

Probability aggregation methods in geoscience

*D. Allard^a, A. Comunian^{b,c}, and P. Renard^b

^aBiostatistique et Processus Spatiaux (BioSP), INRA, Site Agroparc, 84914 Avignon, France.

^bCHYN, Université de Neuchâtel, Neuchâtel, Suisse

^cnow at National Centre for Groundwater Research and Training, University of New South Wales, Sydney, Australia

*The order of the authors is alphabetical

Running title: Probability aggregation methods in geoscience

Corresponding author: D. Allard allard@avignon.inra.fr

Summary

The need of combining in a probabilistic framework different sources of information is a frequent task in earth sciences. This can occur for example when modeling a reservoir using direct geological observations, geophysics, remote sensing, training images etc. For example, the probability of occurrence of a certain lithofacies at a certain location can easily be computed conditionally on the values observed at each source of information. The problem of aggregating these different conditional probability distributions into a single conditional distribution arises as an approximation to the inaccessible genuine conditional probability given all information, since building a full probabilistic model is in general impossible. This paper makes a formal review of most aggregation methods proposed so far in the literature with a particular focus on their mathematical properties. The exact relationships relating the different methods is emphasized. The case of events with more than 2 possible outcomes, never explicitly studied in the literature, is treated in details. It is shown that in this case, equivalence between different aggregation formulas is lost. It is then shown how the optimal parameters of the aggregating methods can be estimated when training data are accessible and statistical tools are proposed for assessing the goodness-of-fit of these different formula.

The concepts of calibration, sharpness and reliability, well known in the weather forecasting community are introduced. On a simulation study mimicking common situations in earth science it is found that the Bordley formula, also known as the Tau model, provides the best prediction, and that linear pooling provides the worst.

Keywords Data integration; Conditional probability pooling; Calibration; Sharpness; Log-linear pooling.

1 Introduction

The problem of aggregating probability assessments coming from different sources of information is probably as old as statistics and stochastic modeling.

In geosciences, Tarantola and Valette (1982) and Tarantola (2005) developed the concept of conjunction and disjunction of probabilities in the context of inverse problems. Benediktsson and Swain (1992) adopted consensus theoretic classification methods to aggregate geographical data like satellite images coming from different sources. Journel (2002) proposed the Tau model in a very broad perspective. This model was subsequently used for example by Strebelle et al. (2003) to map lithofacies using seismic information and multiple-point statistics, and by Caers (2006) to combine the probability assessments derived from different 2D geostatistical models to simulate 3D geological structures. Okabe and Blunt (2004, 2007) used a linear probability combination method to simulate 3D porous medium from 2D multiple-point statistics extracted from microscope images of a rock sample. Mariethoz et al. (2009) used the probability conjunction method to develop a collocated co-simulation algorithm allowing to model any complex probability relationship between the primary and secondary variable. Ranjan and Gneiting (2010) combined weather forecasts coming from different models with the beta-transformed linear opinion pool (BLP).

In the context of risk analysis, Genest and Zidek (1986) and Clemen and Winkler (1999, 2007) provide detailed reviews about probability aggregation methods and their properties.

The diversity of approaches one can find in the literature may be surprising, but this is because aggregating probabilities is usually an ill posed problem: there is often in practice a lack of information and models to describe accurately the interactions between the sources of information. In that framework, we are left with making assumptions and select a method without being able to check the accuracy of the estimations. In other words, there is neither a single method nor a single set of parameters (as several methods are parametric) that can aggregate probabilities accurately under all possible circumstances. Instead, the selection of the most suitable aggregation method depends on the specific problem which is addressed; a clear understanding of the properties characterizing each aggregation method is therefore an important step.

Clemen and Winkler (1999) proposes a classification of the probability aggregation methods into mathematical combination methods and behavioral approaches. Behavioral approaches are based on the interaction among experts; the aggregation process concludes with an agreement about a common probability term. Note that in the context of behavioral approaches the word “interaction” has a meaning strictly related to the fact that the experts are human beings who can exchange advice and discuss their assessments. In the context of geoscience, there are no exchange of information between different sources. We thus restrict ourselves to mathematical aggregation methods which are functions or operators aggregating probability distributions P_i coming from different sources into a global probability distribution P_G .

We shall provide an overview of most of the available techniques to aggregate probability distributions and discuss their properties in the perspective of Earth sciences applications. In section 3, we define the main mathematical properties of the aggregation methods. We then review the aggregation methods of interest for the Earth Sciences (section 4). Most of the methods are parametric, and therefore section 5 contains a brief overview of the standard approaches used to determine the parameters. Through a series of examples, section

6 illustrates the different behaviors of the methods under simple and more complex situations. Section 7 discusses the results and provides guidelines for the selection of a suitable aggregation method.

2 Set-up and notations

We wish to assess the probability of an event, denoted E , conditional on the occurrence of a set of data events, D_i , $i = 1, \dots, n$. This means that we wish to approximate the probability $P(E | D_1, \dots, D_n)$ on the basis of the simultaneous knowledge of the n conditional probabilities $P(E | D_i)$. The event E can for example be a lithofacies category at a specified location, while the data D_i can represent information provided by core samples at surrounding wells, a seismic survey, lithofacies patterns on training images or any other source of information.

For categorical events or finite discrete data, the formal probabilistic set-up is the following. We need to consider a sample space Ω such that all events E and D_i are subsets of Ω . In the case of categorical data, let \mathcal{E} be the finite set of events in Ω such that the events E_1, \dots, E_K of \mathcal{E} are mutually exclusive and exhaustive, i.e. \mathcal{E} forms a finite partition of Ω . For continuous data, the set-up is slightly more technical, but still straightforward in the context of probability measures. For the clarity of exposition, we will focus on the finite discrete set-up above; most if not all results presented in this paper still hold for continuous probability density functions.

The computation of the full conditional probability $P(E|D_1, \dots, D_n)$ necessitates a probabilistic model of the joint distribution of (E, D_1, \dots, D_n) , a task which is rarely achievable. Instead, we will build an approximation of the true conditional probability by the use of an aggregation operator (sometimes also called *pooling operator* or *pooling formula*) P_G , such that

$$P(E|D_1, \dots, D_n) \approx P_G(P(E|D_1), \dots, P(E|D_n)). \quad (1)$$

Aggregating the probabilities is an ill-posed problem because there is not a unique way

of constructing the event $D_1 \cap \dots \cap D_n$ from the knowledge of the conditional probabilities $P(E | D_i)$, $i = 1, \dots, n$. One of the aim of this paper is to discuss the mathematical properties of such operators and, elaborating from a subset of desirable properties, to build and compare some of them, both from a theoretical point of view and on the basis of performances on simulated cases.

In some circumstances, it will be necessary to include a prior probability on the event E , which will be denoted $P_0(E)$. This prior probability is independent on any other probabilities $P(E | D_i)$. It can be thought of as arising from an abstract and never specified information D_0 with $P_0(E) = P(E | D_0)$. Equation (1) can thus be generalized in the following way:

$$P(E|D_0, \dots, D_n) \approx P_G(P(E|D_0), \dots, P(E|D_n)). \quad (2)$$

In geoscience, such a prior probability could be for example a proportion of a lithofacies imposed by the user.

In the following, the more concise notation P_i will sometimes be used used to denote $P(E|D_i)$ and the RHS of the (2) will often be rewritten as: $P_G(P_0, P_1, \dots, P_n)(E)$. At the price of a small abuse of notation, we will adopt the more concise notation $P_G(E)$ for $P_G(P_0, P_1, \dots, P_n)(E)$ when the context permits.

Some probability aggregation methods are formulated in terms of *odd ratios*, denoted O , defined as

$$O(E) = \frac{P(E)}{1 - P(E)}, \quad 0 \leq P(E) < 1, \quad (3)$$

with the convention $O(E) = +\infty$ when $P(E) = 1$. In the simple case of a binary outcome, where $\mathcal{E} = \{E, \bar{E}\}$, it is easy to check that $O(E)O(\bar{E}) = 1$. When there are more than two elements in \mathcal{E} , $\prod_{k=1}^K O(E_k)$ can be any fixed value, but (3) will still be used for defining odd ratios.

3 Mathematical properties

In this section we first recall and discuss the main properties that can be used for characterizing aggregation methods. Axiomatic approaches (see *i.e.* Bordley, 1982; Dietrich, 2010) use some of these properties as starting point to derive classes of aggregation operators.

3.1 Dictatorship

Definition 1 (Dictatorship). *A method is dictatorial (Genest and Zidek, 1986) when the probability P_i provided by the i -th source of information is always taken as the group assessment, that is $P_G(P_1, \dots, P_i, \dots, P_n)(E) = P_i(E)$, for all $E \in \mathcal{E}$.*

Dictatorship is clearly a pathological property. From now on, we will focus on non dictatorial aggregation operators.

3.2 Convexity

Definition 2 (Convexity). *An aggregation operator P_G verifying*

$$P_G \in [\min\{P_1, \dots, P_n\}, \max\{P_1, \dots, P_n\}], \quad (4)$$

is convex.

Definition 3 (Unanimity). *An aggregation operator P_G verifying $P_G = p$ when $P_i = p$ for $i = 1, \dots, n$ is said to preserve unanimity.*

It is easy to check that when P_G is convex, $P_i = p$ for $i = 1, \dots, n$ implies $P_G = p$. Thus, any convex operator preserves unanimity, but the converse is not always true. Unanimity, and thus convexity, is not necessarily a desirable property, as we illustrate now in the two following cases. As a first case, consider that all sources of information yield the same probability because they are all induced by the same event of Ω , *i.e.* $D_1 = \dots = D_n$. Then, the true conditional probability can be calculated exactly: $P(E | D_1 \cap \dots \cap D_n) = P(E | D_1)$. In this first case, unanimity arises because the D_i s are all similar.

As a second case, consider that Ω is finite and consider two information $D_1 \neq D_2$ and an event $E \subset (D_1 \cap D_2)$. Then, $P(E | D_1) = P(E)/P(D_1)$, and $P(E | D_1 \cap D_2) = P(E)/P(D_1 \cap D_2)$. Now, $(D_1 \cap D_2) \subset D_1$ implies that $P(D_1 \cap D_2) < P(D_1)$. Hence $P(E | D_1 \cap D_2) > P(E | D_1)$. Thus, in this second case, the full conditional probability of E is larger than any partial conditional probability. In this situation, unanimity, and thus convexity are not desirable properties.

These examples show that whether the pieces of information are similar or different, one should expect the aggregation operator to preserve unanimity or not. Quite often in geosciences, unanimity (and convexity) is a limitation because the conditional probabilities we want to aggregate correspond to very different sources of information. In other words, in geoscience we are essentially in the second case. Therefore, unanimity, and hence convexity, are properties that should not be sought *per se*.

3.3 Independence preservation

Consider two events E and F of Ω such that $E \cap F \neq \emptyset$. Note that since \mathcal{E} is a collection of disjoint events, F is not an element of \mathcal{E} .

Definition 4 (Independence Preservation). *A method preserves the independence if, whenever we choose two events E and F for which $P_i(E \cap F) = P_i(E) P_i(F)$ is valid for every $i = 1, \dots, n$, the aggregated probability operator P_G preserves independence:*

$$P_G(P_1, \dots, P_n)(E \cap F) = P_G(P_1, \dots, P_n)(E) P_G(P_1, \dots, P_n)(F) \quad (5)$$

holds.

Many authors (Lehrer and Wagner, 1983; Genest, 1984; Wagner, 1984; Genest and Wagner, 1987) faced without success the challenge of finding a pooling formula which preserves independence. Independence preservation is of no direct interest in the context described above, since one usually wants to assess the probability of disjoint events E . Quoting Gen-

est and Zidek (1986), our conclusion is that “independence preservation is not a reasonable requirement to impose on consensus-finding procedures”.

3.4 Marginalization

Consider a vector of events $\mathbf{E} = (E_1, E_2)^t$ and $\mathbf{P}(\mathbf{E}) = (P(E_1), P(E_2))^t$. For each component, $k = 1, 2$ of \mathbf{E} one can define the *marginalization operator* M_k :

$$M_k\{\mathbf{P}(\mathbf{E})\} = P(E_k). \quad (6)$$

Definition 5 (Marginalization). *A pooling operator \mathbf{P}_G verifies the marginalization property if, for each component $k = 1, 2$, the operator M_k commutes with the pooling operator:*

$$P_G\{M_k(\mathbf{P}_1), \dots, M_k(\mathbf{P}_n)\} = M_k\{P_G(\mathbf{P}_1, \dots, \mathbf{P}_n)\}. \quad (7)$$

There is only one pooling operator satisfying the marginalization property, namely the linear pooling method. But we will see below that it does not verify other more interesting properties.

3.5 External Bayesianity

The *external Bayesianity* property is related to the behavior of an aggregation operator when additional information become available. Consider that the probabilities can be updated by a likelihood, L , common to all sources of information. We thus consider now the probabilities

$$P_i^L(E) = \frac{L(E)P_i(E)}{\sum_{E \in \mathcal{E}} L(E)P_i(E)}, \quad i = 1, \dots, n,$$

where $L(E)$ is such that $\sum_{E \in \mathcal{E}} L(E) < \infty$.

Definition 6 (External Bayesianity). *An aggregation operator is said to be external Bayesian if the operation of updating the probabilities with the likelihood L commutes with the aggregation operator, i.e. if*

$$P_G(P_1^L, \dots, P_n^L)(E) = P_G^L(P_1, \dots, P_n)(E). \quad (8)$$

In words, it means that it should not matter whether new information arrives before or after pooling. This property is equivalent to the *weak likelihood ratio* property in Bordley (1982). External bayesianity is a very compelling property, both from a theoretical point of view and from an algorithmic point of view. We will see that imposing this property leads to a very specific class of pooling operators.

3.6 Certainty effect

An interesting feature of an aggregation method is its response to situations where a source of information provides a conditional probability equal to 0 (impossible event) or 1 (certain event). Let us suppose that there exists i such that $P(E | D_i) = 0$ and $P(E | D_j) \neq 1$ for $j \neq i$.

Definition 7 (0/1 forcing property). *An aggregation operator which returns $P_G(E) = 0$ in the above-mentioned case is said to enforce a certainty effect, a property also called the 0/1 forcing property (Allard et al., 2011).*

Note that the same is true if $P(E | D_i) = 1$, since in this case $P(E' | D_i) = 0$, for all $E' \neq E \in \mathcal{E}$. In geoscience, this property is convenient to reproduce depositional sequences or catenary patterns. The drawback is that deadlock situations are possible, when $P(E | D_i) = 0$ and $P(E | D_j) = 1$ for $j \neq i$. Deadlocks can arise when data are inconsistent with each other. A practical solution can be to consider probabilities in a constrained interval, e.g. $[0.001, 0.999]$.

4 Aggregation methods

Aggregation methods can be divided into methods derived from axiomatic approaches and methods derived from model considerations.

Genest and Zidek (1986), Bordley (1982) and Dietrich (2010) restricted themselves to the binary case, i.e. when there are only two possible outcomes, namely E and \bar{E} in \mathcal{E} .

Bordley (1982) showed that there is only one class of aggregation operator verifying at the same time a set of structural axioms always verified in geoscience (weak ordering of the $O_i(E)$ s with respect to E , non interaction between source of information, continuity) and the weak likelihood ratio condition (or external Bayesianity). The associated pooling formula, hereafter called Bordley formula, combines odds multiplicatively. In the same spirit, Genest and Zidek (1986) show that the unique aggregation operator verifying the same structural axioms and external bayesianity is the log-linear pooling. These two results turn out to be exactly similar in the binary case, but lead to different pooling formulas in the general case of more than two possible outcomes. Still in the binary case, Dietrich (2010) shows that for a very close set of structural axioms, the only pooling formula verifying the property of “independent information” is a particular case of the log-linear pooling formula.

Following a model-based approach, Journel (2002) proposed the tau-model, which turns out to be exactly similar to the Bordley formula (Krishnan, 2008). In Polyakova and Journel (2007), the Nu-model is proposed as an alternative to the Tau model. Although no mentions are explicitly made in these papers to any restriction to the binary case, it must be noted that it is in fact the case for all considered examples. It turns out that it is equivalent to work with probabilities or with odds in the binary case. This equivalence is lost if there are more than two possible outcomes in \mathcal{E} . We will show that there are two quite different routes for generalizing the Nu model to the general case. We will also show how this Nu-model is related to log-linear pooling methods and that following a maximum entropy principle or equivalently a conditional independence assumption entails a specific, parameter-free form of the Bordley formula. The resulting pooling formula is similar to the Markovian-type Categorical Prediction (MCP) equations in Allard et al. (2011).

There is yet another enlightening dichotomy. Some methods combine the information in an additive way, leading to linear pooling formula and its generalization, in the spirit of the *disjunction* operation of probability distributions (Tarantola and Valette, 1982; Taran-

tola, 2005). Other methods combine probabilities or odds in a multiplicative way, which corresponds to the *conjunction* operation of probability distributions (Tarantola and Valette, 1982; Tarantola, 2005). This last criterion defines two very different groups within which the aggregation methods share many common properties.

The next subsections, following and extending the work of Genest and Zidek (1986), Clemen and Winkler (1999) and Clemen and Winkler (2007), provide a summary of some of the most important aggregation methods in Earth sciences.

4.1 Additive methods and transformed additive methods

Linear pooling

Probably the most intuitive way of aggregating the probabilities P_1, \dots, P_n is the *linear pooling*, proposed by Stone (1961) and attributed to Laplace by Bacharach (1979):

$$P_G(E) = \sum_{i=1}^n w_i P_i(E), \quad (9)$$

where the w_i are positive weights verifying $\sum_{i=1}^n w_i = 1$ in order to have a meaningful global probability. Since the linear pooling is simple to understand and to implement, it is probably the most common aggregation method. However Ranjan and Gneiting (2010) demonstrated that the linear pooling is intrinsically suboptimal. This point will be illustrated in the next sections.

It does not verify independence preservation and 0/1 forcing properties, nor external bayesianity unless it is dictatorial (i.e., $w_i = 1$ for one source D_i and $w_j = 0$, for all $j \neq i$). It is a convex aggregation method, and as a consequence, it does preserve unanimity. As already discussed in Section 3.2, this property might be considered as a serious limitation in the context of geoscience modeling. If we provide an equal weight w_i to every probability P_i the method reduces to an arithmetic average; in this case it coincides with the *disjunction* of probabilities (Tarantola and Valette, 1982; Tarantola, 2005).

Genest (1984) proved that all pooling operators verifying the marginalization property

are of the form

$$P_G(E) = \sum_{i=0}^n w_i P_i(E), \quad (10)$$

where P_0 is a prior probability and where the weights $w_0, \dots, w_n \in [-1, 1]$ add up to one and must satisfy other consistency conditions to ensure that P_G is a probability measure. The aggregation operator defined by (10) is called *generalized linear pooling*. The possibility of negative weights is in theory interesting, but we are faced with the problem of finding weights w_i insuring that P_G is a probability on \mathcal{E} . A safe option is to restrict ourselves to weights $w_0, \dots, w_n \in [0, 1]$ adding to 1. If $w_0 = 0$ we are back to the linear opinion pool.

The resulting probability distribution P_G is very often multi-modal, a not so desired situation. The reasons are profound. From a probabilistic point of view (9) and (10) represent mixture models in which each probability P_i represents a different population; the aggregated probability P_G is then the result of the following hierarchical random experiment: first select a population i with the probability distribution defined by $\mathbf{w} = (w_0, \dots, w_n)$; then select an event E according to probability distribution P_i . In general, this mixture of population model does not correspond to our geoscience context in which we wish to aggregate partial information on the same object.

Beta-transformed Linear Pooling

Ranjan and Gneiting (2010) proposed to apply a Beta transformation to linear pooling operators in order to improve their performance, thereby defining the Beta-transformed Linear Pooling (BLP):

$$P_G(E) = H_{\alpha,\beta} \left(\sum_{i=1}^n w_i P_i(E) \right), \quad (11)$$

where the weights must be positive and add up to one. The function $H_{\alpha,\beta}$ is the cumulative density function of a beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$:

$$H_{\alpha,\beta}(x) = B(\alpha, \beta)^{-1} \int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt \quad (12)$$

$$\text{with } x \in [0, 1] \text{ and } B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt.$$

BLP includes the linear pooling (LP) when $\alpha = \beta = 1$, since $H_{1,1}(x) = x$, for $0 \leq x \leq 1$. For other values of the parameters, the marginalization property verified by LP is lost because of the Beta transformation. However, as it is the case for LP, the 0/1 forcing property is not verified unless dictatorship holds. In general, this transformation leads to non convex aggregation probabilities.

In their work, Ranjan and Gneiting (2010) show, on simulations and on real case studies, that the BLP constantly outperforms LP and that it presents very good performances.

4.2 Methods based on the multiplication of probabilities

We have seen in the previous section that additive aggregation methods correspond to mixture models. They are related to union of events and to the logical operator “or”. In our context, the information conveyed by the events D_i should rather be aggregated by the logical operator “and” related to the intersection of events. Intuitively, aggregation operators based on multiplication seem therefore more appropriate than those based on addition. We now present and discuss different aggregation methods based on the multiplication of probabilities.

Log-linear pooling

Definition 8. *A log-linear pooling operator is a linear operator of the logarithms of the probabilities:*

$$\ln P_G(E) = \ln Z + \sum_{i=1}^n w_i \ln P_i(E), \quad (13)$$

or equivalently

$$P_G(E) \propto \prod_{i=1}^n P_i(E)^{w_i}, \quad (14)$$

where Z is a normalizing constant.

Genest and Zidek (1986) showed that all pooling operators verifying external Bayesianity must be of the form (14) with the additional condition that $\sum_{i=1}^n w_i = 1$. This condition also implies that unanimity is preserved. Log-linear pooling does not preserve independence and does not verify the marginalization property. Unlike linear pooling, it is typically unimodal and less dispersed. Since it is based on a product, it verifies the 0/1 forcing property. One particular possibility consists in setting $w_i = 1$ for each $i \neq 0$. This corresponds to the *conjunction of probabilities* (Tarantola and Valette, 1982; Tarantola, 2005).

If a prior probability $P_0(E)$ must be included, equation (14) becomes $P_G(E) \propto \prod_{i=0}^n P_i(E)^{w_i}$ with the restriction $\sum_{i=0}^n w_i = 1$ to verify external bayesianity, yet better written

$$P_G(E) \propto P_0(E)^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n P_i(E)^{w_i}. \quad (15)$$

In (15), there is no restriction on the weights $\mathbf{w} = (w_1, \dots, w_n)$, and $\sum_{i=0}^n w_i = 1$ is always verified.

The sum $S_{\mathbf{w}} = \sum_{i=1}^n w_i$ plays an important role in (15). If $S_{\mathbf{w}} = 1$, the prior probability P_0 is filtered out since $w_0 = 0$ and unanimity is preserved. In all other cases, unanimity may not be preserved. Suppose that $P_i = p$ for each $i = 1, \dots, n$. Now, if $S_w > 1$, the prior probability has a negative weight and P_G will always be further from P_0 than p . This corresponds to the second case illustrating convexity in Section 3. Conversely, if $S_w < 1$, P_G is always closer from P_0 than p . And of course, $P_G = p$ if $S_w = 1$. The influence of the prior probability P_0 on the aggregated result P_G can thus be tuned changing the value of the sum $S_{\mathbf{w}} = \sum_{i=1}^n w_i$.

Finally, note that if neither external Bayesianity nor unanimity are properties that must be verified, there are no constraints whatsoever on the weights w_i , $i = 0, \dots, n$.

Generalized logarithmic pooling

Genest and Zidek (1986) showed that if we allow the explicit form of P_G to depend upon E ,

that is if we allow P_G to be of the form

$$P_G(P_1, \dots, P_n)(E) \propto G(E, P_1(E) \dots, P_n(E)),$$

the only pooling operator verifying external Bayesianity is

$$P_G(E) \propto H(E) \prod_{i=1}^n P(E|D_i)^{w_i}, \quad (16)$$

with $\sum_{i=1}^n w_i = 1$ and $H(E)$ being an arbitrary bounded function playing the role of a likelihood on the elements of \mathcal{E} . In this case, if all conditional probabilities are equal, the aggregated probability is proportional to p updated by $H(E)$: $P_G(E) \propto H(E)p$.

Maximum entropy approach

Instead of establishing a pooling formula from an axiomatic point of view, one can choose to optimize a criterion, for example to minimize the distance between the distribution P and its approximation. The Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), or relative entropy, between a distribution P and another distribution (here its approximation P_G) is

$$D(P_G || P) = \mathbb{E}_{P_G} \left[\ln \frac{P_G}{P} \right]. \quad (17)$$

Although not a distance in the mathematical sense (it is not symmetrical), the KL divergence is a measure of “how much different” two probability distributions are. It is always positive and it is equal to zero if, and only if, $P_G = P$. There are strong connections between entropy and KL divergence (see Cover & Thomas, 2006). In particular, let us assume that some quantities related to P are known, such as moments or conditional probabilities. A natural approach, very common in information theory, computer science, image and language processing is to find the distribution P_G that shares properties (moments or conditional probabilities) with P and minimizes the KL divergence $D(P_G||P)$. This can be shown equivalent to finding the distribution P_G maximizing its entropy $H(P_G) = \mathbb{E}_{P_G}[P_G]$, subject to the imposed constraints. Allard et al. (2011) developed such an approach for the prediction

of spatial categorical variables leading to a *Markovian-type categorical prediction* (MCP), which was shown to be a very good approximation of the Bayesian maximum entropy (BME) principle (Christakos, 1990) with the advantage of being computationally efficient. Following a similar route, we obtain the following result. Here, we need to use the full notation $P_G(P_1, \dots, P_n)(E)$.

Proposition 1. *The pooling formula P_G maximizing the entropy subject to the following univariate and bivariate constraints $P_G(P_0)(E) = P_0(E)$ and $P_G(P_0, P_i)(E) = P(E | D_i)$ for $i = 1, \dots, n$ is*

$$P_G(P_1, \dots, P_n)(E) = \frac{P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}{\sum_{E \in \mathcal{E}} P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}. \quad (18)$$

The proof of this proposition is given in the Appendix. Notice that the maximum entropy approximation (18) is a special case of the logarithmic pooling formula with $w_i = 1$, for $i = 1, \dots, n$.

The same formula can also be obtained as a result of the conditional independence assumption. Let us introduce the following convenient notation. We will denote $D_{<i} = \{D_1 \cap \dots \cap D_{i-1}\}$, with the convention $D_{<1} = \Omega$. Then,

$$\begin{aligned} P(E | D_1, \dots, D_n) &= \frac{P_0(E)P(D_1, \dots, D_n | E)}{\sum_{E \in \mathcal{E}} P_0(E)P(D_1, \dots, D_n | E)} \\ &= \frac{P(E) \prod_{i=1}^n P(D_i | E, D_{<i})}{\sum_{E \in \mathcal{E}} P(E) \prod_{i=1}^n P(D_i | E, D_{<i})}. \end{aligned} \quad (19)$$

Let us assume that P_G verifies a conditional independence assumption. Then,

$$P_G(D_0, \dots, D_n | E) = \prod_{i=0}^n P(D_i | E), \quad (20)$$

entails

$$P_G(D_i | E, D_{<i}) = P(D_i | E).$$

Hence,

$$\begin{aligned}
P_G(E) &= \frac{P(E) \prod_{i=1}^n P(D_i | E)}{\sum_{E \in \mathcal{E}} P(E) \prod_{i=1}^n P(D_i | E)} \\
&= \frac{P_0(E)^{1-n} \prod_{i=1}^n P(E | D_i) P(D_i)}{\sum_{E \in \mathcal{E}} P_0(E)^{1-n} \prod_{i=1}^n P(E | D_i) P(D_i)} \\
&= \frac{P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}{\sum_{E \in \mathcal{E}} P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}.
\end{aligned}$$

Put together this last result and the equation (18) allows us to state the following equivalence.

Proposition 2. *Regarding the aggregation of probabilities considered in this work, Maximum Entropy is equivalent to Conditional Independence.*

Probability Multiplication formulas in summary

Multiplication of the probabilities offers a large class of pooling operators, with interesting subclasses which can be summarized in the following way:

$$\{\text{Max. Ent.} \equiv \text{Cond. Ind. pooling}\} \subset \{\text{Ext. Bayes. pooling}\} \subset \{\text{Log-linear pooling}\}. \tag{21}$$

The pooling formula corresponding to the maximum entropy principle / conditional independence assumption (18) is particularly easy to implement since it is parameter free. The larger class of pooling formula (15) corresponds to pooling operators verifying the external bayesianity condition in which the weights are constrained to add up to 1. For this class, the value of $S_{\mathbf{w}}$ is the key factor regarding the behavior with respect to the prior probability P_0 . The largest class of pooling operators is of the same form but does not impose any restriction on the weights. This largest class does not verify any mathematical properties presented in Section 3, but the 0/1 forcing property.

4.3 Methods based on the multiplication of odds

When using odds $O(E)$, it will be important to distinguish two cases:

1. In the first, more restrictive, case there are only two possible outcomes, i.e. $\mathcal{E} = \{E, \bar{E}\}$.

In this case, $P(E) + P(\bar{E}) = O(E) \cdot O(\bar{E}) = 1$. This case will be called the binary case hereafter.

2. In the second case, there are more than two possible outcomes in \mathcal{E} . In this case, there is no general relationships between the odds $O(E)$, and in general $\prod_{E \in \mathcal{E}} O(E) \neq 1$.

We will see that in the binary case, it is completely equivalent to consider operators based on the product of odds and operators based on products of probabilities. In the general case, this equivalence is lost.

Bordley formula and Tau model

Binary case

We first restrict ourselves to the binary case. Bordley (1982) showed that in this case, the only pooling operator verifying the weak likelihood ratio axiom (see Definition 6) in addition to other natural axioms is a pooling formula based on the product of the odd ratios:

$$O_G(E) = O_0(E)^{w_0} \prod_{i=1}^n \left(\frac{O_i(E)}{O_0(E)} \right)^{w_i} = O_0(E)^{w_0 - \sum_{i=1}^n w_i} \prod_{i=1}^n O_i(E) \quad (22)$$

where the weights w_i can vary in $[0, \infty)$. Now, using $P_i(E) = O_i(E)/(1 + O_i(E))$, and denoting $P_i(E) = P(E | D_i)$, (22) becomes

$$P_G(E) = \frac{P_0(E) \prod_{i=1}^n (P_i(E)/P_0(E))^{w_i}}{P_0(E) \prod_{i=1}^n (P_i(E)/P_0(E))^{w_i} + (1 - P_0(E)) \prod_{i=1}^n [(1 - P_i(E))/(1 - P_0(E))]^{w_i}}, \quad (23)$$

or equivalently,

$$P_G(E) \propto P_0(E)^{1 - \sum_{i=1}^n w_i} \prod_{i=1}^n P_i(E)^{w_i}, \quad (24)$$

which is nothing but (15). Hence, we can state the following equivalence:

Proposition 3. *In the binary case, the Bordley formula is equivalent to a log-linear pooling formula verifying external bayesianity.*

Journal (2002) derived a formula for aggregating probabilities that has been later named the Tau model. For presenting this model we will use our usual notations, which are slightly different than those in Journal (2002), Polyakova and Journal (2007) and Krishnan (2008). In particular, these authors use the inverse of odds-ratio instead of odds-ratio, but since the formulae are purely multiplicative this point is of secondary importance.

In a first step, Journal (2002) sets as an axiom the *permanence of ratio* principle, which states (using our notations) that “the incremental contribution of data event D_2 to the knowledge of E is the same after or before knowing D_1 ”. Mathematically,

$$\frac{O_G(E | D_1, D_2)}{O_G(E | D_1)} = \frac{O_G(E | D_2)}{O_G(E)}. \quad (25)$$

From this principle, one easily establish that

$$O_G(E) = O_0(E)^{1-n} \prod_{i=1}^n O_i(E),$$

i.e. a Bordley formula with $w_i = 1$, for $i = 1, \dots, n$. Replacing $O_i(E)$ by $P_i(E)/(1 + P_i(E))$, one gets $P_G(E) \propto P_0(E)^{1-n} \prod_{i=1}^n P_i(E)$, which is nothing but (18). Hence, we established the following proposition:

Proposition 4. *In the case of a binary event, the permanence of ratio principle is equivalent to conditional independence, which is equivalent to a maximum entropy principle.*

In a second step, Journal (2002) reintroduced dependence between the source of information by generalizing this formula thus obtaining the general Bordley formula (22). Krishnan (2008) provides the expression of the parameters w_i as a function of conditional probabilities obtained from the full joint probability, but this exercise is unfortunately only of academic interest, since if the full joint model was known an approximate formula such as the Tau model would not be necessary anymore.

General case

The general case with more than two possible outcomes in \mathcal{E} , was not considered in Bordley (1982). In Journal (2002), Polyakova and Journal (2007) and Krishnan (2008), the Tau model is exclusively presented in the case of binary event, either explicitly or implicitly. What happens in the general case with $K > 2$ possible outcomes is rarely addressed explicitly. In this case, the quantities $O(E_1), \dots, O(E_K)$ in (22), although computable when the probabilities belong to $[0, 1)$, are not odds in the usual sense. Back-transforming the odds into probabilities using $P_G(\cdot) = O_G(\cdot)/(1 + O_G(\cdot))$ does not lead to quantities adding to one. A normalization step is thus required to obtain a regular probability distribution. A complete formulation of the Tau model in the general case is thus

$$P_G(E) \propto O_G(E)/(1 + O_G(E)), \quad \text{with} \quad O_G(E) = O_0(E)^{1 - \sum_{i=1}^n w_i} \prod_{i=1}^n O_i(E)^{w_i}, \quad E \in \mathcal{E}. \quad (26)$$

We thus obtain the following equivalence.

Proposition 5. *The Tau model is equivalent to the Bordley formula; only in the case of a binary event, they both are equivalent to a log-linear pooling.*

Note that since $O_G(E) = 0 \Leftrightarrow P_G(E) = 0$, the Tau model (26) verifies the 0/1 forcing property, both in the binary and in the general case.

4.3.1 The nu model

The *Nu model* was proposed in Polyakova and Journal (2007) as an alternative to the Tau model. We first re-derive its expression using our notations before discussing its relationships with the other pooling methods. It will be useful to distinguish the binary case from the general case.

Binary case

Let us first consider the binary case. We start from the exact decomposition (19)

$$P(E \mid D_1, \dots, D_n) = \frac{P(E) \prod_{i=1}^n P(D_i \mid E, D_{<i})}{\sum_{E \in \mathcal{E}} P(E) \prod_{i=1}^n P(D_i \mid E) D_{<i}},$$

and we denote $\nu_i^*(E) = P(D_i | E, D_{<i})/P(D_i | E)$. Then, defining $\nu^*(E) = \prod_{i=1}^n \nu_i^*(E)$, one can write

$$\begin{aligned} P(E | D_1, \dots, D_n) &= \frac{P(E) \prod_{i=1}^n \nu_i^*(E) P(D_i | E)}{\sum_{E \in \mathcal{E}} P(E) \prod_{i=1}^n \nu_i^*(E) P(D_i | E)} \\ &= \frac{P(E)^{1-n} \nu^*(E) \prod_{i=1}^n P(E | D_i)}{\sum_{E \in \mathcal{E}} P(E)^{1-n} \nu^*(E) \prod_{i=1}^n P(E | D_i)}. \end{aligned} \quad (27)$$

From this we obtain the Nu model:

$$P_G(E) \propto P_0(E)^{1-n} \nu^*(E) \prod_{i=1}^n P(E | D_i). \quad (28)$$

In terms of odds, denoting $\nu(E) = \nu^*(E)/(1 - \nu^*(E))$,

$$O_G(E) = \frac{O(E)^{1-n} \nu(E) \prod_{i=1}^n O_i(E)}{\sum_{E \in \mathcal{E}} O(E)^{1-n} \nu(E) \prod_{i=1}^n O_i(E)}, \quad (29)$$

which is the Nu model. Note that in (28) the factors $\nu^*(E)$ are defined slightly differently than in Polyakova and Journal (2007). After transformation into $\nu(E)$, they lead however to the same analytical expression (29) the only difference being that our $\nu(E)$ is the inverse of the factor ν_0^{-1} in Polyakova and Journal (2007, Eq., 5). Remember that when applying the Nu model in practice, the quantities $\nu_i(E)$ are not known since $P(D_i | E, D_{<i})$ are unknown. They must be considered as parameters to be estimated or set by the user. From (28), one can see that $\nu^*(E)$ acts as a kind of likelihood which updates the probability $P(E)$ to $P^*(E)^{1-n} = \nu^*(E)P(E)^{1-n}$. The Nu model thus verifies the external Bayesianity condition. Since we are in the binary case, $O_G(\cdot)$ must satisfy $O_G(E).O_G(\bar{E}) = 1$, which implies that $\nu(E).\nu(\bar{E}) = 1$, i.e., $\nu(E)$ are odds.

Proposition 6. *For the binary case $\mathcal{E} = \{E, \bar{E}\}$, the Nu model is equivalent to*

i) a maximum entropy pooling formula updated by the odds $(\nu(E), 1/\nu(E))$.

ii) a generalized logarithmic pooling formula with $w_i = 1$, for $i = 1, \dots, n$

The maximum entropy formula corresponds to (28) with $\nu(E) = 1$ for all $E \in \mathcal{E}$. Conditional independence (20) is a sufficient condition for this, but in theory it is not necessary.

If $\nu(E)$ is close to a constant c for all E , the maximum entropy pooling formula (18) is an excellent approximation of (28). Note that in (29) the particular status of $\nu(E)$ as compared to $P_0(E)$ is a little bit unclear.

General case

In the the general case with $K > 2$ possible outcomes in \mathcal{E} , equations (28) and (29) are not equivalent. Two routes are possible for generalizing the Nu model.

1. **The first route** (Nu-1) consists in generalizing the pooling of the probabilities, as in (28), thus obtaining a generalized or updated maximum entropy formula. Would the full joint probability be accessible, the quantities $\nu^*(E)$ could be exactly computed. This not being the case, $\nu^*(E)$, if not set equal to 1, acts as a kind of likelihood, as already seen in the binary case.
2. **The second route** (Nu-2) considered in Polyakova and Journal (2007) consists in generalizing the pooling of the odds, as in (29), thus leading to

$$P_G(E) \propto O_G(E)/(1 + O_G(E)), \quad O_G(E) = \frac{O(E)^{1-n}\nu(E) \prod_{i=1}^n O_i(E)}{\sum_{E \in \mathcal{E}} O(E)^{1-n}\nu(E) \prod_{i=1}^n O_i(E)}. \quad (30)$$

In this second route, $\nu(E)$ acts as an odd updating the product of odds. Increasing $\nu(E)$ leads to an increase of the probability $P_G(E)$.

It is important to stress that, when not in the binary case, these two routes will lead to different values of the aggregated probability $P_G(E)$ for given values of $\nu(E)$. This is illustrated in Table 1, in which $P_G(E)$ is computed according to the Nu-1 or Nu-2 representation for several values of $\nu(E)$. Note that since $w_1 + w_2 = 2 > 1$, the aggregated probability will always be further away from the prior P_0 than the probabilities P_i (see proposition 6 *ii*). Hence, for all considered cases, P_G is the highest for E_3 . One can also see that when $\nu(E)$ is evenly distributed, the value of $\nu(E)$ does not play any role when following the first route, which can be seen from (28), while it does play an role when following the second route.

Table 1: Aggregated probability computed according to the two possible generalization of the Nu model.

		E_1	E_2	E_3
	P_0	0.6	0.3	0.1
	P_1	1/3	1/3	1/3
	P_2	0.6	0.15	0.25
$(\nu(E_1), \nu(E_2), \nu(E_3))$		P_G		
(1, 1, 1)	Nu-1	0.250	0.125	0.625
(1, 1, 1)	Nu-2	0.302	0.155	0.543
(2, 2, 2)	Nu-1	0.250	0.125	0.625
(2, 2, 2)	Nu-2	0.324	0.189	0.487
(1, 2, 3)	Nu-1	0.105	0.105	0.790
(1, 2, 3)	Nu-2	0.231	0.202	0.567
(0.28, 0.68, 8)	Nu-2	0.105	0.105	0.790

These results illustrate the fact that the first route corresponds to the external Bayesianity condition, with $\nu(E)$ playing the role of an external likelihood. When $\nu(E)$ is uneven, higher values of $\nu(E)$ yield to larger aggregated probabilities. For a given vector for $\nu(E)$, the first route ($\nu(E)$ multiplying probabilities) leads to more extreme probabilities, while the second route ($\nu(E)$ multiplying odds) leads to more equilibrated probabilities. It is however possible to find a vector of values along the second route leading to approximately the same aggregated probabilities.

It is also important to understand the profound difference between Bordley/Tau and Nu aggregations. While in the former there is for each source of information a single parameter w_i independent on the event E , in the latter there is a one parameter per event E without any mention to the source of information.

4.4 Multiplication methods at a glance

As seen in the previous sections, methods based on the multiplication of probabilities or multiplication of odds are intimately related. Presenting all methods in Table 2 makes it possible to grasp the relationships between the multiplication methods in one glance (table

2). At the first level we make a distinction between the binary case and the general case. We re-emphasize that most of the literature is concerned with the binary case, either explicitly or implicitly, for which methods based on odds are equivalent to methods based on probabilities. On the contrary, it is important to distinguish these two cases when dealing with non binary events.

A general formulation of all pooling methods is possible:

$$T_G(E) = Z + U(E) + \left(1 - \sum_{i=1}^n w_i\right) T_0(E) + \sum_{i=1}^n w_i T_i(E), \quad (31)$$

in which T is related to probabilities in the following way: $T \equiv P$ for all linear pooling methods; $T \equiv \ln P$ for methods based on the product of probabilities, and $T \equiv \ln O = \ln P - \ln(1 - P)$ for methods based on the product of odds. $U(E)$ is an updating likelihood when considering the general log-linear pooling; it is the logarithm of the Nu parameter for the Nu model. $T_0(E)$ is the prior probability and Z is a normalizing constant. The weight w_0 has been set equal to $1 - \sum_{i=1}^n w_i$ in order to respect external Bayesianity. Note that $w_i = 1$ for the Nu model and the maximum entropy. When $T \equiv P$, the Beta transformed model can also be included by transforming the right-hand-side of (31) with the Beta cumulative probability function $H_{\alpha,\beta}$.

Table 2: General presentation of non linear aggregation methods.

Weights	Likelihood	2 alternatives	> 2 alternatives	
		Probs \equiv Odds	Probabilities	Odds
When $\sum_{i=1}^n w_i = 1$, Ext. Bayesianity and Unanimity are verified	$\nu(E) = 1$	Log-linear = Bordley = Tau model	Log-linear	Tau model
	$\nu(E) \neq 1$	Gen. log-linear	Gen. log-linear	—
all $w_i = 1$	$\nu(E) = 1$	Cond. Indep. = Max. Entropy	Cond. Indep. \equiv Max. Entropy	—
	$\nu(E) \neq 1$	Nu model	Nu-1 \equiv updated Max. Ent.	Nu-2 (Polyakova and Journal, 2007)

5 Choosing a pooling formula, estimating the weights and assessing the forecast

5.1 Introduction

Table 3 recapitulates the previous sections about the aggregation methods and their properties. A first dichotomy is between methods based on addition and those based on multiplication. BLP is intermediate. Unlike linear pooling it is not convex and does not verify marginalization, but unlike multiplicative methods it does not verify the 0/1 forcing property. This last property is verified by all multiplicative methods. External Bayesianity is verified by the generalized log-linear model, the Nu model and the Bordley formula for binary events. In the more general case, it is only verified by the first route generalizing the Nu model.

The role of the prior deserves some discussion. All aggregation formula allow to take into account some form of prior probability, which could for example represent non stationary proportions. As it can be seen in equation (31), in multiplicative methods the role of prior is multiplicative. More precisely, since ratios P_i/P_0 are aggregated, these methods can be very sensitive to the specification of the prior. In the Bordley formula, the influence of the prior depends on the sum $S_{\mathbf{w}} = \sum_{i=1}^n w_i$. When $S_{\mathbf{w}} = 1$, the prior is filtered out. When $S_{\mathbf{w}} > 1$ the aggregated probability P_G will be further away from P_0 than the P_i s. Contrarily, if $S_{\mathbf{w}} < 1$, P_G will be closer from P_0 than the P_i s. Since Maximum entropy is a model with $S_{\mathbf{w}} = n$, we can expect this method to greatly amplify the departure to the prior.

At the exception of the maximum entropy approach which is parameter free, all methods presented above have some parameters that need either to be estimated or set by the user. In the Nu model there are $K - 1$ parameters, where K is the cardinality of \mathcal{E} , while for the log-linear formula and the Bordley/Tau model there are n parameters. The most general model is the generalized log-linear, with $K + n - 1$ parameters if not imposing external Bayesianity.

In theory, if the full probability model was known, expressions for the parameters would be accessible. But in this case, the conditional probability would also be accessible, and a

	Lin.	BLP $((\alpha, \beta) \neq (1, 1))$	ME	$\nu(1)$	$\nu(2)$ $K > 2$	Bordley	Gen. Log-lin
convexity	yes	no	no	no	no	no	no
marginalization	yes	no	no	no	no	no	no
0/1 forcing	no	no	yes	yes	yes	yes	yes
ext. Bayes.	no	no	no	yes	no	no*	no*
# of param.	$n - 1$	$n + 1^\dagger$	0	$K - 1$	$K - 1$	n	$n + K - 1$

Table 3: Recapitulation of the properties of methods for aggregating n sources of information and a prior term when there are K alternatives. Note that some properties not verified in the general case are verified for some very specific values, which either reduce the method to a different method or to dictatorship. The “no*” are “yes” when $K = 2$. [†]Number of parameters in BLP is n if we impose $\alpha = \beta$.

pooling formula would not be sought in the first place.

In the context of aggregating expert opinion, Winkler (1968) suggests four ways of assessing the weights for the linear pool, which could also be applied to the other methods:

- i) equal weights;
- ii) weights proportional to a ranking based on expert’s advice;
- iii) weights proportional to a self-rating (each source of information provide a rank for itself);
- iv) weights based on some comparison of previously assessed distributions with actual outcomes.

Setting equal weights is sometimes relevant when there is no element which allows to prefer one source of information to another, or when symmetry of information justifies it. But even in this case the sum $S_{\mathbf{w}}$ needs to set or estimated. Suggestions ii) and iii) might be relevant in the context of human judgments, but of no great use in a geoscience context.

When training data are available (case iv) it is possible to estimate the optimum weights according to the optimization of some criterion. Heskes (1998) proposed an algorithm based

on the minimization of a Kullback-Leibler distance for selecting weighting factors in logarithmic opinion pools. The optimal weights are found by solving a quadratic programming problem. Ranjan and Gneiting (2010) minimized the likelihood for finding the optimal shape parameters for the Beta transformed linear opinion pool. Cao et al. (2009) used Ordinary kriging to estimate the parameters of the Tau model, but the concept of “distance” between source of information and that of variogram of probabilities is not at all obvious.

In the next section, we provide some details about a likelihood approach for estimating the parameters for methods based on the multiplication of probabilities. A similar derivation for the linear opinion pool and its Beta transform can be found Ranjan and Gneiting (2010).

5.2 Scoring rules and divergence

The aggregated probability distribution $P_G(E)$ must be as close as possible to the (unknown) conditional probability distribution $P(E | D_1, \dots, D_n)$, $E \in \mathcal{E}$. Scoring rules (Gneiting and Raftery, 2007) provide summary measures for the evaluation of the aggregated probability distributions, by assigning a numerical value, a *score*, based on P_G and on the event that materializes. Specifically, a scoring rule is a function that associate a value $S(P_G, E_k) \in (-\infty, \infty)$ for each event E_k in \mathcal{E} , when the forecasting probability distribution is P_G . $S(P_G, P)$ will denote the expected value of $S(P_G, E_k)$ under the true probability distribution P , i.e. $S(P_G, P) = \sum_{E_k \in \mathcal{E}} S(P_G, E_k)P(E_k)$. In the following, we will only consider *strictly proper scoring rules*, for which $S(P, P) \geq S(Q, P)$ for all probability distribution Q , where equality holds if and only if $Q = P$. In words, the highest score is achieved when the aggregated probability distribution is equal to the true distribution. Under mild conditions, if S is a proper scoring rule,

$$d(Q, P) = S(P, P) - S(Q, P)$$

is the associated *divergence function*. It is non-negative and it is equal to 0 if and only if $Q = P$. Note that the order plays an important role in the definition of the divergence, which

is thus not necessarily symmetrical. Gneiting and Raftery (2007) review some of the most important scoring rules for categorical variables. We mention two scoring rules which will be important for us in the rest of this work.

Definition 9 (Quadratic or Brier score). *The quadratic or Brier score (Brier, 1950), is defined by*

$$S(P, E_k) = - \sum_{j=1}^K (\delta_{jk} - p_j)^2 \quad (32)$$

where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. The associated divergence is the squared Euclidean distance, $d(Q, P) = \sum_{k=1}^K (p_k - q_k)^2$. In this particular case, the divergence is symmetrical (and hence is a distance).

Definition 10 (Logarithmic score). *The logarithmic score corresponds to*

$$S(P, E_k) = \ln p_k \quad (33)$$

The associated divergence is the Kullback-Leibler divergence, $d(Q, P) = \sum_{k=1}^K q_k \ln(p_k/q_k)$. The highest achievable score is $S(P, P) = \sum_{k=1}^K p_k \ln(p_k)$, which is nothing but the entropy of the distribution P .

Scoring rules can be used for estimating the parameters of a pooling operator according to the following general approach. Consider a pooling operator $P_{G,\theta}$ depending on some parameters θ and a proper scoring rule, tailored to the problem considered. The estimator $\hat{\theta} = \arg \max_{\theta} S(\theta)$, where $S(\theta)$ is the empirical score built from the data set, is the optimum score estimator. Maximum likelihood estimation is a special case of optimum score estimation. The logarithmic score is related to the maximum likelihood estimation of the parameters of the pooling formulas presented in the next section. The Brier score is closely related to the calibration and the sharpness of P_G presented in the section after next.

5.3 Likelihood for log-linear pooling formulas

We now describe the maximum likelihood approach for estimating the parameters for pooling the formula based on the product of probabilities, which is recalled in its most general form:

$$P_G(E_k) = \frac{\nu(E_k)P_0(E_k)^{1-\sum_{i=1}^n w_i} P_i(E_k)^{w_i}}{\sum_{k=1}^K \nu(E_k)P_0(E_k)^{1-\sum_{i=1}^n w_i} P_i(E_k)^{w_i}}. \quad (34)$$

This model includes the log-linear model, when all $\nu(E_k) = 1$ and the Nu model (route 1), when all $w_i = 1$. In the binary case it also includes all pooling operators based on the product of odds.

The setting is the following. We denote $\mathbf{w} = (w_1, \dots, w_n)$ and $\boldsymbol{\nu} = (\nu(E_1), \dots, \nu(E_K))$ the parameters of the pooling formula and consider M repetitions of a random experiment. For each experiment $m = 1, \dots, M$, the information $D_i^{(m)}$ is available, allowing to compute the individual conditional probabilities $P_i^{(m)}(E_k)$, and to estimate the aggregated probabilities $P_G^{(m)}(E_k)$ of occurrence of any event E_k . For the sake of lighter notations, we will denote $P_{i,k}^{(m)} = P_i^{(m)}(E_k)$, $P_{G,k}^{(m)} = P_G^{(m)}(E_k)$. In addition to the input information, we also have access to the outcome of the experiments, i.e. the real occurrence of one of the various possible outcomes. We denote it $Y_k^{(m)}$, $Y_k^{(m)} = 1$ if the outcome is E_k and $Y_k^{(m)} = 0$ otherwise. In the same spirit, we will further denote $\nu_k = \nu(E_k)$. The full log-likelihood is:

$$L(\mathbf{w}, \boldsymbol{\nu}) = \ln \prod_{m=1}^M \prod_{k=1}^K (P_{G,k}^{(m)})^{Y_k^{(m)}} = \sum_{m=1}^M \sum_{k=1}^K Y_k^{(m)} \ln P_{G,k}^{(m)}. \quad (35)$$

Notice that the log-likelihood is nothing but the empirical score of the data-set when applying the logarithmic scoring rule. Replacing $P_{G,k}^{(m)}$ in (35) by its expression yields (34)

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\nu}) &= \sum_{m=1}^M \sum_{k=1}^K Y_k^{(m)} \left\{ \ln \nu_k + \left(1 - \sum_{i=1}^n w_i\right) \ln P_{0,k} + \sum_{i=1}^n w_i \ln P_{i,k}^{(m)} \right\} \\ &\quad - \sum_{m=1}^M \ln \left\{ \sum_{k=1}^K \nu_k P_{0,k}^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n (P_{i,k}^{(m)})^{w_i} \right\}. \end{aligned} \quad (36)$$

The parameters $\hat{\mathbf{w}}$ and $\hat{\boldsymbol{\nu}}$ maximizing the log-likelihood (36) are the maximum likelihood (ML) estimators of \mathbf{w} and $\boldsymbol{\nu}$. They are found by numerical methods.

The equations obtained when equating the derivatives of (36) with respect to the parameters ν_k to zero:

$$\sum_{m=1}^M Y_k^{(m)} (\nu_k)^{-1} - \sum_{m=1}^M P_{0,k}^{1-\sum_{i=1}^n w_i} \prod_{i=1}^n (P_{i,k}^{(m)})^{w_i} / \left(\sum_{l=1}^K \nu_l P_{0,l}^{1-\sum_{i=1}^n w_i} (P_{i,l}^{(m)})^{w_i} \right) = 0, \quad (37)$$

provide an interpretation of the ML estimators. Denoting $M_k = \sum_{m=1}^M Y_k^{(m)}$ the number of occurrence of E_k , and recognizing that the second term is nothing but the probability (34), Equation (37) can be better written

$$M_k = \sum_{m=1}^M P_{G,k}^{(m)}, \quad (38)$$

for all $k = 1, \dots, K$.

Likewise, setting the derivatives with respect to w_i to zero leads after some simplifications to

$$\sum_{m=1}^M \sum_{k=1}^K Y_k^{(m)} \ln P_{i,k}^{(m)} = \sum_{m=1}^M \sum_{k=1}^K P_{G,k}^{(m)} \ln P_{i,k}^{(m)}. \quad (39)$$

In words, the parameters maximizing the log-likelihood are such that: i) the observed number of occurrence of E_k is equal to the sum of the probabilities of E_k computed with the ML parameter estimates, ie. the aggregated probabilities are globally calibrated; ii) the sum of the observed log-probabilities is equal to the sum of the aggregated log-probabilities computed with the ML parameter estimates. Note that for the Nu model, only the k conditions (38) must be verified, since the w_i s are set to 1.

In theory, it is possible to follow a similar approach for the pooling formulas based on the multiplication of Odds, but the expressions are lengthy, without bringing new insight. They are not shown here.

When fitting models, adding parameters leads to increased values of the log-likelihood. But doing so may lead to over-fitting. The Bayesian Information Criterion (BIC) introduced in Schwartz (1978) resolves this problem by adding a penalty term for the number of

parameters in the model:

$$BIC = -2L + J \ln M, \quad (40)$$

where L is the log-likelihood, J the total number of parameters of the model considered and M the number of repetitions. Given any two estimated models, the model with the lower value of BIC is the one to be preferred. Lower BIC implies either fewer explanatory variables, better fit, or both. The models being compared need not be nested, unlike the case when models are being compared using an F or likelihood ratio test.

5.4 Calibration and sharpness

Calibration and sharpness are two particular aspects of the pooling operators which can be used to evaluate their quality. We will follow Ranjan and Gneiting (2010) for a brief introduction to these notions. We need the following set-up. One considers a random experiment, leading to random information D_1, \dots, D_n and thus random probabilities P_i . It is convenient to introduce (Y_1, \dots, Y_K) the random vector corresponding to the outcome, in which $Y_k = 1$ if the outcome is E_k and $Y_k = 0$ otherwise, i.e. $P(Y_k = 1) = P(E_k) = E[Y_k]$.

Definition 11 (Calibration). *The aggregated probability $P_G(E)$ is said to be calibrated if*

$$P(Y_k | P_G(E_k)) = P_G(E_k), \quad k = 1, \dots, K. \quad (41)$$

This definition is in accordance with economic, meteorological and statistical forecasting literature (Ranjan and Gneiting, 2010). Ranjan and Gneiting (2010) proved that linear opinion pools lack calibration, even though all conditional probabilities $P(E_k | D_i)$ are calibrated.

Sharpness refers to the concentration of the aggregated distribution. The more concentrated $P_G(\cdot)$ is, the sharper it is.

Calibration and sharpness of the pooling formulas will be assessed on simulations. They arise naturally considering the Brier score. The empirical mean Brier score is defined as

$$BS = \frac{1}{M} \left\{ \sum_{k=1}^K \sum_{m=1}^M (P_G^{(m)}(E_k) - Y_k^{(m)})^2 \right\}, \quad (42)$$

where the superscript refers to the m th random experiment.

Suppose that the probability $P_G(E_k)$ takes discrete values $f_k(j)$ (e.g., from 0 to 1 by step of 0.01), where $j = 1, \dots, J$. Let $n(j)$ be the number of times $P_G(E_k) = f_k(j)$ and let $q_k(j)$ be the empirical event frequency for E_k when $P_G(E_k) = f_k(j)$. If the pooling formula is calibrated, one must have $q_k(i) = P(E_k | P_G(E_k) = f_k(i)) = f_k(i)$. Reliability diagrams plot the empirical event frequency against the aggregated probabilities (Bröcker and Smith, 2007). Significant deviation from the diagonal must be interpreted as a lack of calibration.

The Brier score can be decomposed in the following way:

$$BS = \sum_{k=1}^K \left\{ \frac{1}{M} \sum_{j=1}^J n_k(j) (f_k(j) - q_k(j))^2 \right\} - \sum_{k=1}^K \left\{ \frac{1}{M} \sum_{j=1}^J n_k(j) (q_k(j) - \bar{q}_k)^2 \right\} + \sum_{k=1}^K \bar{q}_k (1 - \bar{q}_k), \quad (43)$$

where $\bar{q}_k = \frac{1}{M} \sum_{m=1}^M Y_k^{(m)}$ is the marginal event frequency.

The first term of the decomposition is the reliability term. It corresponds to the calibration. The lower is this term, the better the pooling formula is calibrated. The second term is a deviation around the recalibrated probability. For a calibrated pooling formula, it corresponds to the sharpness; in this case the higher the sharpness, the better. The last term depends on the observation alone; it is independent on the pooling formula. To address the performance of the aggregation methods, Ranjan and Gneiting (2010) proposed diagnostics based on the paradigm of maximizing the sharpness, subject to calibration. With this paradigm, optimal weights can be found using other scoring rules, such as the logarithmic scoring rule.

6 Simulation study

We now conduct some simulation studies in order to compare the features of the different aggregation methods. We will consider three cases. In the first case, we consider the aggregation of independent information for the prediction of the binary outcome. In this case, maximum entropy (equivalent to conditional independence) should perform reasonably well.

In the second case, also concerned with the prediction of a binary outcome, we will consider a truncated gaussian model with correlation between the sources of information to be aggregated. In the third case, we consider a pluri-gaussian model in which there are three possible categories. In these examples, since we have access to the analytical expressions of all conditional probabilities, we will be able to compare the performances of the different aggregation methods with the conditional probability. For doing so, we will use the Brier scores (Equation (43)), BIC (Equation (40)) and the reliability plots presented in section 5.

6.1 Binary case: independent sources of information

For this first example, we adopt the same analytical setting as in Ranjan and Gneiting (2010), in which the Beta-transformed Linear pooling is shown to be superior to linear poolings. The sources of information are two independent $(0, 1)$ Gaussian random variables D_1 and D_2 . Let Φ denotes the standard normal cumulative distribution function and define $p = \Phi(D_1 + D_2)$. Suppose Y is a Bernoulli random variable with success probability p , and consider the event $E = \{Y = 1\}$. Then,

$$P(E | p) = P(Y = 1 | p) = E[Y | p] = p, \quad (44)$$

and

$$P_1(E) = P(E | D_1) = E[Y | D_1] = E[\Phi(D_1 + D_2 | D_1)] = \Phi(D_1/\sqrt{3}) = P_2(E). \quad (45)$$

Note that $P(E)$, $P_1(E)$ and $P_2(E)$ are naturally calibrated. A training sample of size $M = 10,000$ is generated by simulating D_1, D_2 and Y . The prior is the constant value $p_0 = E[p] = 1/2$. Table 4 presents the log-likelihood, the BIC and the Brier scores with their reliability and sharpness component for different pooling formula. The log-likelihood is computed according to

$$L = \sum_{m=1}^M Y^{(m)} \ln P_G^{(m)}(E) + (1 - Y^{(m)}) \ln(1 - P_G^{(m)}(E)).$$

For the sake of comparison, it is also computed for $P_G(E) = P_1(E)$ and $P_G(E) = P_{12}(E) = P(E | D_1, D_2)$. The model with the lowest Brier score, or with the lowest BIC should be

preferred. In the case of binary events, remember that the Bordley formula and the Tau model are equivalent, and that the Nu model is the generalized log-linear model with weights $w_i = 1$ for all $i = 1, \dots, n$. Optimal weights were obtained with the maximum likelihood approach described in the previous section, with the additional constraints of equality $w_1 = w_2$ to account for the symmetry between D_1 and D_2 . For the same reason, for the BLP parameters, we imposed $\alpha = \beta$. From Table 4, one can see that although P_1 being calibrated, it lacks sharpness. The exact conditional probability P_{12} has the lowest log-likelihood, the lowest Brier score and the highest sharpness. Linear pooling has a lower Brier score than a single information, but at the price of a loss of calibration, and it lacks sharpness. It has the highest BIC among all models considered. As expected from Ranjan and Gneiting (2010), BLP is well calibrated with a high sharpness and the BIC decreases dramatically. Note that the parameter α is quite high, indicating that a strongly bimodal Beta density is necessary to calibrate the linear pooling. Among the multiplicative pooling formula, Maximum Entropy performs surprisingly well considering that it is parameter free. This result is explained by the fact that D_1 and D_2 are drawn independently. Introducing one parameter in the pooling formula, either for the Nu model or for the Bordley formula decreases the Brier score and the log-likelihood when they are estimated using maximum likelihood, while they can increase when the parameters are away from their optimal values (results not shown here). The Bordley formula leads to better scores than the Nu model. Finally, note that log-likelihoods and Brier scores generally decrease or increase together. The lowest BIC is obtained for the Bordley formula, indicating that the extra parameter in the Generalized log-linear formula leads to some over-fitting.

6.2 Truncated Gaussian model with three sources of information

As a second case, we consider a truncated model, in the same spirit as the construction described in Chugunova and Hu (2008). The prediction point s_0 is located at the origin. The location of the three information point are defined by their distances (d_1, d_2, d_3) and

	Weight	Shape	– Loglik	BIC	BS	REL	SH
P_1	—	—	5751.505		0.1973	0.0011	0.0538
P_{12}	—	—	4135.747		0.1352	0.0010	0.1158
Lin	—	—	5208.667	10417.3	0.1705	0.0346	0.1141
BLP	—	7.8963	4168.667	8346.5	0.1362	0.0011	0.1148
ME	—	—	5028.669	10057.3	0.1391	0.0045	0.1154
Nu	—	-0.0108	4294.852	8598.9	0.1388	0.0043	0.1155
Bordley/Tau	1.4640	—	4139.390	8289.0	0.1353	0.0010	0.1156
GL	1.4646	-0.0140	4138.860	8296.1	0.1354	0.0008	0.1154

Table 4: Maximum likelihood and Brier scores for different pooling formulas: two sources of close to independent information.

their angles $(\theta_1, \theta_2, \theta_3)$ with the horizontal axis. We consider a random function $X(s)$ with an exponential covariance matrix; the range is set equal to 1 throughout. We define a threshold t and we are interested in the event $E = \{X(s_0) \leq t - a\}$ given the information $D_i = \{X(s_i) \leq t\}$, $i = 1, 2, 3$. Since we know the full model all conditional probabilities can be numerically computed. A total of 10,000 thresholds t are drawn according to a $(0, 1)$ Gaussian random variable, and we set $a = 1.35$. A Gaussian random vector $(X(s_i))_{i=0, \dots, 3}$ is then simulated conditionally on $X(s_i) \leq t$, for $i = 1, 2, 3$. With this setting we sample the whole range of probabilities for the event $E = \{X(s_0) \leq t - 1.35\}$, which on average will be close to 0.5. Figure 1 shows the histograms of the marginal and the conditional probabilities of E in one of the cases considered below. Clearly, the whole range of probabilities is sampled, allowing us a good calibration of the different pooling formulas.

Symmetrical information

In this experiment the three data points s_1, s_2, s_3 are located on a circle of radius equal to 1 (hence equal to the range), s_1 being on the horizontal axis. In this configuration the correlation between $X(s_0)$ and $X(s_i)$ is equal to 0.37 for all s_i . Hence we impose an equal weight to each data. Two cases were considered. In the first case, the angles between $s_{2,3}$ and s_1 is $2\pi/3$. There is thus a perfect symmetry between all information. In the second

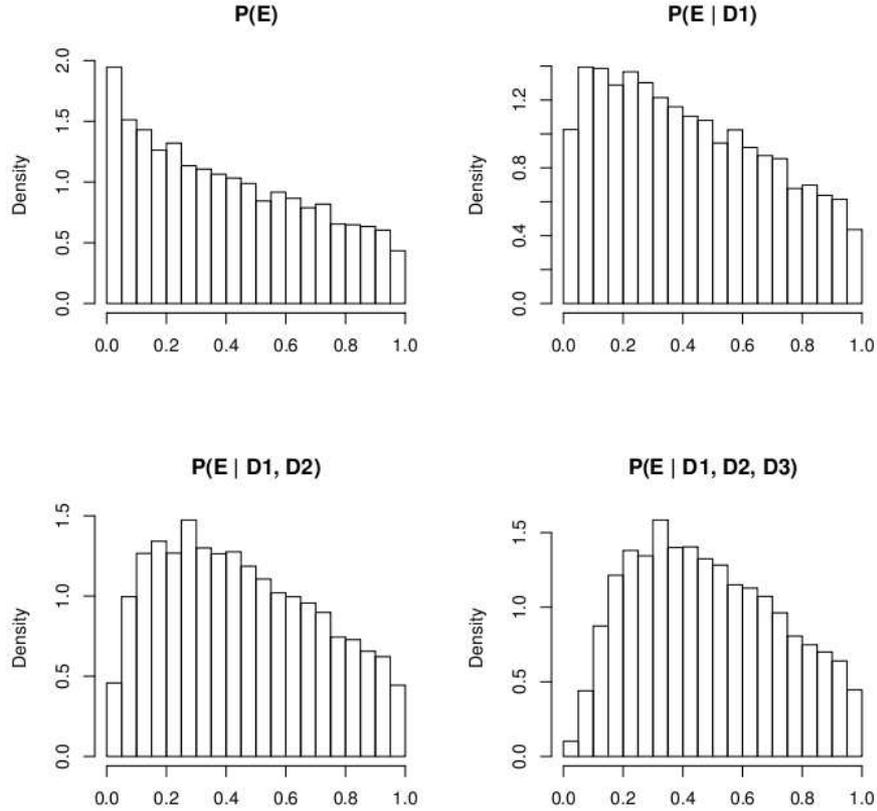


Figure 1: Histograms of $P(E)$, $P(E | D_1)$, $P(E | D_1, D_2)$ and $P(E | D_1, D_2, D_3)$.

case, the angles between $s_{2,3}$ and s_1 is $\pi/3$: the redundancy between the points is increased. Results are presented in Table 5. The Brier scores, very close to each other, are not a good criterion for assessing the quality of a pooling formula. But the log-likelihood shows a more contrasted behavior. Because of the symmetry, the linear pooling is equivalent to a single source of information. It is, by far, the poorest pooling method. The BLP formula leads to a log-likelihood which is higher than the one obtained when having access to the conditional probability with two sources of information and it is very sharp. Among the multiplicative formula, the Maximum entropy is a relatively poor pooling method; the Bordley formula performs significantly better than the Nu model. The most general model is only marginally better and its BIC is higher than the Bordley formula, which indicates over-fitting.

Angle($s_{2,3}, s_1$) = $2\pi/3$							
	Weight	Shape	- Loglik	BIC	BS	REL	SH
P_1	—	—	5873.695		0.1978	0.0024	0.0543
P_{12}	—	—	5751.289		0.1960	0.0006	0.0544
P_{123}	—	—	5724.907		0.1966	0.0012	0.0544
Lin	—	—	5873.695	11747.4	0.1978	0.0024	0.0543
BLP	—	0.6328	5774.155	11557.5	0.1958	0.0006	0.0545
ME	—	—	5787.132	11574.3	0.1999	0.0038	0.0537
Nu	—	-0.0744	5764.451	11538.1	0.1978	0.0018	0.0538
Bordley	0.7706	—	5726.661	11462.5	0.1961	0.0004	0.0545
GL	0.7470	0.0173	5726.060	11470.5	0.1965	0.0004	0.0542
Angle($s_{2,3}, s_1$) = $\pi/3$							
	Weight	Shape	-Loglik	BIC	BS	REL	SH
P_1	—	—	5782.200		0.1943	0.0019	0.0573
P_{12}	—	—	5686.825		0.1929	0.0006	0.0574
P_{123}	—	—	5650.034		0.1935	0.0007	0.0569
Lin	—	—	5782.200	11564.4	0.1943	0.0019	0.0573
BLP	—	0.6681	5704.759	11418.7	0.1932	0.0006	0.0570
ME	—	—	5720.108	11440.2	0.1974	0.0042	0.0564
Nu	—	-0.0769	5695.921	11391.8	0.1952	0.0021	0.0566
Bordley	0.7506	—	5651.377	11312.0	0.1931	0.0006	0.0571
GL	0.7126	0.02674	5649.963	11318.35	0.1937	0.0008	0.0568

Table 5: Maximum likelihood and Brier scores for different pooling formulas; three symmetrical sources of information.

Uneven information

In this third situation, the three points are at distances $(d_1, d_2, d_3) = (0.8, 1, 1.2)$. The distances being different, we will consider different weights for the three sources of information. For comparison purpose we will also include equal weights solutions. Here again the Brier score does not vary a lot; the log-likelihood shows larger variations. The method with the best indicators related to the Brier score is the BLP. Interestingly the optimal solution consists in having a 100% weight for the closest source of information and null weights for all others. It is also the case for the Bordley formula. When equal weights are imposed in the Bordley formula, the Brier score and the log-likelihood remain almost identical; but because

the number of free parameters decreases, the BIC reaches a minimum. This model achieves the best compromise between the adequation (as measured by the logarithmic score) and parsimony. The calibration curve of four pooling formulas are shown in Figure 2. On these plots, deviation from the first diagonal indicates a lack of calibration. It is visible for the linear pooling and very clear for the maximum entropy, for which there is a clear lack of predicted low probabilities.

	Weights	Shape	−Loglik	BIC	BS	REL	SH
P_1	—	—	5786.6		0.1943	0.0022	0.0575
P_{12}	—	—	5730.8		0.1927	0.0007	0.0577
P_{123}	—	—	5641.4		0.1928	0.0009	0.0579
Lin.eq	(1/3, 1/3, 1/3)	—	5757.2	11514.4	0.1940	0.0018	0.0575
Lin	(0, 0, 1)	—	5727.2	11482.0	0.1935	0.0015	0.0577
BLP	(0, 0, 1)	0.6633	5680.5	11397.8	0.1921	0.0004	0.0580
ME	—	—	5727.7	11455.4	0.1972	0.0046	0.0571
Nu	—	0.923	5791.4	11592.0	0.1950	0.0023	0.0570
Bordley.eq	(0.724, 0.724, 0.724)	—	5646.1	11301.4	0.1928	0.0006	0.0576
Bordley	(0, 0, 1.873)	—	5645.3	11318.3	0.1928	0.0007	0.0576
GL	(0.001, 0.530, 1.281)	1.039	5643.1	11323.0	0.1930	0.0010	0.0576

Table 6: Maximum likelihood and Brier scores for different pooling formulas; three symmetrical sources of information.

In summary

From these examples, one can draw the following partial conclusions

- Logarithmic scores (log-likelihood) and quadratic scores (Brier scores) can be used, alone or together, to estimate parameters and select a model. They usually lead to consistent model selection rules.
- Linear pooling is by far the poorest method. When there is a symmetry in information it is equivalent to selecting one source of information; otherwise selecting only the closest (i.e. the most correlated) information is always better than any other linear pooling.
- BLP is a real improvement of linear pooling. It achieves calibration, the log-likelihood

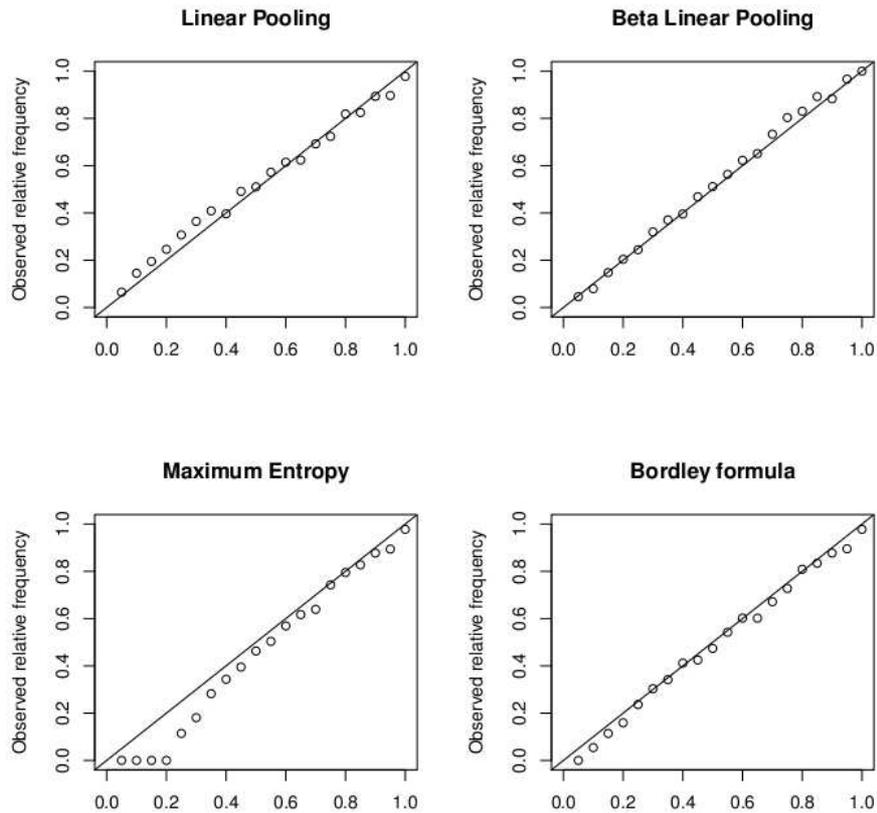


Figure 2: Calibration curve when information is uneven. BLP and Bordley formula are computed with their optimal parameters. Deviation from first diagonal indicates lack of calibration.

and the Brier score are dramatically reduced.

- Multiplicative pooling formulas lead to better scores than additive ones.
- Bordley/Tau formula shows a better scoring and a better calibration than the Nu model.
- Among all methods considered, the Bordley formula leads generally to the lowest value of the BIC. When there is a symmetry in the information (e.g., equal distance to the data points), it is an interesting option to impose equal weight.
- The generalized log-linear pooling, which is the most general model with the largest number of parameters is only marginally better than the Bordley formula which should

therefore be preferred to avoid over-fitting.

6.3 Trinary events

We keep the same geometrical configuration as above. The trinary events are defined by means of two independent $(0, 1)$ Gaussian random functions $U(s)$ and $V(s)$. The category $C(s)$ will depend on $U(s)$ and $V(s)$ according to the following scheme

$$I(s) = 1 \quad \text{if } U(s) \leq t \text{ and } V(s) \leq t \quad (46)$$

$$I(s) = 2 \quad \text{if } U(s) > t \text{ and } U(s) > V(s) \quad (47)$$

$$I(s) = 3 \quad \text{if } V(s) > t \text{ and } U(s) \leq V(s) \quad (48)$$

where t is a threshold. The marginal probabilities are the following. First, $P(1 = k) = P(U \leq t)P(V \leq t) = G^2(t)$ where $G(\cdot)$ denotes the cdf of a $(0,1)$ Gaussian random variable, and $g(t)$ its density. Then, symmetry imposes $P(I = 2) = P(I = 3)$, which leads to $P(I = 2) = 0.5[1 - P(I = 1)] = 0.5[1 - G^2(t)]$. The conditional probabilities $P(I(s_0) | I(s_i))$ are detailed in the Appendix.

The thresholds t are drawn such that the probability of category 1 is uniformly sampled between 0.1 and 0.8. A total of 20,000 random samples were drawn. It should be remembered that for trinary experiments, the equivalence between methods based on the product of probabilities and those based on the product of odds is lost. It is thus necessary to distinguish between these methods. The Nu-1 route corresponds to a product of probabilities updated by a likelihood on the events, while the Nu-2 route corresponds to a product of odds updated by odds. Linear methods do not perform particularly well, even the BLP. This is probably due to the fact that a single parameter α was used for all categories. A generalized version, with one parameter per category would probably lead to better performances. ME and $\nu(1)$ do not perform well at all. The reason is that they lead to probabilities very close to 0 or 1, thus strongly penalizing the scores when the prediction is wrong. Methods based on the product of probabilities tend to perform better than the corresponding ones based on the

product of odds (at the exception of the ν models). The optimal method is the Generalized log-linear pooling formula. Unlike the binary case, the extra parameters of this model, as compared to the log-linear model, offers the flexibility needed to fit to non binary outcomes. The generalized Loglinear pooling of odds is a model, not yet proposed in the literature, that combines w_i , the weights on the sources of information with the parameters $\nu(E)$. It performs slightly better than the Bordley/Tau model, but its is outperformed by the Generalized Log linear model on probabilities.

	-Loglik	BIC	BS	REL	SH
Lin	24123.8	24123.9	0.2219	0.0271	0.0262
BLP	21517.9	43045.8	0.2187	0.0218	0.0241
ME	44358.3	88716.6	0.2736	0.0780	0.0254
Nu-1	44278.0	88575.9	0.2770	0.0812	0.0253
Log-Lin.	18744.4	37518.6	0.1890	0.0025	0.0345
Gen. Log-lin.	18554.1	37157.8	0.1868	0.0004	0.0351
Bordley/Tau	18846.1	37721.9	0.1904	0.0019	0.0325
Nu-2	21732.6	43494.8	0.2242	0.0300	0.0269
Gen. Log-Odds	18733.2	37525.8	0.1896	0.0011	0.0326

Table 7: Maximum likelihood, BIC and Brier scores for the trinary case.

7 Discussion and Conclusion

We reviewed most methods proposed in the literature for aggregating probability distributions with a focus on their mathematical properties. By doing so, we were able to better understand the relationships between these methods. We were able to show that Conditional independence is equivalent to a particular maximum entropy principle. It is also equivalent to a Nu model with $\nu(E) = 1$ for all $E \in \mathcal{E}$ and to a Bordley/Tau formula with $w_i = 1$ for all source of information.

We showed that binary experiments must be distinguished from non binary ones. In the latter case, the equivalence between Bordley/Tau models (based on odds) and the Log-linear models (based on probabilities) is lost. For this case also, there are two different ways for

generalizing the Nu model. An interesting result of Table 2 is that it shows that one model has not yet been proposed in the literature. This model would combine weights w_i and $\nu(E)$ on odds. It would be a generalized Log-linear model on odds. It would be at the same time a generalization of the Tau model and a generalization of the Nu-2model.

When training data do exist, we also showed that Maximum Likelihood provides an efficient method for estimating the parameters of any chosen model. Using BIC and Brier Score is efficient for selecting the model leading to the best forecasts. Maximum Likelihood is rarely used in geoscience for estimating parameters and/or for selecting aggregation methods. We strongly recommend to use Maximum Likelihood methods.

Simulations presented here and other ones not presented here (Comunian, 2010) have shown that among methods based on multiplication, the Nu model performs generally worst than other methods. This can be explained from the equations: the parameters $\nu(E)$ acts as a likelihood on the events regardless of the information at hand, while the other methods provide a transformation of the conditional probabilities (the same for all events) which accounts for the redundancy or the interaction between information. For non-binary events, methods based on product of probabilities have been shown to perform slightly better than those based on product of odds. In conclusion, we do recommend the use of log-linear combination of probabilities, or its generalized version. If no data is available, the parameter free conditional independence (i.e., maximum entropy) is an acceptable approximation.

A striking result of this study is that linear methods should not be used for aggregating probability distribution. Methods based on product of probabilities are largely superior. This has profound implications on the practice of spatial prediction and simulation of indicator functions. It implies that the kriging paradigm based on linear combinations of bivariate probabilities and its sequential indicator simulation (SIS) counterpart should probably be replaced by a different paradigm base on the product of probabilities. Allard et al. (2011) arrived at a somehow similar conclusion. We hope that this contribution, together with those

cited in this work will help geoscientists to adopt this new paradigm.

8 Acknowledgments

Funding for P. Renard and A. Communian was mainly provided by Swiss National Science foundation (Grants PP002-106557 and PP002-124979) and the Swiss Confederation's Innovation Promotion Agency (CTI Project No. 8836.1 PFES-ES) and partially supported by the Australian Research Council and the National Water Commission.

References

- Allard, D., D'Or, D., and Froidevaux, R. (2011). An efficient maximum entropy approach for categorical variable prediction. *European Journal of Soil Science*, 62(3):381–393.
- Bacharach, M. (1979). Normal bayesian dialogues. *Journal of the American Statistical Association*.
- Benediktsson, J. and Swain, P. (1992). Consensus Theoretic Classification Methods. *IEEE Transactions on Systems Man and Cybernetics*, 22(4):688–704.
- Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, 28(10):1137–1148.
- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.
- Bröcker, J. and Smith, L. A. (2007). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661.
- Caers, J. (2006). A general algorithm for building 3D spatial laws from lower dimensional structural information. Technical report, Stanford Center for Reservoir Forecasting.

- Cao, G., Kyriakidis, P., and Goodchild, M. (2009). Prediction and simulation in categorical fields: a transition probability combination approach. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09*, pages 496–499, New York, NY, USA. ACM.
- Christakos, G. (1990). A bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22(7):763–777.
- Chugunova, T. and Hu, L. (2008). An assessment of the tau model for integrating auxiliary information. In Ortiz, J. M. and Emery, X., editors, *Geostats 2008 - VIII International Geostatistics Congress*. Mining Engineering Department, University of Chile.
- Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*.
- Clemen, R. T. and Winkler, R. L. (2007). Aggregating probability distributions. In *Advances in Decision Analysis*. Edwards, W and Miles, R.F. and von Winterfeldt, D.
- Comunian, A. (2010). *Probability aggregation methods and multiple-point statistics for 3D modeling of aquifer heterogeneity from 2D training images*. PhD thesis, University of Neuchtel, Switzerland.
- Dietrich, F. (2010). Bayesian group belief. *Social Choice and Welfare*, 35:595–626. 10.1007/s00355-010-0453-x.
- Genest, C. (1984). Pooling operators with the marginalization property. *The Canadian Journal of Statistics*.
- Genest, C. and Wagner, C. G. (1987). Further evidence against independence preservation in expert judgement syntheses. *Aequationes Mathematicae*, 32(1):74–86.

- Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Heskes, T. (1998). Selecting weighting factors in logarithmic opinion pools. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems 10*, pages 266–272, Cambridge. MIT Press.
- Journel, A. (2002). Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, 34:573–596.
- Krishnan, S. (2008). The tau model for data redundancy and information combination in earth sciences: Theory and application. *Mathematical Geosciences*, 40:705–727.
- Lehrer, K. and Wagner, C. (1983). Probability amalgamation and the independence issue: A reply to Laddaga. *Synthese*.
- Mariethoz, G., Renard, P., and Froidevaux, R. (2009). Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation. *Water Resources Research*, 45.
- Okabe, H. and Blunt, M. J. (2004). Prediction of permeability for porous media reconstructed using multiple-point statistics. *Physical Review*, 70(6).
- Okabe, H. and Blunt, M. J. (2007). Pore space reconstruction of vuggy carbonates using microtomography and multiple-point statistics. *Water Resources Research*, 43(W12S02).
- Polyakova, E. I. and Journel, A. G. (2007). The nu expression for probabilistic data integration. *Mathematical Geology*.

- Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):261–464.
- Stone, M. (1961). The opinion pool. *The Annals of Mathematical Statistics*.
- Strebelle, S., Payrazyan, K., and Caers, J. (2003). Modeling of a deepwater turbidite reservoir conditional to seismic data using principal component analysis and multiple-point geostatistics. *SPE Journal*, 8(3):227–235.
- Tarantola, A. (2005). *Inverse Problem Theory*. SIAM.
- Tarantola, A. and Valette, B. (1982). Inverse problems = quest for information. *Journal of Geophysics*, 50:159–170.
- Wagner, C. (1984). Aggregating subjective probabilities: some limitative theorems. *Notre Dame J. Formal Logic*, 25(3):233–240.
- Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science*, 15(2):B61–B75.

A Maximum entropy

Let us define $Q(E, D_0, D_1, \dots, D_n)$ the joint probability distribution maximizing its entropy $H(Q) = -\sum_{E \in \mathcal{E}} Q(D_0, D_1, \dots, D_n)(E) \ln Q(D_0, D_1, \dots, D_n)(E)$ subject to the following constraints

1. $Q(E, D_0) = Q(E | D_0)Q(D_0) \propto P_0(E)$, for all $E \in \mathcal{E}$
2. $Q(E, D_0, D_i) = Q(E | D_i)Q(D_i)Q(D_0) \propto P_i(E)$, for all $E \in \mathcal{E}$ and all $i = 1, \dots, n$.

We will first show that

$$Q(E, D_0, D_1, \dots, D_n) \propto P_0(E)^{1-n} \prod_{i=1}^n P_i(E),$$

from which the conditional probability

$$P_G(P_0, P_1, \dots, P_n) = \frac{Q^*(E, D_0, D_1, \dots, D_n)}{\sum_E Q^*(E, D_0, D_1, \dots, D_n)} = \frac{P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}{\sum_E P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}$$

is immediately derived. For ease of notation, we will use \sum_E as a short notation for $\sum_{E \in \mathcal{E}}$,

Proof The adequate approach is to use the Lagrange multiplier technique on the objective function

$$\begin{aligned} J &= - \sum_E Q(E, D_0, D_1, \dots, D_n) \ln Q(E, D_0, D_1, \dots, D_n) \\ &+ \sum_E \mu_E \{Q(E, D_0) - aP_0(E)\} + \sum_{i=1}^n \sum_E \lambda_{E,i} \{Q(E, D_0, D_i) - b_i P_i(E)\}, \end{aligned}$$

where μ_E and $\lambda_{E,i}$ are Lagrange multipliers. For finding the solution Q^* optimizing the constrained problem, we set all partial derivatives to 0. This leads to the system of equations:

$$\ln Q^*(E, D_0, D_1, \dots, D_n) = -1 + \sum_E \mu_E + \sum_E \sum_{i=1}^n \lambda_{E,i}, \quad (49)$$

$$Q^*(E, D_0) = aP_0(E), \quad (50)$$

$$Q^*(E, D_0, D_i) = b_i P_i(E), \text{ for } i = 1, \dots, n. \quad (51)$$

From (49) and (50), we get

$$Q^*(E, D_0) = e^{-1} \prod_E e^{\mu_E} \propto P_0(E).$$

Similarly, from (49) and (51), we get

$$Q^*(E, D_0, D_i) = Q^*(E, D_0) \prod_E e^{\lambda_{E,i}} \propto P_i(E), \text{ for } i = 1, \dots, n,$$

from which we find

$$\prod_E e^{\lambda_{E,i}} \propto P_i(E)/P_0(E), \text{ for } i = 1, \dots, n.$$

Plugging this in (49) yields

$$Q^*(E, D_0, D_1, \dots, D_n) \propto P_0(E) \prod_{i=1}^n \frac{P_i(E)}{P_0(E)}.$$

Hence,

$$P_G(P_0, P_1, \dots, P_n)(E) = \frac{Q^*(E, D_0, D_1, \dots, D_n)}{\sum_E Q^*(E, D_0, D_1, \dots, D_n)} = \frac{P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}{\sum_E P_0(E)^{1-n} \prod_{i=1}^n P_i(E)}.$$

B Conditional probabilities for the trinary event example

1. Let us first compute the conditional probability:

$$\begin{aligned} P(I(s') = 1 \mid I(s) = 1) &= P(U' \leq t, V' \leq t \mid U \leq t, V \leq t) \\ &= P(U' \leq t, V' \leq t, U \leq t, V \leq t) / P(U \leq t, V \leq t) \\ &= P(U' \leq t, U \leq t) P(V' \leq t, V \leq t) / [P(U \leq t) P(V \leq t)] \\ &= G_2^2(t, t; \rho) / G^2(t), \end{aligned}$$

where $G_2^2(t, t; \rho)$ is the bivariate cpf of a $(0, 1)$ bi-Gaussian random vector with correlation ρ .

For symmetry reasons, one has $P(I(s') = 2 \mid I(s) = 1) = P(I(s') = 3 \mid I(s) = 1)$, from which it follows immediately

$$P(I(s') = 2 \mid I(s) = 1) = P(I(s') = 3 \mid I(s) = 1) = 0.5[1 - G_2^2(t, t; \rho) / G^2(t)].$$

2. We consider now:

$$\begin{aligned} P(I(s') = 1 \mid I(s) = 2) &= P(I(s) = 2 \mid I(s') = 1) \frac{P(I(s') = 1)}{P(I(s) = 2)} \\ &= 0.5 \left[1 - \frac{G_2^2(t, t; \rho)}{G^2(t)} \right] \frac{G^2(t)}{0.5[1 - G^2(t)]} \\ &= \frac{G^2(t) - G_2^2(t, t; \rho)}{1 - G^2(t)}. \end{aligned}$$

3. The picture is slightly more complicated for $P(I(s') = 2 \mid I(s) = 2)$:

$$\begin{aligned}
P(I(s') = 2 \mid I(s) = 2) &= P(U' > t, U' > V', U > t, U > V)P(I(s) = 2) \\
&= 0.5[1 - G^2(t)] \int_t^{+\infty} \int_t^{+\infty} g_2(u, u'; \rho) \int_{-\infty}^{u'} \int_{-\infty}^u g_2(v, v'; \rho) dv dv' du du' \\
&= 0.5[1 - G^2(t)] \int_t^{+\infty} \int_t^{+\infty} g_2(u, u'; \rho) G_2(u, u'; \rho) du du'
\end{aligned}$$

There is no closed form expression for the double integral which must be evaluated numerically. Then, $P(I(s') = 3 \mid I(s) = 2)$ is computed as the complement to 1.

4. The Conditional probabilities of $I(s')$ given that $I(s) = 3$ are then obtained by symmetry.