications of (10). Obviously the rate of convergence to zero is much slower than $(1-\mu)^m$, the corresponding rate for independent experts, especially when the correlation is large. The slowest convergence, slower than m^{-1}, occurs for negative α; i.e. when π has a mode at zero. The bimodal beta distributions (α and β negative) model situations in which there is considerable agreement among the experts, who are sometimes right (perhaps for certain kinds of fires) and sometimes wrong (perhaps for other kinds of fires).

As an illustration of the slow convergence, inspired by the example in the introduction, we may ask

> How many dependent experts, with prescribed correlation, ρ, can detect an event as reliably as k independent experts ($\rho = 0$)?

The answer here depends on μ as well as ρ; from (10) it is

$$m(\rho,\mu) = \min\left\{ m : \prod_{j=1}^{m-1}\left(1 - \frac{1-\rho}{1-\rho+j\rho}\mu\right) \le (1-\mu)^{k-1} \right\}. \tag{12}$$

A lower bound is the smallest m for which

$$\prod_{j=1}^{m-1}\left(1 - \frac{1-\rho}{1-\rho+j\rho}\right) \le (1-\mu)^{k-1}, \tag{13}$$

which shows that, for fixed $\rho < 1$, $m(\rho,\mu) \uparrow \infty$ as $\mu \uparrow 1$. Thus the effect of dependence is nearly as dramatic here as in the Berger (1985) example: The necessary number of dependent experts, while never infinite, is, in a sense, unbounded. Some explicit values for $\rho = 0.2$ are as follows: we need $m \ge 10$ for $\mu = 0.25$, $m \ge 12$ for $\mu = 0.50$, $m \ge 18$ for $\mu = 0.75$, and $m \ge 37$ for $\mu = 0.875$.

3.2 Partial Prior Knowledge

How much can be said when only the first and second moments of π_A and $\pi_{\tilde{A}}$ are specified, but the exact form of the distributions is unknown? Formally, the problem is to find extremal values of $\mathbf{E}(1-\pi)^m$ among *all* distributions of π with support in $[0,1]$ and prescribed moments:

$$\mathbf{E}\pi = \mu \qquad \text{and} \qquad \mathbf{E}\pi^2 = \nu = \mu^2 + \rho\mu(1-\mu).$$

Because the four functions: $1, \pi, \pi^2$, and $(1-\pi)^m$ form a *Tchebycheff system*, the techniques in Karlin and Studden (1966) yield, easily, the following solution: The upper bound,

$$\frac{\rho(1-\mu)}{\mu+\rho(1-\mu)} + \frac{\mu}{\mu+\rho(1-\mu)}\left(1-\mu-\rho(1-\mu)\right)^m,$$

puts all mass on zero and $\mu+\rho(1-\mu)$, and has a non-zero limit, as $m \to \infty$; while the lower bound,

$$\frac{1-\mu}{1-\mu+\rho\mu}\left(1-\mu+\rho\mu\right)^m,$$

puts all mass at 1 and $\mu-\rho\mu$, and goes to zero exponentially fast.

Thus there is a very wide range of possible asymptotic behavior when only the two moments are specified. Might we get more useful bounds if we, say, restricted the class of mixing distributions to just the unimodal ones? The answer is no; but we can get the following interesting upper bound on the correlation:

Proposition 1 *If π is unimodal, then $\rho \leq 2/3$.*

Proof. For any unimodal density, f, on $(0,1)$, with mode ϕ, and mean μ, let

$$dH = -(x-\phi)df.$$

Then dH is a legitimate measure, with

$$\int_0^1 dH = 1, \qquad \int_0^1 x\, dH = 2\mu - \phi,$$

and

$$\int_0^1 x^2\, dH = 3\int_0^1 x^2\, f(x)dx + 2\phi\mu.$$

A very general *supporting hyperplane* method of obtaining moment inequalities, described in Karlin and Studden (1966, chapter 12), yields

$$\max_H \int_0^1 x^2\, dH(x) = \min_{\{A,B:\ A+Bt \geq t^2 \text{ on } [0,1]\}} \left[A + B(2\mu-\phi)\right].$$

The above minimum is $2\mu-\phi$, attained for $A=0$ and $B=1$, so

$$\nu \equiv \int_0^1 x^2\, f(x)dx \leq \frac{2\mu-\phi+2\phi\mu}{3}.$$

If we now maximize with respect to ϕ, we have

$$\nu \leq \begin{cases} 2\mu/3 & \text{if } \mu \leq 1/2 \\ (4\mu - 1)/3 & \text{if } \mu \geq 1/2 \end{cases}$$

so, using (5),

$$\rho = \frac{\nu - \mu^2}{\mu(1-\mu)} \leq \begin{cases} (2 - 3\mu)/3(1-\mu) & \text{if } \mu \leq 1/2 \\ (3\mu - 1)/3\mu & \text{if } \mu \geq 1/2 \end{cases}$$

In both cases, the supremum over μ is $2/3$. □

(Notice that, for the unimodal beta densities: $\alpha \geq 0, \beta \geq 0$, (6) shows that the correlation is at most $1/3$.)

3.3 General Asymptotic Behavior

The fact that β does not appear, asymptotically, in (10), is indicative of the more general fact: the asymptotic behavior of $\mathbf{E}(1-\pi)^m$ depends only incidentally on the correlation; what it really depends on is how the distribution of π looks near zero. In general, if the support of π is bounded away from zero, then $\mathbf{E}(1-\pi)^m$ goes to zero exponentially fast, just as in the independent experts case. At the other extreme, if there is mass at zero, then there is no convergence to zero, as in the example from Berger (1985). The only remaining case is where the support includes a neighborhood of zero, in which case there is convergence to zero, but at a rate much slower than exponential. This is illustrated by—but by no means restricted to—the beta distributions shown above.

Finally, we remark that the above results generalize immediately to the situation of exchangeable experts giving dichotomous opinions about an exchangeable *sequence* of events rather than just a single event. As modeled here, in that situation we have merely a single sequence of indicator random variables, with each successive m of them referring to one of the events.

References

[1] Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis, 2nd edition.* Springer–Verlag, New York.

[2] Feller, William (1971). *An Introduction to Probability Theory and Its Applications, Volume II, 2nd edition.* Wiley, New York.

[3] Karlin, S. J., and Studden, W. J. (1966). *Tchebycheff Systems: with Applications in Analysis and Statistics.* Wiley Interscience, New York.

[4] Lindley, Dennis V. (1985). Reconciliation of discrete probability distributions. In *Bayesian Statistics* **2** (J. M. Bernardo et al., eds.) 375-390. North Holland, Amsterdam.

[5] Morris, Peter A. (1986). Observations on expert aggregation. *Management Science* **32** 321-328.

[6] Shortliffe, Edward H. (1976). *Computer-Based Medical Consultations: MYCIN.* Elsevier, New York.

[7] Winkler, Robert L. (1986). Expert resolution. *Management Science* **32** 298-303.