On the choice of the weights for the logarithmic pooling of probability distributions

Luiz Max F. de Carvalho[a,b,c], Daniel A. M. Villela[a], Flavio Coelho[c] & Leonardo S. Bastos[a]
a – Program for Scientific Computing (PROCC), Oswaldo Cruz Foundation.
b – Institute of Evolutionary Biology, University of Edinburgh.
c – School of Applied Mathematics, Getulio Vargas Foundation (FGV).

September 22, 2016

Abstract

Combining different prior distributions is an important issue in decision theory and Bayesian inference. Logarithmic pooling is a popular method to aggregate expert opinions by using a set of weights that reflect the reliability of each information source. The resulting pooled distribution however heavily depends set of weights given to each opinion/prior. In this paper we explore three approaches to assigning weights to opinions. Two methods are stated in terms of optimisation problems and a third one uses a hierarchical prior that accounts for uncertainty on the weights. We explore several examples of interest, such as proportion and rate estimation and combining Normal distributions. Our findings...

Key-words: logarithmic pooling; expert opinion; maximum entropy; Kullback-Leibler divergence; Dirichlet prior.

Background

Combining probability distributions is a topic of general interest, both in the statistical (West, 1984; Genest et al., 1986; Genest and Zidek, 1986) and decision theory literatures (Genest et al., 1984). On the theoretical front, studying opinion pooling operators may give important insights on consensus belief formation and group decision making (Genest and Zidek, 1986). Among the various opinion pooling operators proposed in the literature, logarithmic pooling has enjoyed much popularity, mainly due to its many desirable properties such as relative propensity consistency (RPC) and external Bayesianity (EB) (Genest et al., 1986). In a practical setting, logarithmic pooling finds use in a range of fields, from infectious disease modelling (Coelho and Codeço, 2009) and wildlife conservation (Poole and Raftery, 2000) to engineering (Lind and Nowak, 1988; Savchuk and Martz, 1994).

A common situation of interest is that of combining expert opinions, represented as proper probability distributions, about a quantity of interest $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n$. To combine these opinions using logarithmic pooling requires assigning weights to each of the experts. These weights represent the reliability of each opinion (Genest et al., 1984). This requirement naturally leads to the question of how to choose the weights in a meaningful fashion, according to some well-accepted optimality criterion. There are a few proposals in the literature that build methods using different approaches. One proposal is to maximise the entropy the pooled distribution (Myung et al., 1996), whereas another one is to minimise Kullback-Leibler (KL) divergence between the pooled distribution and the individual opinions (Abbas, 2009) or between the pooled (prior) distribution and the posterior distribution (Rufo et al., 2012a,b).

These approaches, while moving away from the problem of arbitrarily assigning the weights, arrive at single point solutions, similar to point estimates in Statistical theory. Albeit acknowledging that these approaches have merit, we argue that in many settings, where one has substantial prior information on the relative reliabilities of the information sources (experts), it would be desirable to incorporate this information into the pooling procedure while accommodating uncertainty about the weights (Poole and Raftery, 2000). Moreover, assigning a probability distribution over the weights allows one to obtain a posterior distribution using a Bayesian procedure, which in turn enables learning about the weights from data. Therefore, it makes possible to sequentially update knowledge about the reliability of each expert/source in the face of new data.

In this paper we explore previous approaches for deriving the weights for logarithmic pooling, namely by maximising the entropy of the resulting distribution and minimising the KL divergence between the pooled distribution and each individual distribution. Additionally, we propose a hierarchical prior approach in which we place a distribution on the weights. We also discuss the case where the decision maker is interested in a transformation y of the quantity θ for which the expert opinions were elicited. To conclude, we present an example on proportion estimation by combining Beta priors. [EXEMPLO]

Properties of the logarithmic pooling operator

In what follows, we introduce the necessary theory and notation and motivate the use of the logarithmic pooling operator by presenting some of its desirable properties.

First let us define the logarithmic pooling (LP) operator. Let $\mathbf{F}_\theta := \{f_0(\theta), f_1(\theta), f_2(\theta), \ldots, f_K(\theta)\}$ be a set of prior distributions representing the opinions of $K+1$ experts and let $\boldsymbol{\alpha} := \{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_K\}$ be the vector of weights, such that $\alpha_i > 0 \, \forall i$ and $\sum_{i=0}^{K} \alpha_i = 1$. Then the log-pooled prior is

$$\mathcal{LP}(\mathbf{F}_\theta, \boldsymbol{\alpha}) := \pi(\theta) = t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\theta)^{\alpha_i}, \tag{1}$$

where $t(\boldsymbol{\alpha}) = \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} d\theta$.

Logarithmic pooling will only yield proper probability distributions if it is possible to normalise the expression in (1). While Poole and Raftery (2000) provide a proof for the case of two densities (see Theorem 1 therein), Genest et al. (1986) (pg.489) prove the result for a finite number of densities.

Theorem 1. ***Normalisation Genest et al. (1986).*** *Let A be the $(K+1)$-dimensional open simplex on $[0,1]$. For all $\boldsymbol{\alpha} \in A$ there exists a constant $t(\boldsymbol{\alpha})$ such that $\int_{\boldsymbol{\Theta}} \pi(\theta) d\theta = 1$.*

A simple proof using Hölder's inequality is given in the Appendix of this paper. This result ensures any (finite) number of proper distributions can be combined using the logarithmic pooling operator to yield a normalisable (proper) density.

Another interesting property of the logarithmic pooling operator is the fact that combining log-concave densities will produce a log-concave pooled distribution.

Remark 1. ***Log-concavity.*** *Let $\mathbf{F}_\theta$ be a set of log-concave distributions, i.e., each f_i can be represented as $f_i(\theta) \propto e^{\psi_i(\theta)}$, where $\psi_i(\cdot)$ is a concave function. Then $\pi(\theta)$ is also log-concave.*

Log-concavity of the pooled prior may be important to consider in order to guarantee unimodality and certain conditions on tail behaviour.

Exponential family

Suppose we are interested in a random variable, Y, from a exponential family with parameter θ and probability density function given by

$$f(y|\theta) = h(y) e^{\theta y - \psi(\theta)}. \tag{2}$$

Let $\mathbf{F}_y$ be a set of distributions on y of the form in (2), $f_i(y|\theta_i)$, $i = 0, 1, \ldots, K$. The combined (log-pooled) distribution also belongs to the exponential family:

$$\pi(y|\boldsymbol{\alpha}) = t(\boldsymbol{\alpha}) h^*(y) e^{\theta^* y - \psi^*(\tilde{\theta})}. \tag{3}$$

where $h^*(y) = \prod_{i=0}^{K} h_i(y)^{\alpha_i}$, $\theta^* = \sum_{i=0}^{K} \alpha_i \theta_i$ and $\psi^*(\tilde{\theta}) = \sum_{i=0}^{K} \alpha_i \psi_i(\theta_i)$.

The entropy function of the log-pooled distribution is

$$H_\pi(y) = -\log t(\boldsymbol{\alpha}) + \psi^*(\tilde{\theta}) - \mathbb{E}_\pi[\log h^*(y)] - \theta^* \mathbb{E}_\pi[Y], \tag{4}$$

and the Kullback-Leibler divergence between each distribution in $\mathbf{F}_y$ and $\pi(y|\boldsymbol{\alpha})$ can be written as:

$$KL(f_i||\pi) = -\psi_i(\theta_i) + \psi^*(\tilde{\theta}) + \mathbb{E}_\pi[\log h_i(y)] - \mathbb{E}_\pi[\log h^*(y)]. \tag{5}$$

These expressions allow for easy computation of information measures for a broad class of distributions, which will be useful in the remainder of this paper.

Conjugate priors

A conjugate prior family for $f(y|\theta)$ is the following

$$g(\theta|a,b) = K(a,b)e^{\theta a - b\psi(\theta)} \,, \tag{6}$$

where $K(a,b)$ is a normalising constant. Here, we employ the definition and notation for exponential family and its conjugate prior family from Robert (2001, chapter 3). Similar to the above, let $\mathbf{G}_\theta$ be a set of conjugate prior distributions representing the opinions of $K+1$ experts, and $g_i(\theta) = g(\theta|a_i, b_i)$ from equation (6).

The log-pooled prior is also a conjugate prior for $f(y|\theta)$ with hyperparameters given by an weighted mean of the experts hyperparameters, i.e., $\pi(\theta|\boldsymbol{\alpha}) = g(\theta|a^*, b^*)$. Note that the posterior distribution of θ is given by $\pi(\theta|y) = g(\theta|a^* + y, b^* + 1)$.

The entropy function of the log-pooled prior is given by

$$H_\pi(\theta) = -\log(K(a^*, b^*)) + a^* E[\theta] - b^* E[\psi(\theta)] \,. \tag{7}$$

And the Kullback-Leibler divergence, $KL(f_i||\pi)$, is the following

$$KL(g_i||\pi) = \log(K(a_i, b_i)) - \log(K(a^*, b^*)) + (a_i - a^*)E[\theta] - (b_i - b^*)E[\psi(\theta)] \,. \tag{8}$$

Combining distributions on a transformed quantity

So far we have discussed the situation where the decision maker combines distributions over θ. We now turn our attention to the case where the decision maker is interested in a quantity y, given by a deterministic transform of θ. This is a slightly different setting to that discussed by Poole and Raftery (2000), who discuss invertible and non-invertible transforms in the context of pooling distributions on the input and output of a deterministic model.

Let $\theta \in \Theta \subseteq \mathbb{R}^p$ and $y \in \mathcal{Y} \subseteq \mathbb{R}^q$. Define the transform (model) as $M : \Theta \to \mathcal{Y}$. This paper concerns the setting where a set $\mathbf{F}_\theta$ of distributions on θ is available but one is interested in obtaining a pooled distribution on y. There are two ways of obtaining a pooled distribution $\pi(y)$: (a) combine the distributions in $\mathbf{F}_\theta$ and then apply $M(\cdot)$ to the resulting distribution to get the induced distribution on y ("pool-then-induce" approach) or; (b) apply $M(\cdot)$ to each component of $\mathbf{F}_\theta$ and combine the resulting induced distributions using the logarithmic pooling operator ("induce-then-pool" approach). First we note that under some conditions on the transformation $M(\cdot)$, the order in which one performs the pooling and transforming operations does not matter. Consider $M(\theta) = y$ and suppose $M(\cdot)$ is invertible, i.e., there exists $M^{-1} : \mathcal{Y} \to \Theta$ such that,

$$A_0 : \quad M(M^{-1}(v)) = v, \; v \in \mathcal{Y} \text{ and } M^{-1}(M(t)) = t, \; t \in \Theta$$

and $J := det\left[\frac{\partial M^{-1}}{\partial y_1}, \frac{\partial M^{-1}}{\partial y_2}, \ldots, \frac{\partial M^{-1}}{\partial y_p}\right]$ is the Jacobian.

With some abuse of notation, define

$$\begin{aligned} \mathbf{F}_{\mathbf{y}}^{\mathbf{-1}} &:= \{h_1(y), h_2(y), \ldots, h_K(y)\} \\ h_i(y) &= f_i(M^{-1}(y))|J| \end{aligned} \tag{9}$$

where $|J|$ is the absolute value of J.

Remark 2. *Define $\pi_\theta(\theta|\boldsymbol{\alpha}) = \mathcal{LP}(\mathbf{F}_\theta, \boldsymbol{\alpha})$ and $\pi_y(y|\boldsymbol{\alpha}) = \pi_\theta(M^{-1}(y))|J|$. Similarly, define $\pi'_y(y|\boldsymbol{\alpha}) = \mathcal{LP}(\mathbf{F}_{\mathbf{y}}^{\mathbf{-1}}, \boldsymbol{\alpha})$. If A_0 is satisfied, then $\pi_y(y|\boldsymbol{\alpha}) = \pi'_y(y|\boldsymbol{\alpha})$, i.e., the order in which one combines and transforms the distributions does not matter.*

Proof. To see that Remark 2 holds, we only need a straightforward calculation from the definition of the logarithmic pooling operator in (1):

$$\pi_y(y|\boldsymbol{\alpha}) \propto \prod_{i=0}^{K} f_i(M^{-1}(y))^{w_i}|J|$$

whereas

$$
\begin{aligned}
\pi'_y(y|\boldsymbol{\alpha}) &\propto \prod_{i=0}^{K} \left[f_i(M^{-1}(y))|J| \right]^{w_i} = \prod_{i=0}^{K} f_i(M^{-1}(y))^{w_i} \prod_{i=0}^{K} |J|^{w_i} \\
&\propto \prod_{i=0}^{K} f_i(M^{-1}(y))^{w_i} |J|,
\end{aligned}
$$

as claimed. This is very similar in nature to the "external Bayesianity" property of the logarithmic pooling operator, by which one can either pool the prior distributions and then combine the resulting distribution with the likelihood or combine each prior with the likelihood separately and then pool the resulting posteriors to obtain the same posterior distribution. □

We note, however, that if $M(\cdot)$ is not invertible,
One can see this result

Can we prove the reverse (Equivalence→Invertibility) ?

Here go my two cents. Suppose $M(\cdot)$ is not invertible on the whole of $\mathcal{Y}$, but instead is **piecewise invertible**. Let $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_T$ be a partition of $\mathcal{Y}$, i.e., $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$, $\forall i \neq j \in \{1, 2, \ldots, T\}$ and $\bigcup_{t=1}^{T} \mathcal{Y}_t = \mathcal{Y}$. Then define the inverse functions $M_t^{-1}(y) : \mathcal{Y}_t \to \Theta$, $t \in \{1, 2, \ldots, T\}$. Last, let J_t be the Jacobian of $M_t^{-1}(\cdot)$. Then we are prepared to write:

$$
\pi_y(y|\boldsymbol{\alpha}) \propto \sum_{t=1}^{T} \left(\prod_{i=0}^{K} f_i(M_t^{-1}(y))^{w_i} \right) |J_t|, \quad \text{and} \tag{10}
$$

$$
\pi'_y(y|\boldsymbol{\alpha}) \propto \prod_{i=0}^{K} \left[\sum_{t=1}^{T} f_i(M_t^{-1}(y))|J_t| \right]^{w_i} \tag{11}
$$

which, I claim, will only be equal if $T = 1$, that is, if $M(\cdot)$ is invertible in the usual sense.

We now move on to study three approaches to assign weights. The first two approaches are based on optimality criteria and a third method is based on assigning a Dirichlet hyperprior on the weights.

Methods

Choosing the weights by maximising entropy

In a context of near complete uncertainty about the relative reliabilities of the experts (information sources) it may be desirable to combine the prior distributions such that $\pi(\theta)$ is maximally uninformative. Such approach would ensure that, given the constraints imposed by $\mathbf{F}_\theta$, the pooled distribution is the one which best represents the current state of knowledge (Jaynes, 1957; Savchuk and Martz, 1994). In order to choose $\boldsymbol{\alpha}$ so as to maximise prior diffuseness, one can maximise the entropy of the log-pooled prior:

$$
H_\pi(\theta) = E_\pi\left[-\ln \pi(\theta)\right] = -\int_{\boldsymbol{\Theta}} \pi(\theta) \ln \pi(\theta) d\theta \tag{12}
$$

In some cases it may be useful to express $H_\pi(\theta)$ as

$$
H_\pi(\theta; \boldsymbol{\alpha}) = \sum_{i=0}^{K} \alpha_i E_\pi[-\ln f_i(\theta)] - \ln t(\boldsymbol{\alpha}) \tag{13}
$$

Formally, we want to find $\hat{\boldsymbol{\alpha}}$ such that

$$
\hat{\boldsymbol{\alpha}} := \arg\max H_\pi(\theta; \boldsymbol{\alpha}) \tag{14}
$$

This approach, however, does not result in a convex optimisation problem, therefore one is not guaranteed to find a unique solution. See Proposition 3, below, for intuition as to why.

Choosing the weights by minimising Kullback-Leibler divergence

One could also want to choose the pooling weights so as to minimise the total Kullback-Leibler divergence between each proposed distribution and the pooled distribution. Let $d_i = \text{KL}(f_i||\pi)$ and let $L(\boldsymbol{\alpha})$ be a loss function such that

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=0}^{K} d_i \qquad (15)\\ &= -K \ln t(\boldsymbol{\alpha}) + \sum_{i=0}^{K}\sum_{j\neq i}^{K} \alpha_j \text{KL}(f_i||f_j) \qquad (16)\\ \hat{\boldsymbol{\alpha}} &:= \arg\min L(\boldsymbol{\alpha}) \qquad (17)\end{aligned}$$

Remark 3. ***Uniqueness*** *(Rufo et al., 2012a). The distribution obtained following (17) is unique, i.e., there is only one aggregated prior $\pi(\theta)$ that minimizes $L(\boldsymbol{\alpha})$.*

One can get some intuition into the proof of this claim by noting that minimising (16) is equivalent to maximising $\ln t(\boldsymbol{\alpha}) = \ln \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} d\theta$. Rufo et al. (2012a) show that $t(\boldsymbol{\alpha})$ is concave, therefore the problem in (17) has a unique solution. By contrast, the problem in (14) requires to minimise $\ln t(\boldsymbol{\alpha})$ hence lacking a sufficient condition for the existence of a unique solution.

Minimising Kullback-Leibler divergence in transformed space

One might argue that procedure (b) makes little sense, given that the set of opinions $\mathbf{F}_\theta$ concerns only θ, i.e, it was not necessarily constructed taking the transformation $M(\cdot)$ into account. An example is a situation where experts are asked to provide distributions on the probability p of a particular event. In general, elicitation for $f_i(p)$ will not take into account the induced distribution on the log-odds, $M(p) = \log p/(1-p)$. Nevertheless, the decision-maker may wish to assign the weights $\boldsymbol{\alpha}$ in a way that takes $M(\cdot)$ into account, e.g., by giving lower weights to experts whose distributions on the log-odds scale are unreasonable.

This decision process can be made more precise. In a similar spirit to the discussed above, one can construct $\boldsymbol{\alpha}$ so as to minimise the Kullback-Leibler divergence between each distribution in $\mathbf{F}_{\mathbf{y}}^{-1}$ and a transformation of the distribution obtained by procedure (a), $\pi_y(y|\boldsymbol{\alpha}) = \pi_\theta(M^{-1}(y)|\boldsymbol{\alpha})|J|$. Let $d_i = \text{KL}(h_i(y)||\pi_y(y|\boldsymbol{\alpha}))$. We then aim at solving the problem

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=0}^{K} d_i \qquad (18)\\ \hat{\boldsymbol{\alpha}} &:= \arg\min L(\boldsymbol{\alpha})\end{aligned}$$

This procedure therefore choses weights for each expert by how coherent the prior provided by each expert is with the pool-then-induce – procedure (a) – prior in the transformed space induced by $M(\cdot)$.

Specifying a prior distribution on the weights

In this section we propose a hierarchical prior for θ conditional on $\boldsymbol{\alpha}$ in order to incorporate uncertainty on the weights. A natural choice for a prior distribution for $\boldsymbol{\alpha}$ is the $(K+1)-$dimensional Dirichlet distribution. The conditional distribution $\pi(\theta|\boldsymbol{\alpha})$ is of the form in (1) and the prior density for $\boldsymbol{\alpha}$ is

$$\pi(\boldsymbol{\alpha}) = \frac{1}{\mathcal{B}(\boldsymbol{X})} \prod_{i=0}^{K} \alpha_i^{x_i - 1} \qquad (19)$$

where $\boldsymbol{X} = \{x_0, x_1, \dots, x_K\}$ is the vector of hyperparameters for the Dirichlet prior and $\mathcal{B}(X)$ is the multinomial beta function. The marginal prior for θ is then

$$\begin{aligned} \pi(\theta) &= \int_A \pi(\theta|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \qquad (20)\\ &= \frac{1}{\mathcal{B}(\boldsymbol{X})} \int_A t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} \alpha_i^{x_i - 1} d\boldsymbol{\alpha} \qquad (21)\end{aligned}$$

Applications

Binomial probabilities

We now turn our attention to combining expert opinions about probabilities and proportions. In this setting we are interested in the random variable $Y|\theta \sim Bernoulli(\theta)$. Assume that we want to obtain a combined prior for a proportion θ. A common choice for $\mathbf{F}(\theta)$ is the Beta family of distributions:

$$f_i(\theta; a_i, b_i) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i b_i)} \theta^{a_i - 1}(1-\theta)^{b_i - 1}$$

The log-pooled prior is then

$$\pi(\theta) = t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\theta; a_i, b_i)^{\alpha_i} \tag{22}$$

$$\propto \prod_{i=0}^{K} \left(\theta^{a_i - 1}(1-\theta)^{b_i - 1}\right)^{\alpha_i} \tag{23}$$

$$\propto \theta^{a^* - 1}(1-\theta)^{b^* - 1} \tag{24}$$

with $a^* = \sum_{i=0}^{K} \alpha_i a_i$ and $b^* = \sum_{i=0}^{K} \alpha_i b_i$. Note that (24) is the kernel of a Beta distribution with parameters a^* and b^*, thus

$$H_\pi(\theta) = \ln B(a^*, b^*) - (a^* - 1)\psi(a^*) - (b^* - 1)\psi(b^*) + (a^* + b^* - 2)\psi(a^* + b^*) \tag{25}$$

For the beta family of distributions, the KL divergence between $f_i(\theta)$ and $\pi(\theta)$ is

$$\begin{aligned} d_i = KL(f_i || \pi) = \ln\left(\frac{\mathcal{B}(a^*, b^*)}{\mathcal{B}(a_i, b_i)}\right) &+ (a_i - a^*)\psi(a_i) + (b_i - b^*)\psi(b_i) \\ &+ (a^* - a_i + b^* - b_i)\psi(a_i + b_i) \end{aligned} \tag{26}$$

The marginal prior for θ is

$$\pi(\theta) = \frac{1}{\mathcal{B}(X)} \int_{\mathcal{A}} \frac{1}{\mathcal{B}(a^*, b^*)} \theta^{a^* - 1}(1-\theta)^{b^* - 1} \alpha_i^{x_i - 1} d\boldsymbol{\alpha} \tag{27}$$

which can also be efficiently approximated through Monte Carlo sampling. We provide a simple implementation using the Stan (Stan Development Team, 2014) probabilistic programming language at `https://github.com/maxbiostat/opinion_pooling`. R code for the methods, figures and tables presented in this paper can also be found at the above link.

Here we analyse an example proposed by Savchuk and Martz (1994) (also discussed in Rufo et al. (2012b)) in which four experts are required supply prior information about the survival probability of a certain unit for which there have been $y = 9$ successes out of $n = 10$ trials. The experts express their opinion as prior means for the survival probability, which Savchuk and Martz (1994) then use to construct prior distributions with maximum variance given the restriction on the means. From the vector of prior means $\mathbf{m} = \{m_0 = 0.95, m_1 = 0.80, m_2 = 0.90, m_3 = 0.70\}$, the authors obtain the parameters of the beta distributions for each expert, $\mathbf{a} = \{a_0 = 18.10, a_1 = 3.44, a_2 = 8.32, a_3 = 1.98\}$ and $\mathbf{b} = \{b_0 = 0.955, b_1 = 0.860, b_2 = 0.924, b_3 = 0.848\}$. The resulting prior densities are show in the top panel of Figure 1. To complete the analysis, we place a diffuse $Dirichlet(\boldsymbol{\alpha}|\boldsymbol{X})$ prior on $\boldsymbol{\alpha}$ with $X_i = 1/4\, \forall i$. Finally, we propose to compare the prior distributions representing the experts' opinions as well as the combined distributions obtained by the different approaches using the integrated (marginal) likelihood (Raftery et al. (2007), eq. 9), $l(y) = \int_0^1 f(y|\theta)\pi(\theta)d\theta$. The marginal likelihood for the $i - th$ expert and J observations of the form $\{y_j, n_j\}$ is:

$$\begin{aligned} l_i(y_j) &= \int_0^1 \mathcal{L}(\theta|y_j, n_j)\pi_i(\theta)d\theta \\ &= \prod_{j=1}^{J} \frac{\Gamma(n_j - 1)}{\Gamma(n_j - y_j + 1)\Gamma(y_j + 1)} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i + b_i + n_j)} \frac{\Gamma(a_i + y_j)}{\Gamma(a_i)} \frac{\Gamma(b_i + n_j - y_j)}{\Gamma(b_i)} \end{aligned} \tag{28}$$

Poisson rate λ

Suppose we are interested in a certain count $Y \sim Poisson(\lambda)$ and $K+1$ experts are called upon to elicit prior distributions for λ. A convenient parametric choice for $\mathbf{F}(\lambda)$ is the Gamma family of distributions, for which densities are of the form

$$f_i(\lambda; a_i, b_i) = \frac{b_i^{a_i}}{\Gamma(a_i)} \lambda^{a_i - 1} e^{-b_i \lambda}$$

The log-pooled prior $\pi(\lambda)$ is then

$$\pi(\lambda) = t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\lambda; a_i, b_i)^{\alpha_i} \tag{29}$$

$$\propto \prod_{i=0}^{K} \left(\lambda^{a_i - 1} e^{-b_i \lambda}\right)^{\alpha_i} \tag{30}$$

$$\propto \lambda^{a^* - 1} e^{-b^* \lambda} \tag{31}$$

where $a^* = \sum_{i=0}^{K} \alpha_i a_i$ and $b^* = \sum_{i=0}^{K} \alpha_i b_i$. Noticing (31) is the kernel of a gamma distribution with parameters a^* and b^*, $H_\pi(\lambda)$ becomes

$$H_\pi(\lambda) = a^* - \ln b^* + \ln \Gamma(a^*) + (1 - a^*)\psi(a^*) \tag{32}$$

where $\psi(\cdot)$ is the digamma function.

The Kullback-Liebler divergence between each density and the pooled density is:

$$d_i = \mathrm{KL}(f_i || \pi) = (a_i - a^*)\psi(a_i) - \ln \Gamma(a_i) + \ln \Gamma(a^*) + a^*(\ln \frac{b_i}{b^*}) + a_i \frac{b^* - b_i}{b_i} \tag{33}$$

blablablablablabla
blablablablablabla
blablablablablabla

The marginal prior is then

$$\pi(\lambda) = \frac{1}{\mathcal{B}(X)} \int_A \frac{{b^*}^{a^*}}{\Gamma(a^*)} \lambda^{a^* - 1} e^{b^* \lambda} \alpha_i^{x_i - 1} d\boldsymbol{\alpha} \tag{34}$$

Normal priors

Now suppose one is interested in combining prior distributions on a quantity $\mu \in \mathbb{R}$. Suppose further that the expert priors are densities in the normal family, i.e.,

$$f_i(\mu; m_i, v_i) = \frac{1}{\sqrt{v_i 2\pi}} \exp\left(\frac{-(\mu - m_i)^2}{2v_i}\right)$$

Thus

$$\pi(\mu) = t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\mu; m_i, v_i)^{\alpha_i} \tag{35}$$

$$\propto \prod_{i=0}^{K} \left[\exp\left(-\frac{(\mu - m_i)^2}{2v_i}\right)\right]^{\alpha_i} \tag{36}$$

$$\propto \exp\left[-\frac{1}{2}\left\{\mu^2 \sum_{i=0}^{K} \frac{\alpha_i}{v_i} - 2\mu \sum_{i=0}^{K} \frac{\alpha_i m_i}{v_i}\right\}\right] \tag{37}$$

which yields a normal distribution with parameters and $m^* = \frac{\sum_{i=0}^{K} w_i m_i}{\sum_{i=0}^{K} w_i}$ and $v^* = [\sum_{i=0}^{K} w_i]^{-1}$, where $w_i = \alpha_i / v_i$. The entropy function is then:

$$H_\pi(\mu) = \frac{1}{2}\left[\ln(2\pi e) - \ln \sum_{i=0}^{K} w_i\right] \tag{38}$$

which achieves its maximum when $\alpha_j = 1$ for $s_j = max(s_1, s_2, \ldots, s_K)$.

Results

Table 1 lists the weights proposed by each method. Figure 1 shows the prior and posterior distributions in each of the methods and also the case in which we assign an equal weight ($1/K$) to each opinion. It is interesting to note that maximum entropy suggests to discard all opinions but one, which effectively leads to the maximum entropy. Since $t(\boldsymbol{\alpha})$ is concave, we expect to find the maximum entropy given by the boundary conditions, which may lead to border points in the simplex. Minimising Kullback-Leibler divergence between each prior and the pooled prior leads to finding a unique solution but in this case also suggests to discard two of the opinions. By contrast, using a hierarchical Dirichlet prior for the weights gives rather different results from the first two methods in proposing almost equal weights to each of the opinions. One can get insight into these results by looking at the integrated likelihoods in Table 2 and the densities in Figure 1, we note that all three methods lead to similar pooled distributions. Note that the only distribution with a substantially different $l(y)$ is that of Expert 3, who gave a rather divergent mean for the survival probability ($m_3 = 0.70$).

In conclusion, if the prior distributions (opinions) are not radically different, all three methods will probably lead to similar combined priors. Although this is the case for the simple univariate example presented, it remains to be seen if this is the case for high-dimensional θ under complex sampling distributions. As the results presented in this paper make clear, future research shall be focused on cases where there is substantial heterogeinity in the available opinions. Moreover, a sensitivity analysis for $\pi(\boldsymbol{\alpha})$ is desirable to understand how much we can lead about the experts reliabilities *a posteriori*.

Table 1: Weights obtained using the three methods for the proportion estimation problem. [1] – Kullback-Leibler [2] – Posterior mean for $\boldsymbol{\alpha}$.

Method	α_0	α_1	α_2	α_3
Maximum entropy	0.00	1.00	0.00	0.00
Minimum KL[1] divergence	0.04	0.96	0.00	0.00
Hierarchical prior[2]	0.26	0.24	0.26	0.23

Table 2: Integrated likelihoods ($l(y)$) for the priors of each expert as well as the combined priors. [1] Calculated using the posterior mean of $\boldsymbol{\alpha}$

Expert priors		Pooled priors	
Expert 0	0.237	Equal weights	0.254
Expert 1	0.211	Maximum entropy	0.211
Expert 2	0.256	Minimum KL	0.223
Expert 3	0.163	Hierarchical[1]	0.255

Method	Prior	Posterior
Equal weights	0.90 (0.64–1.00)	0.90 (0.73–0.99)
Maximum entropy	0.80 (0.37–1.00)	0.87 (0.66–0.98)
Minimum KL	0.82 (0.42–1.00)	0.87 (0.67–0.99)
Hierarchical prior	0.88 (0.53–1.00)	0.90 (0.71–0.99)

Acknowledgements

The authors would like to thank Professor Adrian Raftery (University of Washington) for helpful suggestions. DAMV and LSB were supported in part by Capes under Capes/Cofecub project (N. 833/15). FCC is grateful to Fundação Getulio Vargas for funding during this project.

References

Abbas, A. E. (2009). A Kullback-Leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37.

Coelho, F. C. and Codeço, C. T. (2009). Dynamic modeling of vaccinating behavior as a function of individual beliefs. *PLoS Comput. Biol.*, 5(7):e1000425.

Genest, C., McConway, K. J., and Schervish, M. J. (1986). Characterization of externally bayesian pooling operators. *The Annals of Statistics*, pages 487–501.

Genest, C., Weerahandi, S., and Zidek, J. V. (1984). Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17(1):61–70.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, pages 114–135.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. II. *Physical Review*, 108:171–190.

Lind, N. C. and Nowak, A. S. (1988). Pooling expert opinions on probability distributions. *Journal of engineering mechanics*, 114(2):328–341.

Myung, I. J., Ramamoorti, S., and Bailey Jr, A. D. (1996). Maximum entropy aggregation of expert predictions. *Management Science*, 42(10):1420–1436.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: the bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics*, pages 1–45. Oxford University Press.

Robert, C. P. (2001). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics) by.* Springer-Verlag New York.

Rufo, M., Martin, J., Pérez, C., et al. (2012a). Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach. *Bayesian Analysis*, 7(2):411–438.

Rufo, M. J., Pérez, C. J., Martín, J., et al. (2012b). A bayesian approach to aggregate experts' initial information. *Electronic Journal of Statistics*, 6:2362–2382.

Savchuk, V. P. and Martz, H. F. (1994). Bayes reliability estimation using multiple sources of prior information: binomial sampling. *Reliability, IEEE Transactions on*, 43(1):138–144.

Stan Development Team (2014). Stan: A c++ library for probability and sampling, version 2.6.0.

West, M. (1984). Bayesian aggregation. *Journal of the Royal Statistical Society. Series A (General)*, pages 600–607.

Yeh, C.-C. (2011). Hölder's inequality and related inequalities in probability. *International Journal of Artificial Life Research (IJALR)*, 2(1):54–61.

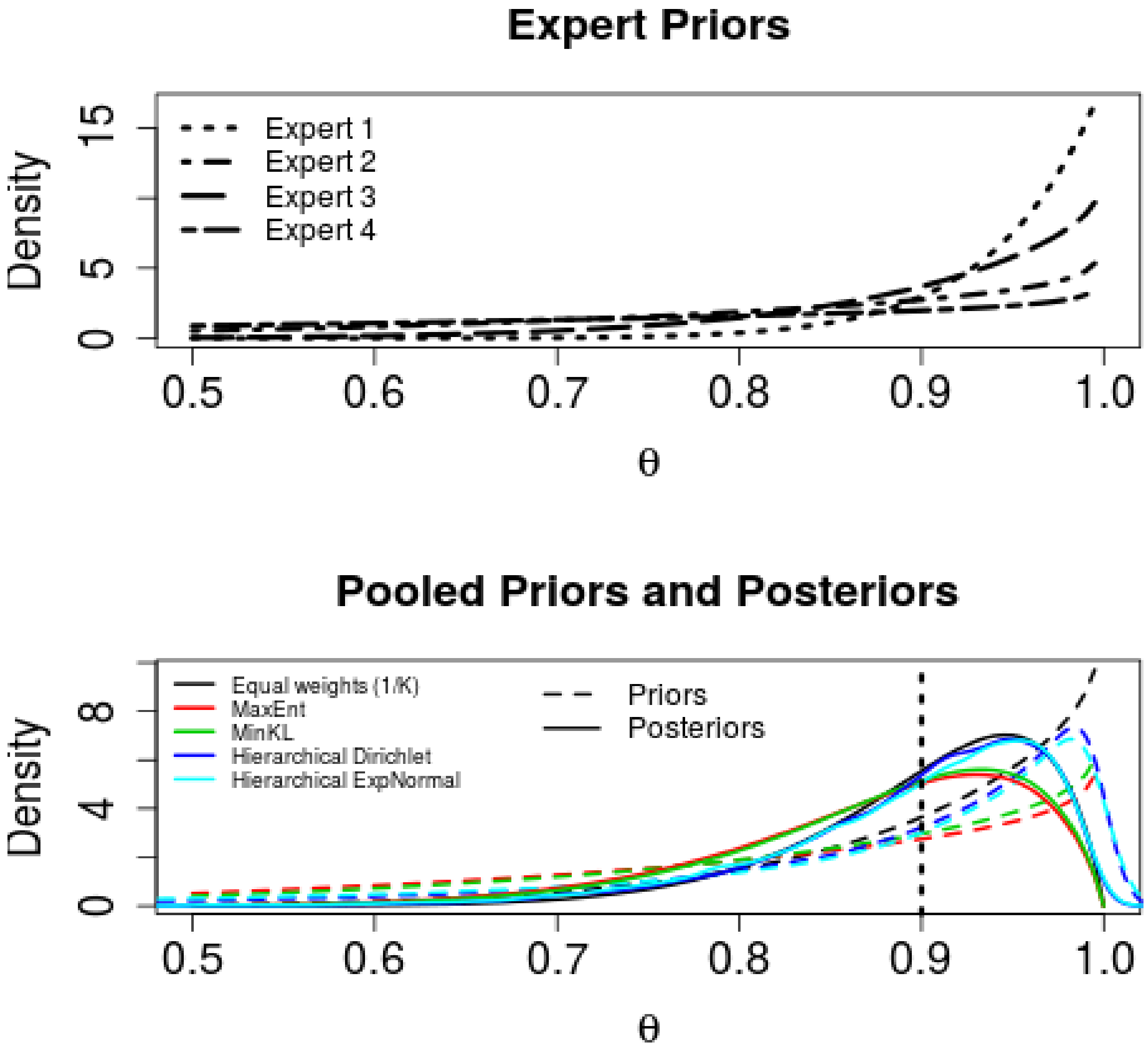

Figure 1: **Prior and posterior densities for θ.** Top panel shows the distributions elicited by each expert (data from Savchuk and Martz (1994)) and the bottom panel shows the pooled priors and posteriors obtained using each of the three methods discussed in this paper. The dashed vertical line marks the maximum likelihood estimate of θ, $\hat{\theta} = 9/10$. [NEEDS CHANGING]

Appendix

Proofs

Here we provide a simple proof of theorem 1 using Hölder's inequality.

Proof. We begin by noting that $\pi(\theta)$ can be re-written as:

$$\pi(\theta) \propto f_0(\theta) \prod_{j=1}^{K} \left(\frac{f_j(\theta)}{f_0(\theta)} \right)^{\alpha_j}. \tag{39}$$

Let $X_j = \frac{f_j(\theta)}{f_0(\theta)}, j = 1, 2, \ldots, K$. Then integrating the expression in (39) is equivalent to finding

$$E_0 \left[\prod_{j=1}^{K} X_j^{\alpha_j} \right] \leq \prod_{j=1}^{K} E_0[X_j]^{\alpha_j}, \tag{40}$$

where $E_0[\cdot]$ is the expectation w.r.t f_0 and (40) follows from Hölder's inequality for expectations (Yeh, 2011). Since $\forall j$ we have $E_0[X_j]^{\alpha_j} = \left(\int_{\boldsymbol{\Theta}} f_0(\theta)\frac{f_j(\theta)}{f_0(\theta)}\right)^{\alpha_j} d\theta = 1^{\alpha_j} = 1$, Theorem 1 is proven. □

It is straightforward to show that remark 1 holds:

Proof.

$$\pi(\theta) \propto \prod_{i=0}^{K}[\exp(\psi_i(\theta))]^{\alpha_i} \tag{41}$$

$$\propto \exp(\psi^*(\theta)), \tag{42}$$

where $\psi^*(\theta) = \sum_{i=0}^{K} \alpha_i \psi_i(\theta)$ is a concave function due to being a linear combination of concave functions. □

To see that equation 7 holds:

$$\begin{aligned} H_\pi(\theta) &= E[-\log(\pi(\theta)] \\ &= -\int \log(\pi(\theta)\pi(\theta)d\theta \\ &= -\int (\log(K(a^*, b^*) + \theta a^* - \psi(\theta)b^*)\pi(\theta)d\theta \\ &= -\log(K(a^*, b^*) + a^*E[\theta] - b^*E[\psi(\theta)] \end{aligned}$$

Likewise for equation 8, we have

$$\begin{aligned} KL(f_i||\pi) &= E_\pi[\log(f_i(\theta) - \log(\pi(\theta)] \\ &= \int [\log(K(a_i, b_i)e^{\theta a_i - b_i\psi(\theta)}) - \log(K(a^*, b^*)e^{\theta a^* - b^*\psi(\theta)})]\pi(\theta)d\theta \\ &= \int [\log(K(a_i, b_i)) - \log(K(a^*, b^*)) + (a_i - a^*)\theta - (b_i - b^*)\psi(\theta)\pi(\theta)d\theta \\ &= \log(K(a_i, b_i)) - \log(K(a^*, b^*)) + (a_i - a^*)E[\theta] - (b_i - b^*)E[\psi(\theta)] \end{aligned}$$