Combining probability distributions: Extending the logarithmic pooling approach

Abstract

Combining 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. However, the resulting pooled distribution depends heavily on set of weights given to each opinion/prior and thus careful consideration must be given to the choice of weights. In this paper we review and extend the statistical theory of logarithmic pooling, focusing on the assignment of the weights using a hierarchical prior distribution. We explore several statistical applications, such as the estimation of survival probabilities, meta-analysis and Bayesian melding of deterministic models of population growth and epidemics. We show that it is possible learn the weights from data, although identifiability issues may arise for some configurations of priors and data. Furthermore, we show how the hierarchical approach leads to posterior distributions that are able to accommodate prior-data conflict in complex models.

Keywords: expert opinion; hierarchical modelling; meta analysis; Bayesian melding.

1 Introduction

Combining probability distributions is a topic of general interest, both in the statistical (West 1984; Genest, McConway, et al. 1986; Genest and Zidek 1986) and decision theory literatures (Genest et al. 1984; French 1985; Pennock et al. 1997; Guardoni 2002). On the theoretical front, studying opinion pooling operators may give important insights on consensus belief formation and group decision making (West 1984; Genest and Zidek 1986; Guardoni 2002). 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, McConway, et al. 1986) – see below. In a practical setting, logarithmic pooling finds use in a wide range of fields, from engineering (Lind et al. 1988; Savchuk et al. 1994) to wildlife conservation (Poole et al. 2000) and infectious disease modelling (F. C. Coelho et al. 2009).

A common situation of interest is combining expert opinions about a quantity of interest $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ when these opinions can be represented as (proper) probability distributions. Combining these opinions using logarithmic pooling requires assigning weights to each of the experts, which represent the (relative) reliability of each opinion (Genest et al. 1984; French 1985). This requirement naturally leads to the question of how to choose the weights in a meaningful way, 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, Martin, et al. 2012; Rufo, Pérez, et al. 2012).

While moving away from the problem of arbitrarily assigning the weights, these approaches arrive at single point solutions, similar to point estimates in statistical theory. While we acknowledge that these approaches have merit, we argue that in many settings it would be desirable to incorporate information on the relative reliabilities of the experts into the pooling procedure while accommodating uncertainty about the weights. 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 (Poole et al. 2000). 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 discuss previous approaches for assigning the weights based on optimality criteria and study assigning hierarchical priors to the weights in order to learn about them from data. This paper is organised as follows: in Section 2 we introduce the necessary concepts and notation on logarithmic pooling, as well as some its key properties. We also prove a new result about log-concavity of the pooled distribution when all distributions are log-concave. In Section 3 we present different approaches to choosing the weights, two methods based on optimality criteria, namely maximising the entropy of the pooled prior and minimising Kullback-Leibler divergence between the pooled distribution and the expert distributions. In addition we also lay out an approach hierarchical modelling of the weights. Section 4 contains applications of logarithmic pooling to reliability analysis (Sections 4.1 and 4.3), meta-analysis (Section 4.2) and Bayesian melding (Section 4.4). We conclude with a discussion of our results in light of the statistical literature in Section 5.

2 Logarithmic pooling: properties and applications

In this section 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 $\mathbf{F}_\theta := \{f_0(\theta), f_1(\theta), \ldots, f_K(\theta)\}$ be a set of (densities of) 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 $\sum_{i=0}^{K} \alpha_i = 1$. The log-pooled density is

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

where $t(\boldsymbol{\alpha}) = \left[\int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i}\, d\theta\right]^{-1}$.

Logarithmic pooling will only yield proper probability distributions if it is possible to normalise the expression in (1). This condition is usually assumed implicitly, without proof. While Poole et al. (2000) provide a proof for the case of two densities (see Theorem 1 therein), Genest, McConway, et al. (1986) (pg.489) prove the result for a finite number of densities:

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

We give a simple proof using Hölder's inequality in the Appendix. This result ensures any (finite) number of proper distributions can be combined using the logarithmic pooling operator to yield a normalisable (proper) density. In addition, log-linear pools enjoy the *external Bayesianity* property (Remark 1), which guarantees that whether one combines the expert opinions before or after observing evidence does not affect the resulting pooled distribution.

Remark 1. ***External Bayesianity (Genest et al. 1984)***. *If the expert opinions are given by densities $f_i(\theta)$ and one observes data x such that one can specify a likelihood $l(x \mid \theta)$, combining the set of posteriors $p_i(\theta \mid x) \propto l(x \mid \theta) f_i(\theta)$ yields the same distribution as combining the densities f_i to obtain a prior $\pi(\theta)$ and then combine it with $l(x \mid \theta)$ to obtain a posterior $p(\theta \mid x) \propto l(x \mid \theta)\pi(\theta)$.*

Proof. Combining the posteriors $p_i(\cdot)$ gives

$$\begin{aligned} p'(\theta \mid x, \boldsymbol{\alpha}) &\propto \prod_{i=0}^{K} \left[l(x \mid \theta) f_i(\theta)\right]^{\alpha_i}, \\ &\propto l(x \mid \theta) \prod_{i=0}^{K} f_i(\theta)^{\alpha_i}, \\ &= \frac{l(x \mid \theta)\pi(\theta \mid \boldsymbol{\alpha})}{m'(x)} \equiv p(\theta \mid x, \boldsymbol{\alpha}), \end{aligned}$$

where the second line follows from $\sum_{i=0}^{K} \alpha_i = 1$. □

Genest et al. (1984) show that the logarithmic pooling operator in (1) is the **only** aggregation (pooling) operator that enjoys external Bayesianity. Moreover, the logarithmic pooling operator has the relative propensity consistency (RPC) property (Remark 2), whereby the pooled opinion preserves relative judgments from the experts.

Remark 2. ***Relative propensity consistency (Genest et al. 1984)***. *Taking $\boldsymbol{F}_X$ as a set of expert opinions with support on a space $\mathcal{X}$, define $\boldsymbol{\xi} = \{\boldsymbol{F}_X, a, b\}$ for arbitrary*

$a, b \in \mathcal{X}$. Let $\mathcal{T}$ be a pooling operator and define two functions U and V such that

$$U(\boldsymbol{\xi}) := \left(\frac{f_0(a)}{f_0(b)}, \frac{f_1(a)}{f_1(b)}, \ldots, \frac{f_K(a)}{f_K(b)}\right) \text{ and } \tag{2}$$

$$V(\boldsymbol{\xi}) := \frac{\mathcal{T}_{\boldsymbol{F}_X}(a)}{\mathcal{T}_{\boldsymbol{F}_X}(b)}. \tag{3}$$

We then say that $\mathcal{T}$ enjoys relative propensity consistency (RPC) if and only if

$$U(\boldsymbol{\xi}_1) \geq U(\boldsymbol{\xi}_2) \implies V(\boldsymbol{\xi}_1) \geq V(\boldsymbol{\xi}_2), \tag{4}$$

for all $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$.

We refer the reader to Genest et al. (1984) for a proof. Informally, this property says that if all experts consider a particular event A more probable than another event B, then the pooled opinion should be consistent with these relative judgments. Genest et al. (1984) show that for mild conditions on $\mathcal{X}$, namely $|\mathcal{X}| \geq 3$, the logarithmic pooling operator is the only pooling operator with RPC (see also Lemma 1 in Appendix A.1).

Another desirable property of the logarithmic pooling operator is log-concavity. Log-concavity of the pooled prior may be important to consider in order to guarantee unimodality and certain conditions on tail behaviour – see Bagnoli et al. (2005). This motivates the following theorem, which is, to the best of our knowledge, a new result:

Theorem 2. ***Log-concavity***. *If $\mathbf{F}_\theta$ is a set of log-concave distributions, then $\pi(\theta \mid \boldsymbol{\alpha})$ is also log-concave. Moreover, logarithmic pooling is the only pooling operator that will always produce a log-concave density when all the elements of $\mathbf{F}_\theta$ are log-concave.*

Proof. See the Appendix. □

Theorem 2 tells us that logarithmic pooling is the only aggregation method to universally preserve log-concavity, for any configuration of the weights ($\boldsymbol{\alpha}$). This universality result is important because it holds for any set of log-concave distributions, $\mathbf{F}_\theta$. In contrast, a linear pool of $K+1$ Gaussian distributions with common mean, $\pi_{\text{linear}}(\theta) = \sum_{i=0}^{K} \alpha_i f_i(\theta)$, would produce a log-concave pooled distribution for any $\boldsymbol{\alpha}$, but this would potentially fail if the means were different.

2.1 Exponential family

The exponential family of probability distributions finds widespread in the modelling of empirical phenomena. In this section we give expressions for the entropy and Kullback-Leibler divergence for the pooled distributions. These will be useful in applications presented later in the paper.

Suppose we are interested in a random variable Y whose distribution belongs to the exponential family with parameter θ and probability density function (pdf) given by

$$f(y|\theta) = h(y)e^{\theta y - s(\theta)}. \tag{5}$$

Let $\mathbf{F}_y$ be a set of densities on y of the form in (5), $f_i(y|\theta_i)$, $i = 0, 1, \ldots, K$. The combined (log-pooled) distribution also belongs to the exponential family:

$$\pi(y|\boldsymbol{\alpha}) = t(\boldsymbol{\alpha})h^*(y)e^{\theta^* y - s^*(\boldsymbol{\theta})}, \tag{6}$$

where $\boldsymbol{\theta} := \{\theta_0, \theta_1, \ldots, \theta_K\}$, $h^*(y) = \prod_{i=0}^{K} h_i(y)^{\alpha_i}$, $\theta^* = \sum_{i=0}^{K} \alpha_i\theta_i$ and $s^*(\boldsymbol{\theta}) = \sum_{i=0}^{K} \alpha_i s_i(\theta_i)$.

The entropy function of the log-pooled distribution is

$$H_\pi(Y; \boldsymbol{\alpha}) := -\mathbb{E}_\pi\left[-\log \pi(Y|\boldsymbol{\alpha})\right] = -\log t(\boldsymbol{\alpha}) + s^*(\boldsymbol{\theta}) - \mathbb{E}_\pi[\log h^*(Y)] - \theta^*\mathbb{E}_\pi[Y]\,, \tag{7}$$

where $\mathbb{E}_\pi\left[g(Y)\right]$ is the expectation of a π-measurable function $g(Y)$ with respect to $\pi(y|\boldsymbol{\alpha})$, when the integral exists.

The Kullback-Leibler divergence between the pooled distribution (6) and each distribution in $\mathbf{F}_y$ can be written as:

$$\mathrm{KL}(\pi||f_i) = -H_\pi(Y;\boldsymbol{\alpha}) - \mathbb{E}_\pi[\log h_i(Y)] - \theta_i \mathbb{E}_\pi[Y] + s_i(\theta_i). \tag{8}$$

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 (see also Appendix A.4).

2.1.1 Conjugate priors to the exponential family

A conjugate prior family for $f(y|\theta)$ (5), has the following form (Diaconis et al. 1979):

$$g(\theta|a,b) = K(a,b)e^{\theta a - bs(\theta)}\,, \tag{9}$$

where $K(a,b)$ is a normalising constant. Similar to the above, let $\mathbf{G}_\theta$ be a set of log-conjugate prior distributions representing the opinions of $K+1$ experts, and $g_i(\theta) = g(\theta|a_i,b_i)$ from equation (9).

The log-pooled prior is also a conjugate prior for $f(y|\theta)$ with hyperparameters given by a weighted mean of the experts's hyperparameters, i.e., $\pi(\theta|\boldsymbol{\alpha}) = g(\theta|a^*,b^*)$, where $a^* = \sum_{i=0}^K \alpha_i a_i$ and $b^* = \sum_{i=0}^K \alpha_i b_i$.

The entropy function of the log-pooled prior (9) is given by

$$H_\pi(\theta;\boldsymbol{\alpha}) = -\log(K(a^*,b^*)) - a^*\mathbb{E}_\pi[\theta] + b^*\mathbb{E}_\pi[s(\theta)]\,. \tag{10}$$

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

$$KL(\pi||g_i) = -H_\pi(\theta;\boldsymbol{\alpha}) - \log(K(a_i,b_i)) - a_i\mathbb{E}_\pi[\theta] + b_i\mathbb{E}_\pi[s(\theta)]\,. \tag{11}$$

2.2 Bayesian melding

Another important application of logarithmic pooling is in the Bayesian melding method of Poole et al. (2000). Deterministic simulation models are widespread in Science and Engineering (see Poole et al. (2000) and references therein). One is often interested in a deterministic model M with inputs $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^p$ and outputs $\phi \in \mathbf{\Phi} \subseteq \mathbb{R}^q$, such that $\phi = M(\theta)$. If one wants to learn about θ from data and a (prior) distribution on ϕ is available, then one needs a method to combine the information between the prior on θ and the prior induced on it through M, which is often non-invertible.

Bayesian melding seeks to draw inference by first employing logarithmic pooling to construct a prior on ϕ of the form

$$\tilde{q}_{\Phi}(\phi) \propto q_1^*(\phi)^{\alpha} q_2(\phi)^{1-\alpha}, \tag{12}$$

where $q_1^*()$ is the **induced** prior on the outputs and q_2 is the prior on ϕ without considering the deterministic model, henceforth called the natural prior on ϕ. The prior in (12) can then be inverted to obtain a *coherised* prior on θ, $\tilde{q}_{\Theta}(\theta)$. Poole et al. (2000) give a way of obtaining $\tilde{q}_{\Theta}$ even when M is non-invertible, which we will not discuss further here.

Standard Bayesian inference may then follow, leading to the posterior

$$p_{\Theta}(\theta) \propto \tilde{q}_{\Theta}(\theta) L_1(\theta) L_2(M(\theta)), \tag{13}$$

which enjoys all the properties of usual posterior distributions. The method allows standard Bayesian inference to be carried out about all quantities of interest in the model, which makes it attractive to application in policy making (Alkema et al. 2008), where proper acknowledgment of uncertainty is crucial.

In Poole et al. (2000) (Section 6.2 therein), the authors fix $\alpha = 1/2$, justifying their choice by the fact that while the weights should reflect the reliability of each expert (information source). In their Bayesian melding analysis, one is combining distributions based on different bodies of evidence, but assessed by the same expert. Another option is to fix $\alpha = 1 - \epsilon$, with ϵ small (Alkema et al. 2007). This can be useful when the prior distribution on outputs is uniform, as it still enforces the constraint, but keeps the prior information about the inputs. Here we relax the restriction of fixing the weight, instead modelling α through a hyperprior – see Sections 4.4.1 and 4.4.2.

3 Assigning the weights in logarithmic pooling

The weights ($\boldsymbol{\alpha}$) play a key role on the logarithmic pooling and hence their choice is critical. Building on work by Poole et al. (2000), Rufo, Martin, et al. (2012), and Rufo, Pérez, et al. (2012) and Abbas (2009), we now move on to study three approaches to assigning the weights in logarithmic pooling. The first two approaches are based on optimality criteria and a third method proposes assigning a (hyper)prior to the weights.

3.1 Choosing weights based on optimality criteria

The first set of approaches we will consider attempt to assign the weights by achieving an optimality condition using only information contained in the expert distributions themselves, without reference to any external information such as observed data.

3.1.1 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 diffuse. According to its proponents, such an 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 et al. 1994). In order to choose $\boldsymbol{\alpha}$ so as to maximise prior diffuseness, one can maximise the entropy of the log-pooled prior, i.e.

$$\begin{aligned} H_\pi(\theta;\boldsymbol{\alpha}) = \mathbb{E}_\pi\left[-\log \pi(\theta)\right] &= -\int_{\boldsymbol{\Theta}} \pi(\theta)\log\pi(\theta)\,d\theta, \\ &= -\sum_{i=0}^{K} \alpha_i \mathbb{E}_\pi[\log f_i] - \log t(\boldsymbol{\alpha}). \end{aligned} \tag{14}$$

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

$$\hat{\boldsymbol{\alpha}} := \arg\max_{\boldsymbol{\alpha}} H_\pi(\theta;\boldsymbol{\alpha}). \tag{15}$$

This approach, however, does not result in a convex optimisation problem, therefore one is not guaranteed to find a unique solution – see Remark 3 for intuition as to why. A possible resolution to the non-uniqueness of the maximum entropy solution would be to add further constraints, for instance requiring that $E_\pi[\theta] = m$. It is however unclear which set of constraints would ensure uniqueness.

3.1.2 Minimising Kullback-Leibler divergence

One could also wish to choose the pooling weights so as to minimise the total Kullback-Leibler divergence between the pooled distribution, π, and each distribution in $\mathbf{F}_\theta$. Let $E_i[g]$ be the expectation of a measurable function $g : \boldsymbol{\Theta} \to \mathbb{R}$ with respect to each density f_i. We can define a loss function such that

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=0}^{K} \mathrm{KL}(f_i||\pi), \\ &= -(K+1)\log t(\boldsymbol{\alpha}) - (K+1)\sum_{i=0}^{K}\alpha_i\mathbb{E}_i[\log f_i] + \sum_{i=0}^{K}\mathbb{E}_i\left[\log f_i\right], \end{aligned} \tag{16}$$

and we want to find

$$\hat{\boldsymbol{\alpha}} := \arg\min_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}). \tag{17}$$

Fortunately, this set up leads to a unique solution, a result we summarise in Remark 3.

Remark 3. ***Uniqueness of the minimum KL solution***. *The distribution obtained following (17) is unique, i.e., there is only one aggregated prior $\pi(\theta \mid \boldsymbol{\alpha})$ that minimizes $L(\boldsymbol{\alpha})$.*

Proof. We begin by noting that the second term in (16) is a linear combinations of the weights, and hence we may restrict attention only to the first term – the third term does not depend on $\boldsymbol{\alpha}$. Next, recall that minimising (16) is equivalent to maximising $\log t(\boldsymbol{\alpha}) = \log \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i}\, d\theta$. Proposition 3.1 in Rufo, Martin, et al. (2012) states that $t(\boldsymbol{\alpha})$ is (log-)concave, therefore any optimisation problem which involves minimising $-\log t(\boldsymbol{\alpha})$ is convex. We thus conclude that the problem in (17) has a unique solution. □

By contrast, the problem in (15) requires one to minimise $\ln t(\boldsymbol{\alpha})$, hence lacking a sufficient condition for the existence of a unique solution. Likewise, using the loss function $L'(\boldsymbol{\alpha}) = \sum_{i=0}^{K} \mathrm{KL}(\pi || f_i)$ would not lead to a unique solution. See Appendix A.2 for implementation details.

3.2 Hierarchical modelling of the weights

As discussed by Poole et al. (2000) and others (Zhong et al. 2015; Li et al. 2017), estimating the weights would be of interest since this would allow one to assess the reliability of each source of information (expert). Li et al. (2017) explore the idea of computing the pooled distribution for several values of the weights. Whilst informative, this approach has two issues: (a) it does not scale well with increasing the number of distributions being combined,

K, and; (b) it fails to account for any (posterior) dependence between model parameters and the weights. In this section we propose placing a hierarchical prior on the weights, allowing for standard Bayesian inference about these quantities.

A natural choice for a prior distribution for $\boldsymbol{\alpha}$ is the $(K+1)-$dimensional Dirichlet distribution

$$\pi_A(\boldsymbol{\alpha}) = \frac{1}{\mathcal{B}(\boldsymbol{x})} \prod_{i=0}^{K} \alpha_i^{x_i - 1}, \tag{18}$$

where $\boldsymbol{x} = \{x_0, x_1, \ldots, x_K\}$ is the vector of hyperparameters for the Dirichlet prior and $\mathcal{B}(X)$ is the multinomial beta function. The Dirichlet offers a simple, albeit potentially inflexible prior.

A more flexible prior for $\boldsymbol{\alpha}$ is the logistic-normal distribution (Aitchison et al. 1980):

$$\pi_A(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} \frac{1}{\prod_{i=0}^{K} \alpha_i} \exp\left(\left(\log\left(\frac{\boldsymbol{\alpha}_{-K}}{\alpha_K} \right) - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}^{-1} \left(\log\left(\frac{\boldsymbol{\alpha}_{-K}}{\alpha_K} \right) - \boldsymbol{\mu} \right) \right), \tag{19}$$

where $\boldsymbol{\alpha}_{-K}$ represents the vector $\boldsymbol{\alpha}$ without the K-th element, $\boldsymbol{\mu}$ is a K-size mean vector, and $\boldsymbol{\Sigma}$ is a $K \times K$ covariance matrix. (Aitchison et al. 1980) propose choosing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ minimizing the KL divergence between the Dirichlet (18) and the logistic-normal (19) distributions, i.e.

$$\mu_i = \psi(x_i) - \psi(x_K), \quad i = 0, 1, \ldots, K-1, \tag{20}$$

$$\Sigma_{ii} = \psi'(x_i) + \psi'(x_K), \quad i = 0, 1, \ldots, K-1, \tag{21}$$

$$\Sigma_{ij} = \psi'(x_K), \tag{22}$$

where $\psi(\cdot)$ is the digamma function, and $\psi'(\cdot)$ is the trigamma function.

The marginal prior for θ,

$$\tilde{\pi}(\theta) = \int_{\mathcal{A}} \pi(\theta \mid \boldsymbol{\alpha}) \pi_A(\boldsymbol{\alpha}) d\boldsymbol{\alpha}, \tag{23}$$

can also be efficiently approximated through Monte Carlo sampling when π can be written in closed-form. Even when it cannot be expressed analytically, it is still possible to sample from the marginal prior by using quadrature-based methods for computing $t(\boldsymbol{\alpha})$ when θ is unidimensional (see Discussion).

Concerning posterior inference, the marginal posterior for θ can be obtained through standard methods and shall not be discussed further. The next object to consider is the marginal posterior for the weights, $p(\boldsymbol{\alpha} \mid \boldsymbol{x})$, which can be obtained through

$$\begin{aligned} p(\boldsymbol{\alpha} \mid \boldsymbol{x}) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\alpha}, \theta \mid \boldsymbol{x})\, d\theta, \\ &= \int_{\boldsymbol{\Theta}} \frac{L(\boldsymbol{x} \mid \theta)\pi(\theta \mid \boldsymbol{\alpha})\pi_A(\boldsymbol{\alpha})}{c(\boldsymbol{x})}\, d\theta, \\ &= \frac{\pi_A(\boldsymbol{\alpha})}{c(\boldsymbol{x})} \int_{\boldsymbol{\Theta}} L(\boldsymbol{x} \mid \theta)\pi(\theta \mid \boldsymbol{\alpha})\, d\theta, \\ &\propto \pi_A(\boldsymbol{\alpha})\kappa(\boldsymbol{\alpha}, \boldsymbol{x}), \end{aligned} \tag{24}$$

where $c(\boldsymbol{x}) := \int_{\mathcal{A}} \int_{\boldsymbol{\Theta}} p(\boldsymbol{\alpha}, \theta \mid \boldsymbol{x})\, d\theta\, d\boldsymbol{\alpha}$.

In some situations, in particular the conjugate situation discussed in Section 2.1.1 and exemplified in the Applications section below, it is possible to write down $\kappa(\boldsymbol{\alpha}, \boldsymbol{x})$ in closed-form. This is very convenient because the posterior expectation of the weights, $E_p[\boldsymbol{\alpha} \mid \boldsymbol{x}]$, becomes $E_{\pi_A}[\boldsymbol{\alpha}\kappa(\boldsymbol{\alpha}, \boldsymbol{x})]$, i.e., the expectation of a known function with respect to the prior on the weights. This expectation can be easily and accurately approximated with simple Monte Carlo techniques rather than MCMC – see Section 4.3 for example applications.

4 Applications

In this section we shall present a wide range of applications for logarithmic pooling, from prior elicitation to meta-analysis to Bayesian melding. Computational details, along with instructions to get reproducible code, are given in Appendix A.2.

4.1 Elicitation: combining expert priors on survival probabilities

The first example we consider is combining expert opinions about probabilities and proportions. We analyse an example proposed by Savchuk et al. (1994) (also discussed in Rufo, Pérez, et al. (2012)) in which four experts are required supply prior information about the survival probability θ of a certain unit. The experts express their opinion as prior means for the survival probability, which Savchuk et al. (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 $\mathbf{b} = \{b_0 = 0.955, b_1 = 0.860, b_2 = 0.924, b_3 = 0.848\}$. Furthermore, an experiment is conducted and $y = 9$ successes out of $n = 10$ trials are observed. Thus, in this application we are able to estimate the posterior distribution for the survival probability and also, with the hierarchical modelling approach, the posterior distribution for the weights in face of the observed data. For the hierarchical priors, we employ a Dirichlet$(1/10, 1/10, 1/10, 1/10)$ and a moment-matching logistic-normal priors (see Section 4.3 for justification).

The probability distribution of the survival probability for the i-th expert is a Beta distribution with (hyper)parameters a_i and b_i. The log-pooled distribution for θ is then

$$\begin{aligned}\pi(\theta) &\propto \prod_{i=0}^{K} f_i(\theta; a_i, b_i)^{\alpha_i}, \\ &\propto \prod_{i=0}^{K} \left(\theta^{a_i-1}(1-\theta)^{b_i-1}\right)^{\alpha_i}, \\ &\propto \theta^{a^*-1}(1-\theta)^{b^*-1}, \qquad (25)\end{aligned}$$

with $a^* = \sum_{i=0}^K \alpha_i a_i$ and $b^* = \sum_{i=0}^K \alpha_i b_i$. Note that (25) is the kernel of a Beta distribution with parameters a^* and b^*. Hence the entropy is the following

$$H_\pi(\theta) = \log \mathcal{B}(a^*, b^*) - (a^* - 1)\psi(a^*) - (b^* - 1)\psi(b^*) + (a^* + b^* - 2)\psi(a^* + b^*). \tag{26}$$

And the KL divergence between $\pi(\theta)$ and $f_i(\theta)$ is

$$\begin{aligned} d_i = KL(f_i||\pi) = \ln\left(\frac{\mathcal{B}(a^*, b^*)}{\mathcal{B}(a_i, b_i)}\right) &+ (a_i - a^*)\psi(a_i) + (b_i - b^*)\psi(b_i) \\ &- (a_i - a^* + b_i - b^*)\psi(a_i + b_i). \end{aligned} \tag{27}$$

In this conjugate setting, the posteriors associated with each expert are also Beta distributions with parameters $a_i' = a_i + y$ and $b_i' = b_i + (n - y)$. This allows us to employ the maximum entropy and minimum KL procedures to combine these posterior distributions and thus make the weights comparable with the posterior means obtained with the hierarchical priors.

Our analysis of this example is thus split into two: weights for the priors and for the posteriors. Before observing any data, we can employ the optimisation procedures discussed above to obtain weights only taking into account information encoded in the expert priors themselves. To these optimisation procedures we add the technique of Rufo, Pérez, et al. (2012) which seeks to minimise KL distance between the pooled prior and the Jeffreys's posterior. When data are available, we can then use maximum entropy and minimum KL to obtain the weights in the same fashion as before, but now also estimate the posterior distribution of weights using a hierarchical prior. Finally, for this example we can also compute the integrated (marginal) likelihood of each expert, meaning that we can, assuming one of the experts is correct, compute "model" probabilities by normalising the marginal likelihoods (see Section 4.3, below).

In Table 1 we present weights obtained with the optimisation methods for the priors, including the solution found by Rufo, Pérez, et al. (2012) (Section 5.2 therein). With regard to the posteriors, we show maximum entropy, minimum KL along with posterior means of the weights under two prior distributions (Dirichlet and logistic-normal). Maximising the entropy of the pooled prior – and posterior – lead to the degenerate solution $\boldsymbol{\alpha} = \{0, 0, 0, 1\}$, which gives all the weight to the most diffuse prior distribution – $\text{Beta}(1.98, 0.848)$. Since $t(\boldsymbol{\alpha})$ is concave, we expect to find the maximum entropy given by the boundary conditions, which may lead to points in the border of the simplex. Unsurprisingly, the same solution was found by Rufo, Pérez, et al. (2012), whose method tends to favour more diffuse distributions. Minimising Kullback-Leibler divergence between the pooled prior and each expert prior leads to a unique solution but in this case also suggests to discard two of the opinions.

The hierarchical priors gave very similar posterior distributions for the weights, which assign the experts nearly equal weight, although the logistic-normal prior lead to results closer to the marginal likelihood-based weights.

Figure 1 shows the prior densities for each expert and pooling method and Table 2 contains the prior and posterior mean and credibility intervals from each of the methods and also the case in which we assign an equal weight $(1/K)$ to each opinion. Assigning equal weights actually gives a prior mean that is the same as the maximum likelihood estimate of θ, $\hat{\theta} = 9/10$. This explains why both hierarchical posteriors resemble equal weights so closely. Finally, we use the integrated (marginal) likelihood (Raftery et al. (2007), eq. 9), $l(y) = \int_0^1 f(y|\theta)\pi(\theta)\, d\theta$, as a univariate summary to compare the priors. The marginal

Table 1: **Weights obtained using different methods for the survival probability example (Savchuk et al. 1994).** [1] – Kullback-Leibler; [2] – Posterior mean for $\boldsymbol{\alpha}$.

	Method	α_0	α_1	α_2	α_3
Prior	Maximum entropy	0.00	0.00	0.00	1.00
	Minimum KL[1]	0.04	0.96	0.00	0.00
	Rufo, Martin, et al. (2012)	0.00	0.00	0.00	1.00
Posterior	Maximum entropy	0.00	0.00	0.00	1.00
	Minimum KL	0.17	0.83	0.00	0.00
	Dirichlet[2]	0.26	0.24	0.27	0.23
	Logistic-normal[2]	0.27	0.24	0.31	0.18
	Marginal likelihoods	0.27	0.24	0.30	0.19

likelihood for the i-th expert and J observations of the form $\{y_j, n_j\}$ is:

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

For the hierarchical priors we take the posterior mean of $(a^\star, b^\star)$ as (a_i, b_i). Results are given in Table 3 and show that, apart from expert 3 – and hence the maximum entropy pooled prior–, all other pooled priors and individual experts' priors give similar marginal likelihoods.

The posterior distribution for the weights estimated under both priors favours expert 2, the expert with the highest marginal likelihood. The logistic-normal gives expert 2 a higher weight when compared with the Dirichlet. This is connected to the increased flexibility of the logistic-normal (see Section 4.3).

We stress that that the marginal likelihoods are not being used here as a means of selecting priors, but rather as a useful univariate summary which is informative about the compatibility with the observed data and hence informative about prior-data conflict. While in this example one can gain insight into prior-data conflict from just the prior means and y/n, in other situations it might be harder to discern which expert gave the best (prior) guess.

Table 2: **Prior and posterior mean and credibility intervals for each method for assigning the weight, survival probability example (Savchuk et al. 1994).** Values for the hierarchical priors are from the marginal prior of θ in (23).

Method	Prior	Posterior
Equal weights	0.90 (0.64–1.00)	0.90 (0.73–0.99)
Maximum entropy	0.70 (0.17–0.99)	0.86 (0.63–0.98)
Minimum KL	0.82 (0.42–1.00)	0.87 (0.67–0.99)
Rufo, Pérez, et al. (2012)	0.70 (0.17–0.99)	0.86 (0.63–0.98)
Dirichlet	0.86 (0.40–1.00)	0.89 (0.70–0.99)
Logistic-normal	0.88 (0.35–1.00)	0.89 (0.71–0.99)

4.2 Meta-analysis: HIV prevalence among MSM populations in Brazil

Another potential application of logarithmic pooling is in meta analysis. Logarithmic pooling can also be used to combine probability distributions of a particular outcome estimated from several studies. In epidemiology, systematic review and meta analysis

Table 3: **Integrated likelihoods for the priors of each expert as well as the combined priors, failure probability example**. For the hierarchical priors we take the posterior expectations of $a^\star$ and $b^\star$ as a_i and b_i, respectively. [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	0.163
Expert 2	0.256	Minimum KL	0.223
Expert 3	0.163	Hierarchical prior[1] (Dirichlet/logistic-normal)	0.255

are popular tools for merging and contrasting results across multiple studies (Rothman et al. 2008, Chapter 33). Moreover, estimation of disease prevalence and the effect of exposure variables are amongst the most important application of meta-analyses in epidemiology. We illustrate the different approaches to assign weights in the logarithmic polling in the systematic review and meta analysis conducted by Malta et al. (2010). They analysed studies published from 1999 to 2009 assessing the HIV prevalence among men who have sex with another men (MSM) in Brazil. The authors have found six studies that estimated HIV prevalence in MSM population in Brazil. Data from each study consists of n_i observed individuals, y_i of which were infected with HIV.

Assuming a uniform prior for the HIV prevalence among MSM, denoted by φ, and a binomial model for each study, i.e. $Y_i \sim \text{Binomial}(n_i, \varphi)$. The posterior distribution for the HIV prevalence conditional on each study is then a Beta distribution with parameters $a_i = y_i + 1$ and $b_i = n_i - y_i + 1$, for $i = 0, 1, \ldots, 5$. For the first part of our analysis of this problem we will assume $\boldsymbol{F}_\varphi^B$ to be composed of these posterior distributions. In meta-analysis it is common for researchers to employ a Gaussian (normal) distribution instead of a distribution with support on $(0,1)$, relying on the large sample normal approximation of the binomial distribution. Here we study how this choice of representation impacts the logarithmic pooling procedure by comparing the Beta and Gaussian distributions as representations of the HIV prevalence among MSM.

If the probability density on φ for each study is now

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

where $m_i = y_i/n_i$ and $v_i = m_i(1-m_i)/n_i$. We have

$$\begin{aligned}
\pi(\varphi \mid \boldsymbol{\alpha}) &= t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\varphi; m_i, v_i)^{\alpha_i}, \\
&\propto \prod_{i=0}^{K} \left[\exp\left(\frac{-(\varphi - m_i)^2}{2v_i}\right)\right]^{\alpha_i}, \\
&\propto \exp\left[-\frac{1}{2}\left\{\varphi \sum_{i=0}^{K} \frac{\alpha_i}{v_i} - 2\varphi \sum_{i=0}^{K} \frac{\alpha_i m_i}{v_i} - \sum_{i=0}^{K} \frac{\alpha_i m_i^2}{v_i}\right\}\right]. \qquad (29)
\end{aligned}$$

Completing the square shows $\pi(\varphi)$ is the density of 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:

$$H_\pi(\varphi) = \frac{1}{2}\left[\ln(2\pi e) - \ln \sum_{i=0}^{K} w_i\right], \qquad (30)$$

which achieves its maximum when $\alpha_j = 1$ for $v_j = max(v_1, v_2, \dots, v_K)$ and thus maximising entropy always leads to degenerate solutions. The Kullback-Leibler divergence between the pooled distribution $\pi(\varphi)$ and each $f_i(\varphi)$ is then

$$\text{KL}(f_i || \pi) = \frac{1}{2} \log\left(\frac{v^\star}{v_i}\right) + \frac{v_i + (m_i - m^\star)^2}{2v^\star} - \frac{1}{2}. \qquad (31)$$

For the second part of our analysis of this example, we will assume that set of distributions to be combined, $\boldsymbol{F}_{\varphi}^{G}$, is composed by the Gaussian distributions described above.

Table 4 contains the sample size for each study, the total of HIV positive observed, and the estimated prevalence using the Beta distribution described above and a Gaussian distribution (see below). Note that the estimated prevalences among MSM are very high when compared with the HIV prevalence in the general population, 0.6% (Malta et al. 2010). In addition, there is considerable heterogeneity between studies, with (mean) estimates ranging from 6% (Tun et al. 2008) to 24%(Sutmoller et al. 2002; Barcellos et al. 2003).

Table 4: **Data extracted from the systematic review and meta analysis conducted by Malta et al. (2010) assessing the HIV prevalence among MSM in Brazil.** n_i is the sample size, y_i is the total of HIV-positive participants in the i-th study. Prevalence estimates are presented as mean and 95% credibility intervals, either from a Beta distribution with parameters $a_i = y_i + 1$ and $b_i = n_i - y_i + 1$ or a Gaussian distribution with $m_i = y_i/n_i$ and $v_i = m_i(1 - m_i)/n_i$ (see text).

				Estimated prevalence, φ (95% CI)	
Study	Reference	n	y	Beta	Gaussian
0	Tun et al. (2008)	658	44	0.068 (0.050–0.089)	0.067 (0.048–0.086)
1	Barcellos et al. (2003)	461	111	0.242 (0.204–0.282)	0.241 (0.202–0.280)
2	Carneiro et al. (2003)	621	61	0.100 (0.077–0.124)	0.098 (0.075–0.122)
3	Sutmoller et al. (2002)	1165	281	0.242 (0.218–0.267)	0.241 (0.217–0.266)
4	Brazilian Ministry of Health (2000)	642	57	0.090 (0.069–0.113)	0.089 (0.067–0.111)
5	Harrison et al. (1999)	849	99	0.118 (0.097–0.140)	0.117 (0.095–0.138)

In a meta-analytic context it also makes sense to consider giving weights to each study proportional to sample size, with larger studies receiving larger weights and hence more credence. We have included this weighting, with $\alpha_i = n_i / \sum_{i=0}^{K} n_i$, in our analysis, along with the maximum entropy and minimum KL solutions. The weights obtained by maximising entropy and minimising KL divergence for both the Beta and Gaussian representations of the prevalence information are given in Table 5. While the maximum entropy method lead to the same degenerate solution for both distributions, giving all the weight to the study by Barcellos et al. (2003), minimising KL divergence lead to the studies by Tun et al. (2008) and Barcellos et al. (2003) being given non-zero weights. Interestingly, while the minimum KL method for the Beta distribution representation lead to roughly equal weights, with the study by Tun et al. (2008) being slightly favoured, the solution for the Gaussian representation assigned a much larger weight to the distribution from Barcellos et al. (2003) (see Discussion).

Table 5: **Weights obtained using different methods for the HIV prevalence example**. Sample size pertains to assigning the weights based on the normalised sample sizes ($\alpha_i = n_i / \sum_{i=0}^{K} n_i$).

Method		α_0	α_1	α_2	α_3	α_4	α_5
Maximum entropy	Beta	0	1	0	0	0	0
	Gaussian	0	1	0	0	0	0
Minimum KL divergence	Beta	0.53	0.47	0	0	0	0
	Gaussian	0.17	0.83	0	0	0	0
Sample size		0.15	0.11	0.14	0.27	0.15	0.19

Table 6 shows the estimates of the HIV prevalence among MSM in Brazil using different methods for obtaining the pooled distributions and Figure 2 shows the resulting densities.

For comparison, we also included results from the log-pooled distributions obtained with equal weights and the marginal prior on φ induced by the Dirichlet (1/10, 1/10, 1/10, 1/10, 1/10, 1/10) and moment-matching logistic-normal priors on $\boldsymbol{\alpha}$. The heterogeneity observed across studies (Table 4) is also reflected in the variation in the combined (pooled) priors under different methods. In contrast to the mean prevalence of 24% yielded by the maximum entropy pooled prior, all other combined distributions have estimated mean prevalences in the range $[13\%, 16\%]$ for the Beta representation and $[12\%, 16\%]$ for the Gaussian representation. As expected, the marginal priors for φ induced by placing a prior on $\boldsymbol{\alpha}$ and then marginalising (*via* Monte Carlo) yield broader distributions, which encompass the range of all original studies. This effect is slightly more pronounced for the logistic-normal prior, pointing towards more flexibility compared to the Dirichlet.

Table 6: **Mean and credibility intervals for each method for assigning the weights under two representations of information, HIV prevalence example.** Sample size pertains to assigning the weights based on the normalised sample sizes ($\alpha_i = n_i / \sum_{i=0}^{K} n_i$).

	Estimated HIV prevalence	
Method	Beta	Gaussian
Equal weights	0.150 (0.125–0.176)	0.122 (0.099–0.145)
Maximum entropy	0.242 (0.204–0.282)	0.241 (0.202–0.280)
Minimum KL	0.134 (0.107–0.163)	0.160 (0.129–0.192)
Sample size	0.162 (0.137–0.188)	0.132 (0.109–0.155)
Dirichlet prior	0.144 (0.066–0.253)	0.133 (0.063–0.250)
Logistic-normal prior	0.143 (0.060–0.261)	0.138 (0.059–0.259)

4.3 Posterior distribution of the weights: interpretability and prior sensitivity

The analysis of the survival probabilities in Section 4.1 instigates the question of how sensitive posterior inference for $\boldsymbol{\alpha}$ can be to prior specification and the compatibility between the expert opinions and the observed data. In this section we study these questions in simple settings using simulations. We employ four hyperpriors in our analyses: a Dirichlet$(1, 1, 1, 1, 1)$ and a more flexible Dirichlet$(1/10, 1/10, 1/10, 1/10, 1/10)$, along with the corresponding moment-matching logistic-normal priors. All computations are done with the simple Monte Carlo method outlined in the end of section 3.2 and agree closely with solutions using MCMC (not shown).

Central to the discussion here is the notion that the marginal likelihoods – of each expert prior–, when suitably normalised, provide a gold standard for the weights, in the sense that one could do no better when learning the weights conditional on the observed data. We shall refer to the vector of weights obtained by dividing the marginal (integrated) likelihood of each expert by the sum of marginal likelihoods by $\boldsymbol{\alpha}''$. If each prior were a model for the data, then $\boldsymbol{\alpha}''$ would be the model probabilities or the Bayesian model averaging (BMA) weights.

The examples explored here are highly stylised and the marginal likelihood is not usually available for comparison. As such, the discussion presented here should be seen as the analysis of a baseline scenario, where we have control over the ground truth and can more clearly analyse the issues of estimating and interpreting the weights in logarithmic pooling. Note also that the goal of this section is not to address the long run (frequentist) properties of the posterior distribution of the weights – we provide the results of a brief experiment addressing some aspects of posterior concentration under repeated sampling in Figure S4.

Rather, the goal of the following sections is to provide some insight into the potential pitfalls of the interpretation of the posterior weights even in the context of simple models.

4.3.1 Example 1: Beta conjugate analysis

For our first example in this section, we will use $K = 5$ experts, who will elicit Beta distributions about a probability p. Some data (x, n) will then be observed and a likelihood $L(x \mid n, p) = \text{binomial}(n, p)$ will summarise the information brought by the data. In what follows we will elicit the parameters of a Beta distribution on p for each expert using the mean $\mu_i := \mathbb{E}_i[p]$ and coefficient of variation $c_i := \sqrt{\text{Var}_i(p)}/\mu_i$. For more information, please see the appendix of Flavio Codeço Coelho et al. (2015).

The setting we will investigate is when one expert provides a distribution that is significantly more compatible with the data that are ultimately observed. The idea here is to evaluate how the posterior distribution of the weights $p(\boldsymbol{\alpha} \mid x, n)$ supports the "correct" expect as we vary (a) the strength of evidence x/n and (b) the coefficient of variation of the "correct" distribution. As we make the correct expert's coefficient of variation smaller, we expect the posterior weight to increase. The intuition is that if one gives the correct answer with more certainty, one should receive more credence *a posteriori*. To study point (a), we evaluate the posterior weights for $x/n = \{5/10, 50/100, 500/1000, 5000/10000\}$. In addition, we choose the "true" $p = 1/2$ and then construct $\boldsymbol{m} = \{0.1, 0.2, 0.5, 0.8, 0.9\}$ and $\boldsymbol{c} = \{0.1, 0.1, c_2, 0.1, 0.1\}$, where we will vary c_2 between 0.001 and 0.75 in order to study point (b) above. The upper value was chosen such that this is the largest value of c_2 for which the weight of the "correct" expert in $\boldsymbol{\alpha}''$ is the highest weight when $x/n = 5/10$.

In Figure 3, we plot two quantities as a function of the coefficient of variation of the correct expert (c_2) for various levels of evidence (x, n): (i) the ratio between the largest and second largest marginal likelihoods (r_l) computed using (23); and (ii) the ratio between the posterior means of the largest and second largest weights (r_w). The marginal likelihoods (Figure 3a) behave as expected, with r_l diminishing as c_2 increases, the effect more pronounced with increasing the strength of evidence (x/n). We show r_w as function of c_2 in Figure 3b. While for larger values of c_2 the ratio decreases as expected, for low values (high precision) it is also low, attaining a maximum at an intermediate value. This somewhat counter-intuitive result is a quirk of Beta distributions: for low values of c_2, the corresponding parameter values $a_2 = b_2$ are large compared to other (a_i, b_i), which means that any configuration of the weights that assigns non-zero weight to expert 2 is likely to lead to a combined prior that is compatible with the data x/n.

If c_2 is large, the "correct" distribution becomes too diffuse and a different expert is favoured. To convey this in Figure 3b, we interrupt the plotted lines for values of c_2 at which expert 2 was not the one with the highest weight. The correct expert does not attain the largest posterior weight for all values of c_2 for three of the four hierarchical priors considered. The "flexible" logistic-normal prior is the only hyperprior for which expert 2 is consistently favoured for all values of c_2 considered. This phenomenon is similar in nature to identifiability issues in linear mixtures, where components need to be well separated in order for it be possible to reliably recover the mixing proportions (Yakowitz et al. 1968). The results show that the "flexible" logistic-normal hyperprior, i.e., a moment-matching prior to the Dirichlet$(1/10, 1/10, 1/10, 1/10, 1/10)$, circumvents these identifiability problems and allows for better discrimination of the "correct" expert.

The results so far make clear that interpreting the posterior distribution of the weights, in particular the posterior means, is not necessarily trivial or intuitive. We give an explicit example to illustrate the inherent problem of interpreting the weights in a log-linear mixture of beta distributions. Suppose $c_2 = 0.2$ and $c_j = 0.1$ for all $j \neq 2$, with $\boldsymbol{m}$ given as before. This setup leads to $\boldsymbol{a} = \{89.9, 79.8, 12.0, 19.2, 9.1\}$ and $\boldsymbol{b} = \{809.1, 319.2, 12.0, 4.8, 1.01\}$. If the data are $x = 5$ and $n = 10$, computing marginal likelihoods and normalising would lead to weights $\boldsymbol{\alpha}'' = \{0.006, 0.095, 0.710, 0.142, 0.048\}$. However, by calculating $a^{\star\star} = \sum_{i=0}^{K} \alpha_i'' a_i = 19.75$ and $b^{\star\star} = \sum_{i=0}^{K} \alpha_i'' b_i = 44.00$, we see that we obtain a pooled prior with $\mathbb{E}_\pi[p] = 0.31$, far off the "optimal" 1/2. Even in this situation where $r_l \approx 5$, weighting experts according to their marginal likelihoods does not lead to a satisfactory solution. Hence, we argue that there is no hope to reliably learn the weights from these data under this configuration of the expert opinions. If the data were, say, $x = 50, n = 100$, then one would obtain marginal likelihood-based weights such that the pooled "prior" expectation would be $\mathbb{E}_\pi[p] = 0.51$.

4.3.2 Example 2: Gaussian conjugate analysis with known variance

Next, we explore an example where the mean and coefficient of variation setup above is easier to interpret. Consider the problem of drawing inference about the mean μ of a Gaussian distribution with known variance σ^2. Similar to the example above, we investigate the estimation of posterior weights under various scenarios for the expert opinions. For this set of experiments we generated $n = 10, 100, 500$ data points from a Gaussian distribution with $\mu = 3$ and $\sigma^2 = 1$. We then considered $\boldsymbol{m} = \{1, 2, 3, 4, 5\}$ and $\boldsymbol{c} = \{0.1, 0.1, c_2, 0.1, 0.1\}$, where we vary c_2 between 0.001 and 1.5, with this upper bound being chosen similarly to what was done above, in this case for the data set with $n = 10$.

Results (see Figure S3) show the same pattern as Figure 3, with the ratio of weights being small for very low cvs and then increasing with cv until it starts to decay, as expected. This phenomenon is rooted in the same numerical cause as what is observed for the Beta example; for very low cvs, the variance for expert 2 is really small, which in turn makes $w_2 = \alpha_2/v_2$ large for pretty much any value of α_2. This in turn means that many weight configurations $\boldsymbol{\alpha}$ will lead to very similar values of the pooled hyperparameters, m^* and v^* and thus will not receive very different probability *a posteriori*.

4.4 Bayesian melding with varying weights

We now turn our attention to applications of logarithmic pooling to the statistical analysis of deterministic models. In their seminal paper, Poole et al. (2000) lay out Bayesian melding as way to achieve full Bayesian inference for deterministic models – see also Section 2.2 above. In this section we explore two Bayesian melding applications and extend their approach by accommodating uncertainty about the weight α.

4.4.1 Bowhead whale population growth

We begin with the analysis of a non-age-structured population deterministic model (PDM) population model for bowhead whales originally carried out by Poole et al. (2000). The model describes the annual population of bowhead whales in terms of the annual number of whales killed , C_t, the maximum sustainable yield rate (MSYR) and the initial bowhead population (P_0) as:

$$P_{t+1} = P_t - C_t \times \text{MSYR} \times P_t \left(1 - (P_t/P_0)^2\right). \tag{32}$$

One of the quantities of interest in the model was P_{1993}, due to 1993 being the last year for which independent abundance measurements were available, allowing for model calibration. Another important model quantity is the rate of population increase from 1978 to 1993, ROI, defined through

$$P_{1993} = P_{1978}(1 + \text{ROI})^{15}.$$

We are then interested in the model outputs $\phi = \{P_{1993}, \text{ROI}\}$. The key idea is to account for the influence of the priors on the inputs $\theta = \{\text{MSYR}, P_0\}$ on P_{1993} through the **induced** distribution. In particular, we aim at composing the prior distribution

$$\tilde{q}_\Phi(P_{1993}) \propto q_1^*(P_{1993})^\alpha q_2(P_{1993})^{1-\alpha}, \tag{33}$$

where q_1^* is the induced distribution and q_2 is the natural prior on P_{1993}. The main innovation we propose here is to place a probability distribution over α in order to relax the need to fix it to particular value. We choose a Beta$(1, 1)$ prior as our π_A. The target posterior is then

$$p_{\Theta,M}(P_0, \text{MSYR}, \alpha \mid C_t) \propto \tilde{q}_\Theta(P_0, \text{MSYR})L_1(P_0, \text{MSYR})L_2(P_{1993})\pi_A(\alpha), \tag{34}$$

where $\tilde{q}_\Theta$ is the suitably inverted distribution over the input space from the prior over the output space, $\tilde{q}_\Phi$ (see Poole et al. (2000), section 3.3.4). The subscript makes reference to the fact that this is a posterior over the inputs $\theta \in \Theta$ which are linked to the outputs $\phi \in \Phi$ by a deterministic model M, given by (32). Further details on priors and likelihoods are given in Poole et al. (2000) and the Appendix of this paper. We note that when α is random, it is important to include all of the normalising constants that depend on it (Neuenschwander et al. 2009), in particular the normalising constant of the expression in (33).

Here we will consider two ways of approximating (34). First, we used the sampling importance-resampling (SpIR) algorithm described in Appendix A.2.3. This method does not rely on any parametric approximation to the induced distribution q_1^*, instead using standard kernel methods to approximate the density at any point. We used $k = l = 100,000$ iterations to produce a sample from $p_{\Theta,M}$. We also explored a Hamiltonian Monte Carlo (HMC) implementation in Stan (Carpenter et al. 2017). However, for this implementation we needed to approximate q_1^* by a parametric form. Since q_2 is a normal distribution, we approximate q_1^* by a normal distribution such that $\tilde{q}_\Phi$ (Equation 33) can be written in closed-form. We give further discussion on this choice in Appendix A.3. Since $p_{\Theta,M}$ is a challenging target distribution, we used four independent chains of $10,000$ iterations each. We observed a low percentage of divergent iterations (<2%), likely caused by the very challenging posterior geometry induced by high correlations between parameters.

In Figure 4 we show the marginal posteriors for various quantities of interest, obtained with both algorithms and for fixed and varying α. As expected, SpIR are a bit noisier, but distributions are largely the same as obtained by MCMC. For α in particular, despite the ruggedness of distribution obtained with SpIR, the mean and 95% credibility intervals of both distributions match very closely: SpIR = 0.39 (0.02–0.87) and MCMC = 0.40 (0.02–0.91). The high posterior uncertainty about α and the substantial overlap between distributions with fixed and varying α could be explained by the lack of sensitivity of the posterior distribution to α. We confirm this is indeed the case by running SpIR (original algorithm by Poole et al. (2000)) for a few values of α (including the endpoints 0 and 1) and verifying very little difference in the resulting posteriors (Figure S2).

4.4.2 Influenza in a boarding school

Another important class of deterministic models are the ordinary differential equation-based models of disease transmission. Here we will consider such a deterministic epidemic model and how one can draw inference about a key epidemiological quantity, the basic reproductive number, R_0. In 1978, an anonymous source reported an influenza H1N1 epidemic at a small boarding school in England (Anon. 1978). In total, 512 boys out of 763 became ill during the outbreak. Due to the population being isolated and having high rates of contact, many of the assumptions of compartimental epidemic models hold. In particular, the Susceptible-Infected-Removed (SIR) model is a good description of disease spread. The model consists of the system of ordinary differential equations

$$\begin{aligned}
\frac{dS}{dt} &= -\beta SI, \\
\frac{dI}{dt} &= \beta SI - \gamma I, \\
\frac{dR}{dt} &= \gamma I,
\end{aligned}$$

where $S(t) + I(t) + R(t) = 1 \, \forall t$, β is the transmission (infection) rate and γ is the recovery rate. The basic reproductive number is

$$R_0 = \frac{\beta}{\gamma}. \tag{35}$$

The goal is to draw inference about β and γ, and consequently about R_0, from data. Data on the number of infected individuals per time ($Y(t)$) were obtained from the **outbreaks** package (Jombart et al. 2019) and we choose to model the deviation from the ODE solution using log-normal errors, i.e.,

$$L(Y(t) \mid \beta, \gamma, \sigma_I^2) = \text{log-normal}(\mu = \log(I(t)), \sigma_I^2), \tag{36}$$

where $I(t)$ is computed *via* an ODE solver. Here we will consider a situation where one has priors on β and γ, which induce a prior q_1^* on R_0, and also a prior q_2 on R_0 directly. This is the case when, for instance, one wants to make q_2 informative so as to incorporate expert knowledge and/or evidence from previous study. For the priors on β and α we choose commonly used, so-called "uninformative" log-normal priors with parameters $\mu_\beta = \mu_\gamma = 0$ and $\sigma^2_\beta = \sigma^2_\gamma = 1$, which induces a log-normal distribution (q_1^*) on R_0 with parameters $\mu_1 = \mu_\beta - \mu_\gamma$ and $\sigma_1^2 = \sigma^2_\beta + \sigma^2_\gamma$. Using the extensive information gathered by Biggerstaff et al. (2014), we constructed an informative log-normal prior (q_2) with mean 1.5 and variance 0.25^2, which gives $\mu_2 = 0.3917656$ and variance $\sigma_2^2 = 0.1655264$. This leads to a prior credibility interval of (1.070–2.047), which covers most of the estimates (and confidence intervals) of R_0 for Influenza found by Biggerstaff et al. (2014). The target posterior is then

$$p(\beta, \gamma, \alpha \mid Y(t)) \propto L(Y(t) \mid \beta, \gamma, \sigma_I^2) q_1^*(R_0)^\alpha q_2(R_0)^{1-\alpha} \pi_A(\alpha), \tag{37}$$

where we again let π_A be a Beta(1,1) distribution. This setup is convenient because it leads to a closed-form expression for the combined prior on R_0 (see Appendix, Section A.4), while the log-normal priors are flexible and useful in practice. We approximate the posterior in (37) using HMC as described in Appendix A.2.

In Figure 5a we show the posterior distribution of the pooling weight α, which favours high values with a mean and 95% credibility interval of 0.77 (0.21–0.99). The posterior distribution for R_0 obtained by letting α vary and also the resulting distributions of fixing $\alpha = 1/2$ or $\alpha = 1$ are shown in Figure 5b. One can see that fixing $\alpha = 1$ and hence excluding the informative prior leads to a higher estimate of R_0 and fixing $\alpha = 1/2$ as per Poole et al. (2000) leads to the lowest estimates. The solution proposed in this paper, namely assigning α a prior and estimating it from data, leads to an intermediate solution. Fixing $\alpha = 1/2$ also leads to underestimating the measured incidence (Figure 5c), whilst setting $\alpha = 1$ leads to mean predictions that are higher, albeit still underestimating the measured incidence. Again, letting α vary leads to an intermediate solution.

Our results agree somewhat with the estimate obtained by Murray (2002), who finds

$\rho = N/R_0 = 202$ and hence $R_0 = 3.78$, using purely numerical methods with no acknowledgment of uncertainty. The highest estimates we obtained were for fixed $\alpha = 1$, $R_0 = 3.02$ (2.27–3.83). This example showcases a desirable consequence of letting α vary: when the "natural" prior q_2 – which is normally informative – is incompatible with the data, it will receive a lower weight (α closer 1) and hence allow the induced prior (q_1), which is usually more diffuse, to dominate. In fact, as discussed by Biggerstaff et al. (2014), the spread of the 1978 boarding school epidemic is unusually fast when compared to regular seasonal Influenza and was likely caused by the lack of previous exposure of the population to the causing strain, H1N1. The varying α approach makes it possible to deal with such an outlier data set by lowering the influence of the informative prior constructed based on previous studies.

5 Discussion

In this paper we have provided an overview of statistical applications of logarithmic pooling (LP), including a new approach based on assigning a prior measure to the weights. In what follows we discuss our findings in light of the rich literature on log-pooling, as well as point out connections to other parts of the statistical literature on model and forecast aggregation.

5.1 Objections to logarithmic pooling and their counter-arguments

West (1984) argues that LP is strictly theoretically justifiable only when the expert opinions agree. Moreover, LP also violates basic coherence in other respects, for instance when one considers marginalisation or other probability manipulations. Genest and Zidek (1986) (pg. 124) explain, however, that these conclusions stem from the restrictive assumption that the group utility is expressed as a function of the individual utilities. In a statistical application context, the expert opinions are usually employed by an independent decision maker, henceforth called the analyst, and she has her own utility function which can be assumed to not depend on the individual utilities. Another quirk of logarithmic pools is that $\text{supp}(\pi) = \cap_{i=0}^{K} \text{supp}(f_i)$, i.e., the pooled distribution will have the smallest support amongst the distributions being combined. This means a single expert can make large portions of the sample space impossible under the pooled distribution. Again, however, the analyst can use external considerations to exclude an expert whose probability density has too narrow a support.

A consequence of encoding opinions as probability densities is that representations of the same information might have different properties depending on the choice of dominating measure. The results of meta analysis in Section 4.2 make this clear: choosing to represent the information brought by the studies as Beta distribution or a Gaussian does not affect the numerical values of means and probability intervals, but does seem to impact the optimality-based methods for choosing the weights, in particular minimum KL (Table 5). On the other hand, as shown by the agreement of the probability intervals in the bottom of Table 6, moving away from optimality criteria and instead assigning a prior distribution to the weights largely removes dependence on specific choices of probability densities by properly accommodating uncertainty about the weights.

One might worry about being able to learn the weights from data, since the weights depend on the likelihood only indirectly. Indeed, as shown in Section 4.3, it might not always be possible to identify the expert whose opinion is most consistent with the observed data. We argue, however, that this happens mostly in situations where one would not expect to learn much anyway. Consider the situation where the data are highly informative (e.g. Gaussian with small known variance): the likelihood will dominate the prior in the posterior, meaning that most weight configurations will lead to similar (joint) posterior densities. This happens unless there is substantial disagreement between the individual priors being combined and we argue it is a desirable property of LP.

As a final caveat, we note that if interest lies on a multivariate quantity $\theta \in \mathbb{R}^d$, $d > 1$, obtaining the normalising constant $t(\boldsymbol{\alpha})$ will entail computing a high-dimensional integral, which is infeasible to do via quadrature. Here, importance sampling techniques can be leveraged to provide stable and accurate estimates of normalising constants (see Future directions).

5.2 The case for (hierarchical) logarithmic pooling

We shall now argue that properties such as external Bayesianity, relative propensity consistency and log-concavity make logarithmic pooling a powerful tool for the analyst.

Mainly due to the simplicity of their construction, linear mixtures are much more popular in statistical applications (Fruhwirth-Schnatter et al. 2019) than their log-linear cousins. As we hope to have shown in this paper, however, log-linear mixtures (logarithmic pooling) can be as useful or more. External Bayesianity means one does not need to worry about combining the priors first and then obtaining the posterior; one can simply take a set of posterior distributions computed with the same likelihood and combine them.

Moreover, LP preserves log-concavity, which might be crucial in computationally demanding settings where slice sampling, variational or other algorithms that assume log-concavity are employed. In summary, we argue that by employing logarithmic pooling to combine probability densities, the analyst is making the best use of the available information by forming a coherent distribution, that preserves many of the features encoded by the experts in their opinions.

After its theoretical properties, the strongest argument in favour of LP is by far is its adaptability. The extra flexibility brought on by the hierarchical prior on the weights might prove crucial in scientific applications where decision under uncertainty is a regular occurrence. For example, a main strength of Bayesian melding is downweighting parameter values based on implausible model outputs. This strength is magnified by using a hierarchical prior that allows the weight parameter to vary. Indeed, Poole et al. (2000) (Section 5.2) argue that estimating α would be a fruitful path to explore and our results corroborate that view. The result in Section 4.4.2 makes clear the potential of varying-weights Bayesian melding for resolving prior-data conflict. In particular, for protecting the analyst from drawing strong conclusions when the "natural" prior on the quantity of interest is in disagreement with the information brought by the data under analysis.

When comparing the hierarchical prior approach to optimality-based procedures, one might argue that excluding a few or even all experts but one is not problematic since a few experts may, when suitably combined, summarise the information provided by the whole group. Whilst the weights are not probabilities, we argue that it would be preferable to have a solution that respects the so-called Cromwell's rule (Lindley 2013, pg. 91), i.e., not assigning zero probability to events that are logically possible. Here this means allowing for the possibility that the opinion of all experts receives non-zero weight. Incidentally, this should also help alleviate some of the problems discussed in the previous section.

5.3 Future directions

Future research will explore further applications of logarithmic pooling in statistical learning such as combining several posterior predictive distributions from different models fitted to the same data. Techniques such as Bayesian predictive synthesis (BPS, McAlinn et al. (2018), McAlinn, Aastveit, Nakajima, et al. (2019), and McAlinn and West (2019)) and stacking (Yao et al. 2018) have focused on generalising linear pools to combine probabilistic predictions, and logarithmic pooling could be explored as possibility that preserves characteristics such as log-concavity and relative propensity consistency. BPS can include logarithmic pooling as a special case, but understanding the conditions under which this holds remains an open question.

Another interesting avenue for the future is studying the interaction between variable transformations and logarithmic pooling. An example is a situation where one has distributions about a probability p but is interested in the log-odds, $\omega = \log(p/(1-p))$. Should the experts be judged by how reasonable their distributions look in transformed space? How to assign the weights in this situation?

In a practical setting, one might have a collection of MCMC (approximate) samples from different posterior distributions. The statistical question then becomes how to sample from the pooled distribution by re-using these samples. Such task would likely necessitate specially-designed MCMC methods, and would constitute a rich area of future inquiry.

In closing, we hope this paper (i) showcases the usefulness – and potential pitfalls – of logarithmic pooling as a way of combining probability distributions and (ii) entices the statistical community to add it to their toolbox.

Acknowledgments

The authors would like to thank Professors Adrian Raftery, Christian Genest and Mike West, as well as Drs. David Poole, Eduardo Mendes and Felipe Figueiredo 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.

SUPPLEMENTARY MATERIAL

Title: Appendix Proofs and details on algorithms and model choices. (PDF)

A Appendix

A.1 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:

$$\pi(\theta) \propto f_0(\theta) \prod_{j=1}^{K} \left(\frac{f_j(\theta)}{f_0(\theta)} \right)^{\alpha_j}. \tag{38}$$

Let $X_j = \frac{f_j(\theta)}{f_0(\theta)}, j = 1, 2, \ldots, K$. Then, integrating the expression in (38) is equivalent to finding

$$\mathbb{E}_0 \left[\prod_{j=1}^{K} X_j^{\alpha_j} \right] \leq \prod_{j=1}^{K} \mathbb{E}_0[X_j]^{\alpha_j}, \tag{39}$$

where $\mathbb{E}_0[\cdot]$ is the expectation w.r.t f_0 and (39) follows from Hölder's inequality for expectations (Yeh 2011). Since $\forall j$ we have $\mathbb{E}_0[X_j]^{\alpha_j} = \left(\int_{\boldsymbol{\Theta}} f_0(\theta) \frac{f_j(\theta)}{f_0(\theta)}\, d\theta \right)^{\alpha_j} = 1^{\alpha_j}$, Theorem 1 is proven. □

To establish Theorem 2, we will need the following result from Genest et al. (1984).

Lemma 1. ***Representation of a pooling operator with RPC*** *(eq. 3.1). The only relative propensity consistent operator can always be represented by*

$$\mathcal{T}(\boldsymbol{F}_\theta)(\theta) = \boldsymbol{B}(\boldsymbol{F}_\theta)\, c(\theta) \prod_{i=0}^{K} [f_i(\theta)]^{\alpha_i}, \tag{40}$$

with $\boldsymbol{B}(\boldsymbol{F}_\theta) > 0$, $c(\theta) > 0$ *and* $\alpha_0, \alpha_1, \ldots, \alpha_K \geq 0$.

We refer the reader to Genest et al. (1984) for the proof. In short, Lemma 1 is a uniqueness result; logarithmic pooling is the only operator with RPC (Remark 2) and every operator with RPC can be represented as in (40). Now we can state the proof of Theorem 2.

Proof. First, we will show by direct calculation that logarithmic pooling (LP) leads to a log-concave distribution. Notice that each f_i can be written as $f_i(\theta) \propto e^{\nu_i(\theta)}$, where $\nu_i(\cdot)$ is a concave function. We can then write

$$\begin{aligned} \pi(\theta \mid \boldsymbol{\alpha}) &\propto \prod_{i=0}^{K} [\exp(\nu_i(\theta))]^{\alpha_i}, \\ &\propto \exp(\nu^*(\theta)), \end{aligned}$$

where $\nu^*(\theta) = \sum_{i=0}^{K} \alpha_i \nu_i(\theta)$ is a concave function because it is a linear combination of concave functions.

We will now show that LP is the only operator that guarantees log-concavity when $\boldsymbol{F}_\theta$ is a set of concave distributions. First, recall that LP is the only pooling operator that enjoys RPC (Remark 2). With the goal of obtaining a contradiction, suppose that there exists a pooling operator $\mathcal{T}$ which is log-concave but does not enjoy RPC. From Lemma 1, we know that, under this assumption, $\mathcal{T}$ cannot be written in the form $\boldsymbol{B}(\boldsymbol{F}_\theta)c(\theta)\prod_{i=0}^{K} f_i(\theta)^{\alpha_i}$. But every non-negative log-concave function $g(\theta)$ can be represented as

$$g(\theta) = a \cdot c(\theta) \cdot h(\theta), \tag{41}$$

with $a \geq 0$ and $c(\theta)$ and $h(\theta)$ non-negative and log-concave, but otherwise arbitrary. Under the assumptions on $\boldsymbol{F}_\theta$, we have that $h(\theta) := \prod_{i=0}^{K} f_i(\theta)^{\alpha_i}$ is non-negative and log-concave. The only restriction on $c(\theta)$ is that it be positive, it may very well be (log-)concave. Therefore $\mathcal{T}$ can in fact be represented in the form of Lemma 1, contradicting our initial assumption that $\mathcal{T}$ can at same time be log-concave but not enjoy RPC. □

We now move on to the exponential family results in Section 2.1. To see that equation (10) holds:

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

Likewise for equation (11), we have

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

A.2 Computational details

The analyses presented in this paper necessitated numerical optimisation to find the weights based on optimality criteria and Markov chain Monte Carlo (MCMC) to approximate posterior distributions. All computations were carried out in the R (R Core Team 2019) statistical computing environment, version 3.6.0. We provide implementations using the Stan (Carpenter et al. 2017) probabilistic programming language and R code for the methods, figures and tables presented in this paper can also be found at `https://github.com/maxbiostat/opinion_pooling`.

A.2.1 Optimisation procedures

In this section we give more detail on the optimisation procedures used to solve the problems in Sections 3.1.1 and 3.1.2. Since both problems are numerically unstable, we employ an strategy that starts the optimisation routine from $J = 1000$ overdispersed points in the unconstrained space, $\mathbb{R}^K$, obtained through the unit simplex transform (Betancourt 2012). The procedure then picks the overall lowest/highest optimised value in order to avoid local minima/maxima. We draw the initial values from a normal distribution with mean 0 and variance 100^2 and then employ the `optim()` function to optimise the target functions (maximum entropy or minimum KL) using the L-BFGS algorithm (Byrd et al. 1995) with default settings. We then choose the minimum/maximum achieved over the J starting points in order to improve the chances of achieving a global optimum.

A.2.2 Markov chain Monte Carlo *via* Stan

Most of the posterior distributions discussed in this paper cannot be computed in closed-form and therefore we resort to Hamiltonian Monte Carlo (Neal et al. 2011), in particular the No-U-Turn (NUTS) dynamic implementations available in Stan (Hoffman et al. 2014; Betancourt 2017).

For most computations, we employed four independent chains of 4000 iterations each with the first 2000 discarded as warm-up/burn-in. Some models presented more challenging target distributions and we increased the number of iterations to 10000. For all results reported, Monte Carlo error (MCSE) was well below 1% the posterior standard deviation for all parameters, allowing for accurate computation of the relevant expectations. In order to cope with challenging posterior geometry and ensure accurate computation, we used an target acceptance probability of 0.99 (`adapt_delta = 0.99`, in Stan parlance) and up to 2^{15} leapfrog steps (`max_treedepth = 15`). All potential scale reduction factors ($\hat{R}$, Gelman et al. (1992)) were below 1.01, indicating no convergence problems.

A.2.3 Sampling-importance-resampling

In the context of Bayesian melding, while it is possible to employ HMC, it is sometimes preferable to employ custom algorithms that can better deal with the constraints imposed by the deterministic model. In the original paper, Poole et al. (2000, sec. 3.4) propose a sampling-importance-resampling (SpIR) algorithm to sample from the posterior in (13), which we extend here to in order to accommodate varying weights.

0. Draw k values from $q_1(\theta)$, constructing $\boldsymbol{\theta}_k = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)})$;

1. Similarly, sample $\boldsymbol{\alpha}_k$ from $\pi(\boldsymbol{\alpha})$;

2. For each $\theta^{(i)} \in \boldsymbol{\theta}_k$ run the model to compute $\psi^{(i)} = M(\theta^{(i)})$, constructing $\boldsymbol{\phi}_k$;

3. Obtain a density estimate of $q_1^{\star}(\phi)$ from $\boldsymbol{\phi}_k$;

4. Form the importance weights

$$w_i = t\left(\boldsymbol{\alpha}^{(i)}\right)\left(\frac{q_2(M(\theta^{(i)}))}{q_1^{\star}(M(\theta^{(i)}))}\right)^{1-\boldsymbol{\alpha}^{(i)}} L_1(\theta^{(i)})L_2(M(\theta^{(i)})), \tag{42}$$

   where $t\left(\boldsymbol{\alpha}^{(i)}\right) = \left(\int_{\Phi} q_1^{*}(\phi)^{\alpha^{(i)}} q_2(\phi)^{1-\alpha^{(i)}} d\phi\right)^{-1}$ is computed using standard quadrature methods;

5. (Re)Sample l values from $\boldsymbol{\theta}_k$ according to the weights $\boldsymbol{w}_k$.

The quadrature-based normalisation in step in 4 can be replaced with an importance sampling or MCMC estimate when the dimension of either ϕ or θ is large, but this is not explored here.

A.3 Bowhead population growth model: details

For convenience, here we will describe the priors and likelihoods used by Poole et al. (2000) in their analysis of the bowhead whale population model, as well as some of our modelling choices. For P_0 only a shifted gamma prior is available, i.e.,

$$q(P_0) = \frac{b_{P_0}^{a_{P_0}}}{\Gamma(a_{P_0})}(P_0 - s_{P_0})^{a_{P_0}-1} \exp\left(-b_{P_0}(P_0 - s_{P_0})\right),\ P_0 > s_{P_0},$$

with $s_{P_0} = 6400$, $a_{P_0} = 2.8085$ and $b_{P_0} = 0.0002886$. The maximum sustainable yield rate (MSYR) is assigned Gamma prior with parameters $a_{\text{MSYR}} = 8.2$ and $a_{\text{MSYR}} = 372.7$.

The size of the bowhead population in 1993, P_{1993}, is an output of the model for which there are both a prior and a likelihood. The prior (q_2) is a Gaussian distribution with mean $\mu_{1993} = 7800$ and standard deviation $\sigma_{1993} = 1300$, while the likelihood (L_2) is also a Gaussian distribution but with mean $\mu'_{1993} = 8293$ and standard deviation $\sigma'_{1993} = 626$. For the rate of increase (ROI), Poole et al. (2000) use a likelihood that is proportional to $\exp(a + b \times t_8) - 1$, with $a = 0.0302$ and $b = 0.0068$ where t_8 is a random variable with Student t distribution with $\nu = 8$ degrees of freedom. This leads to the density

$$L(\text{ROI} \mid \nu, a, b) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} \left(1 + ((\log(\text{ROI}+1) - a)/b)^2\right)^{-(\nu+1)/2} \frac{1}{|b(\text{ROI}+1)|}, \text{ROI} > -1.$$

For the Stan implementation, we approximate the distribution induced on P_{1993} by the prior on P_0 and MSYR and transformation in (32), q_1^*, by a normal distribution with mean $\mu_{\text{ind}} = 18137.70$ and standard deviation $\sigma_{\text{ind}} = 6146.85$. This step deserves a bit more consideration. As discussed by Poole et al. (2000), q_1^* is very diffuse and likely has heavier tails than a normal distribution. Hence it would make sense also to consider the skew-normal and log-normal families as approximating distributions. On the other hand, we note that approximating q_1^* with a normal distribution allows closed-form computation of the coherised prior $\tilde{q}_{\Phi}(P_{1993})$. In Figure S1 we show the densities of a normal, skew-normal and log-normal distributions fitted to $100,000$ simulations from q_1^* by maximum likelihood. While the skew-normal provides better fit (AIC: 1150814) we do not feel the difference in fit to the normal (AIC: 1152288) justifies the increased technical overhead of not being able to compute the coherised prior in closed-form. Both distributions provide much superior fit than the log-normal (AIC: 1181753).

Note that the sampling-importance-resampling discussed in Section A.2.3 does not necessitate any parametric approximation, employing a density estimation method instead.

In Figure S2 we show the posterior distributions obtained with different values of α using SpIR.

A.4 Pooling of common distributions

In the main text we give the pooled distributions if one assumes a set of Beta (Section 4.1) or Gaussian (Section 4.2) distributions as the expert opinions. In this section we give further results for the pooling of commonly used distributions.

A.4.1 Gamma

Suppose $K+1$ experts are called upon to elicit prior distributions for a quantity $\lambda \in \mathbb{R}^+$. 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

$$\begin{aligned}
\pi(\lambda) &= t(\boldsymbol{\alpha}) \prod_{i=0}^{K} f_i(\lambda; a_i, b_i)^{\alpha_i}, \\
&\propto \prod_{i=0}^{K} \left(\lambda^{a_i - 1} e^{-b_i \lambda}\right)^{\alpha_i}, \\
&\propto \lambda^{a^* - 1} e^{-b^* \lambda}, \qquad (43)
\end{aligned}$$

where $a^* = \sum_{i=0}^{K} \alpha_i a_i$ and $b^* = \sum_{i=0}^{K} \alpha_i b_i$. Noticing (43) is the kernel of a gamma distribution with parameters a^* and b^*, $H_\pi(\lambda)$ becomes

$$H_\pi(\lambda; \boldsymbol{\alpha}) = a^* - \log b^* + \log \Gamma(a^*) + (1 - a^*)\psi(a^*), \qquad (44)$$

where $\psi(\cdot)$ is the digamma function. The Kullback-Leibler divergence between the pooled density π and each density is:

$$\mathrm{KL}(\pi || f_i) = (a_i - a^*)\psi(a_i) - \log \Gamma(a_i) + \log \Gamma(a^*) + a^* \left(\log \frac{b_i}{b^*}\right) + \frac{a_i}{b_i}(b^* - b_i). \qquad (45)$$

A.4.2 Log-normal

Another popular choice for modelling a quantity $\eta \in \mathbb{R}^+$ is the log-normal family. Following the results given in Section 4.2, we know that the pool of log-normal distributions with parameters μ_i and σ_i^2 is a log-normal distribution with parameters $\mu^\star = \frac{\sum_{i=0}^K w_i m_i}{\sum_{i=0}^K w_i}$ and $\sigma^{\star 2} = [\sum_{i=0}^K w_i]^{-1}$, where $w_i = \alpha_i / \sigma_i^2$.

The entropy function is then:

$$
\begin{aligned}
H_\pi(\eta; \boldsymbol{\alpha}) &= \log_2(e) \log\left(\sigma^\star \exp\left(\mu^\star + \frac{1}{2}\right)\sqrt{2\pi}\right), \\
&= \log_2(e)\left[\log\left(\sigma^\star\right) + \mu^\star + \frac{1}{2} + \log(\sqrt{2\pi})\right]. \qquad (46)
\end{aligned}
$$

The KL divergence evaluates to

$$
\mathrm{KL}(\pi || f_i) = \frac{1}{2\sigma_i^2}\left[(\mu^\star - \mu_i)^2 + \sigma^{\star 2} - \sigma_i^2\right] + \log\left(\frac{\sigma_i^2}{\sigma^{\star 2}}\right). \qquad (47)
$$

A.4.3 Poisson

If the quantity of interest is a count $y = 0, 1, \ldots,$ and $\boldsymbol{F}_y$ is a set of Poisson distributions with rate parameters $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \ldots, \lambda_K\}$. We have

$$
\begin{aligned}
\pi(y) &\propto \prod_{i=0}^K \left(\frac{\lambda_i^y}{y!}\right)^{\alpha_i}, \\
\pi(y) &= \frac{\exp(-\lambda^\star)\lambda^{\star y}}{y!}, \text{ with } \lambda^\star = \prod_{i=0}^K \lambda_i^{\alpha_i}. \qquad (48)
\end{aligned}
$$

The entropy of the pooled distribution is

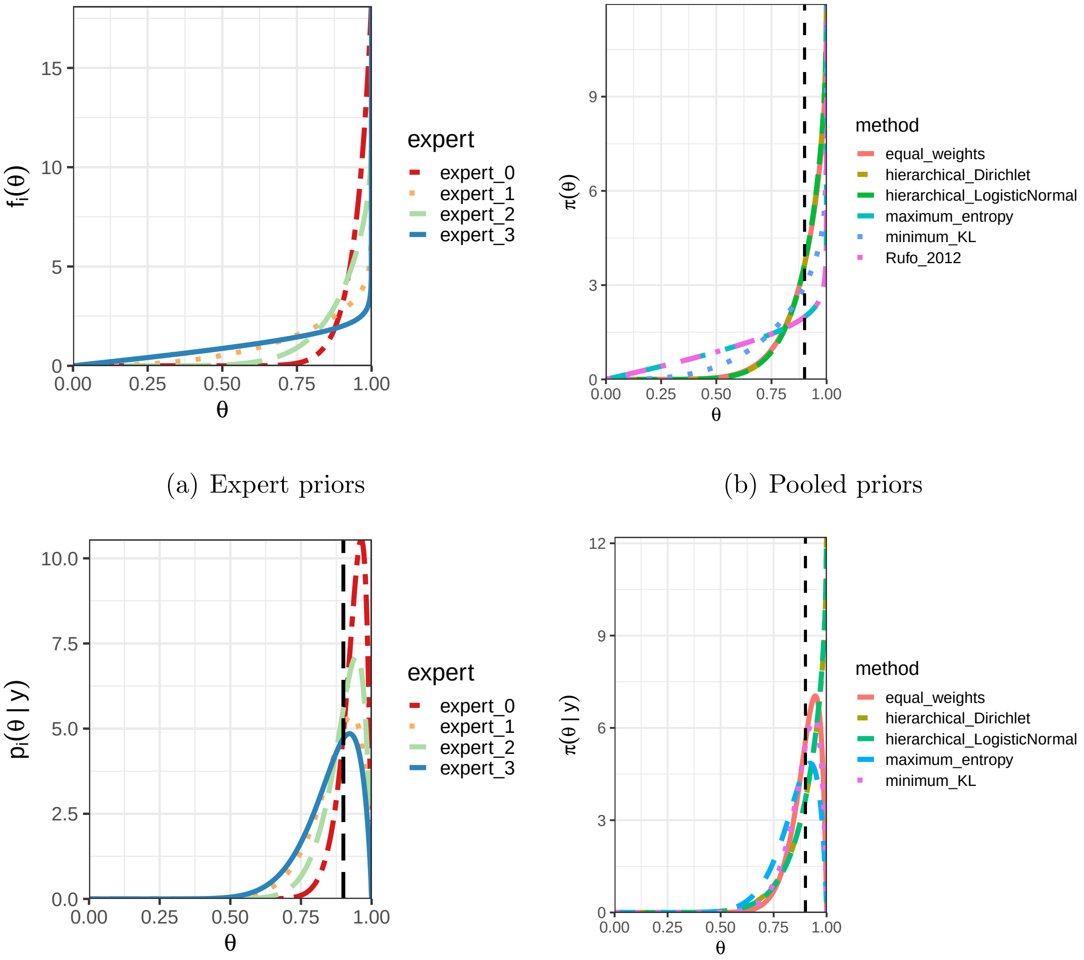
$$
H_\pi(y; \boldsymbol{\alpha}) = -\lambda^\star \log\left(\frac{\lambda^\star}{e}\right) + E_\pi\left[\log(k!)\right], \qquad (49)
$$

where the latter term cannot be evaluated in closed-form, but efficient approximations exist (Evans et al. 1988).

The KL divergence is

$$\begin{aligned} \mathrm{KL}(\pi || f_i) &= \lambda^\star \log\left(\frac{\lambda^\star}{\lambda_i}\right) + \lambda_i - \lambda^\star, \\ &= \lambda^\star \left[\sum_{k=0}^{K} \alpha_k \log(\lambda_k) - \log(\lambda_i)\right] + \lambda_i - \lambda^\star. \end{aligned} \tag{50}$$

A.5 Supplementary figures



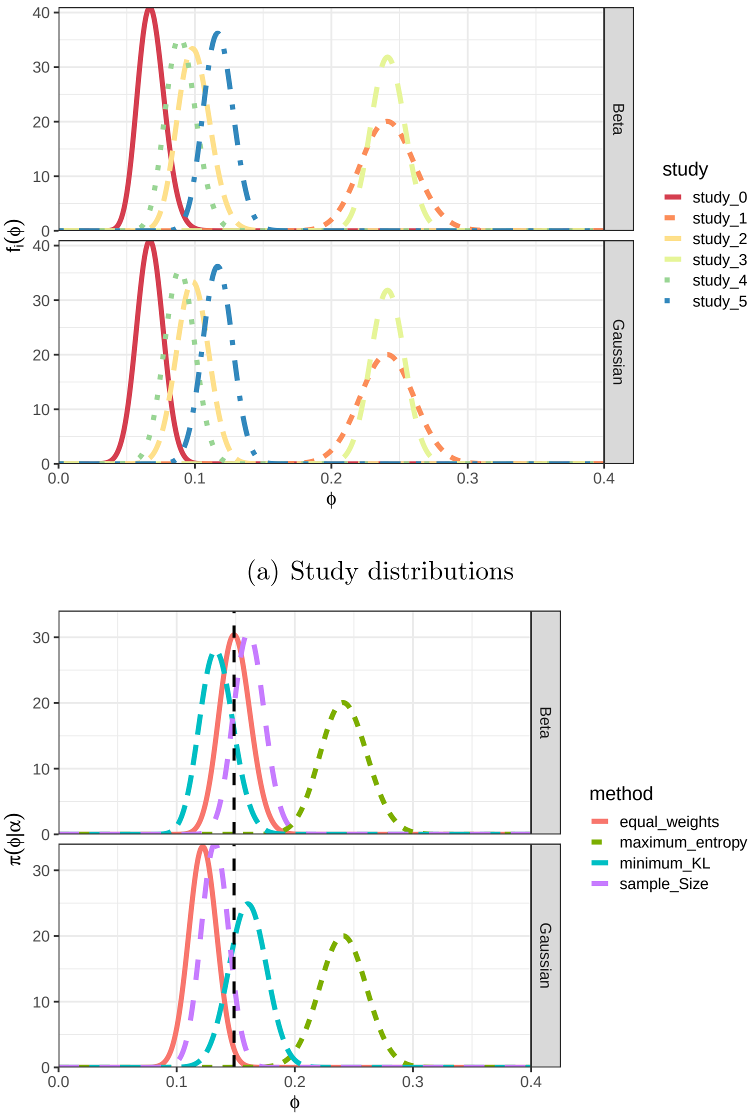
(a) Expert priors

(b) Pooled priors

(c) Expert posteriors

(d) Pooled posteriors

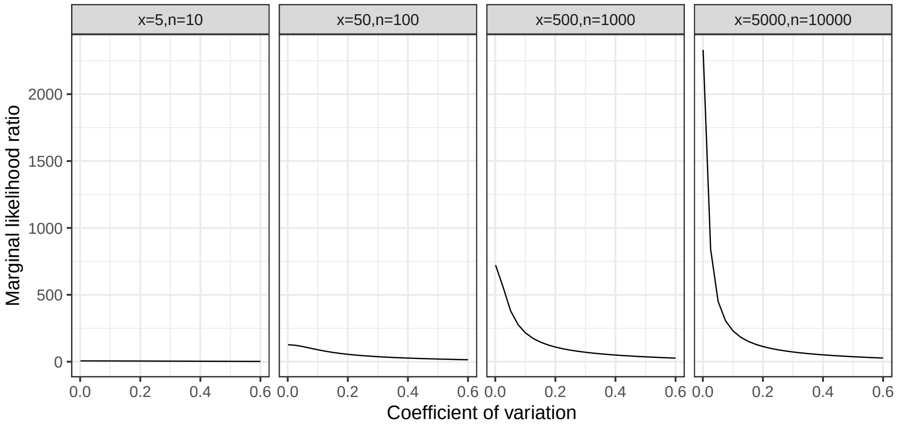
Figure 1: **Prior and posterior densities for the survival probability** θ. Panel (a) shows the distributions elicited by each expert (data from Savchuk et al. (1994)) and panel (b) shows the pooled priors and posteriors obtained using the methods discussed in this paper and the solution found by Rufo, Pérez, et al. (2012). Panels (c) and (d) show the corresponding plots for the posterior distributions after observing the data, $y = 9, n = 10$. The black dashed vertical line marks the maximum likelihood estimate $\hat{\theta} = 9/10$.



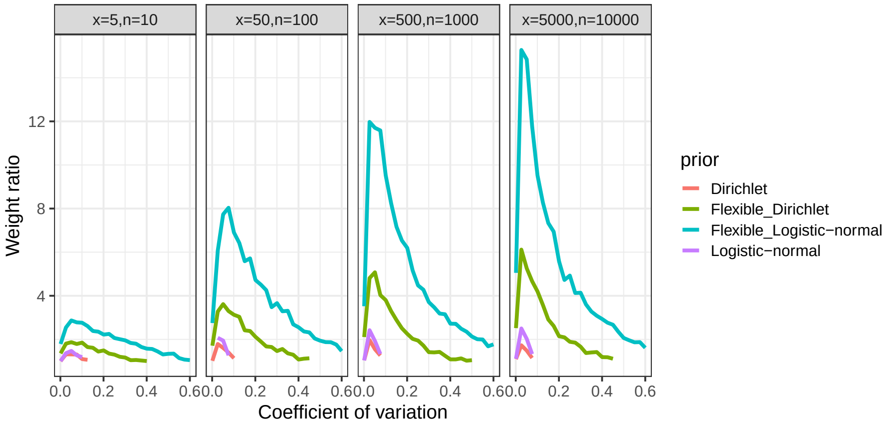
(a) Study distributions

(b) Pooled distributions

Figure 2: **Densities for the HIV prevalence among MSM φ using Beta and Gaussian distributions**. Panel (a) shows the distributions obtained from each study (see Table 4) and panel (b) shows the pooled distributions obtained using the methods discussed in this paper. Vertical tiles show the distribution of choice (Beta or Gaussian) and the vertical dashed black line in panel (b) shows the estimate obtained by combining all studies, $\hat{\varphi} = \sum_{i=0}^{K} y_i / \sum_{i=0}^{K} n_i$.



(a) Ratios of marginal likelihoods



(b) Ratios of posterior weights

Figure 3: **Marginal likelihood and weight ratios for the simulated situation with one correct expert, various strengths of evidence**. Panel (a) shows the ratio between the largest and second largest marginal likelihoods (r_l) as the correct expert's coefficient of variation (c_2) changes, while panel (b) shows the ratio between the largest and second largest posterior mean weights (r_w) in the same settings. Vertical tiles show the observed data and colours in panel (b) show the hyperprior on $\boldsymbol{\alpha}$. "Flexible" priors are a Dirichlet(1/10, 1/10, 1/10, 1/10, 1/10) and the corresponding moment-matching logistic-normal. We interrupt the lines for values of c_2 for which expert 2, the correct expert, does not attain the largest posterior weight (see text).

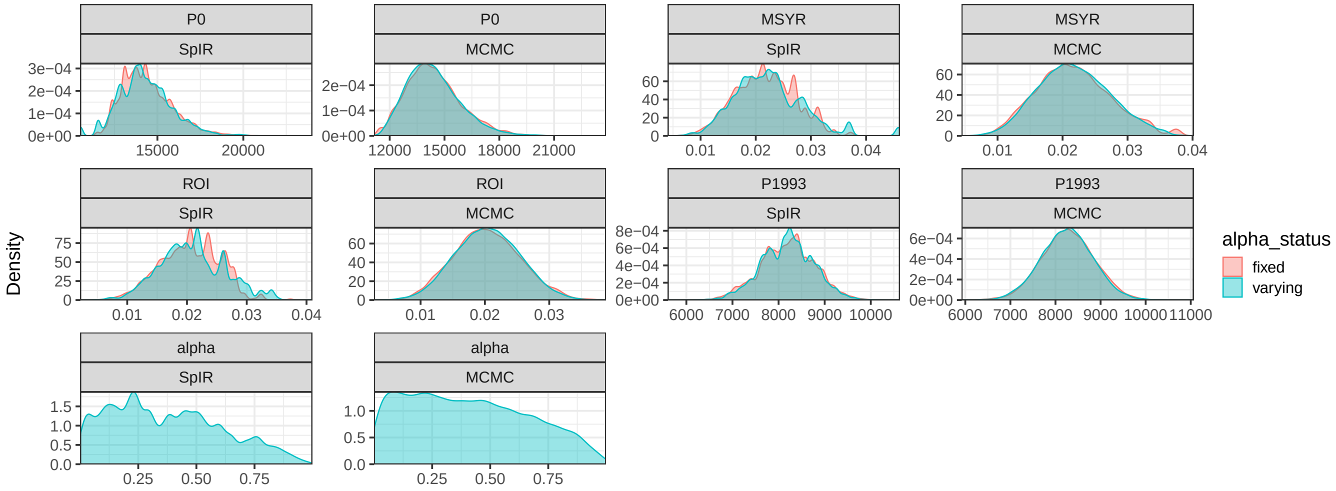
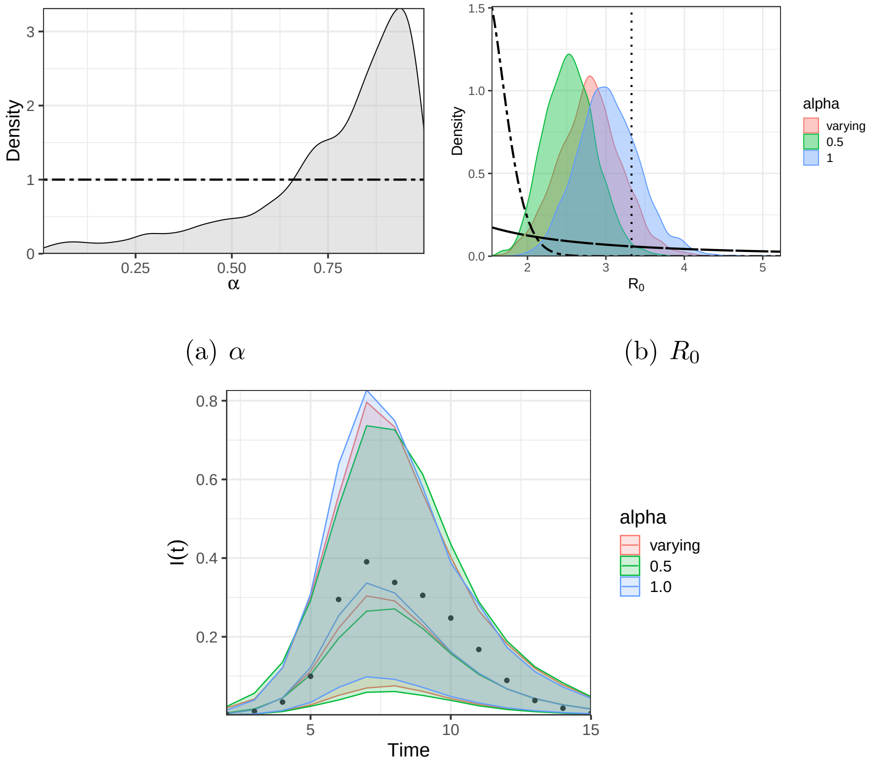


Figure 4: **Marginal posterior distributions for various quantities of interest in the bowhead population model**. We show the posterior distributions obtained by using sampling importance-resampling (SpIR) and Markov chain Monte Carlo (HMC-MCMC), for fixed $\alpha = 1/2$ and placing a prior π_A on α ("varying").



(a) α

(b) R_0

(c) Predictions

Figure 5: **Estimates of the pooling weight (α), the basic reproductive number (R_0) and predictions of the number of infected individuals**. The posterior distribution for the pooling weight α is shown in panel (a), where the horizontal dashed line shows the prior density, a Beta(1, 1). Panel (b) shows the posterior distribution for R_0 obtained with estimating α ("varying") or fixing it to either 1/2 or 1. Vertical line shows $R_0 = 3.78$ (Murray 2002), the dot-dashed line shows the informative prior q_2 and the long-dash line shows the induced prior q_1^*. In panel (c) we show the posterior mean and 95% credibility intervals for the proportion of infected individuals, again by either letting α vary or fixing it to either 1/2 or 1.

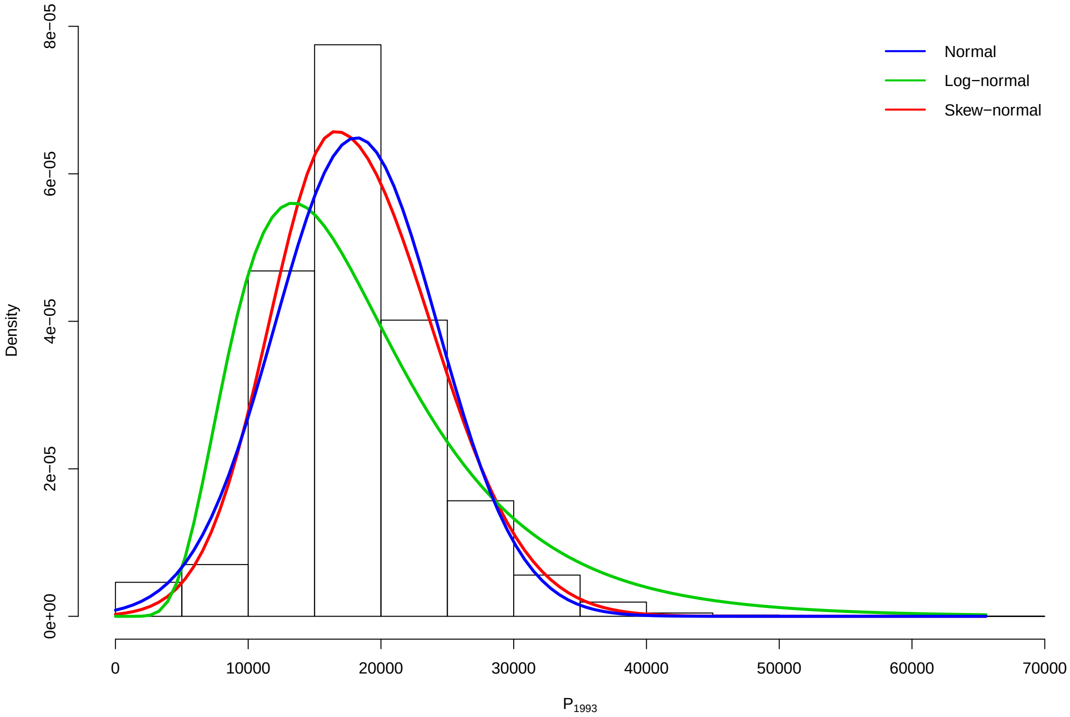


Figure S1: **Induced distribution on P_1993 (q_1^*) and approximating distributions**. We present the histogram of $100,000$ simulations from the prior. Lines show the densities of the three distributions considered.

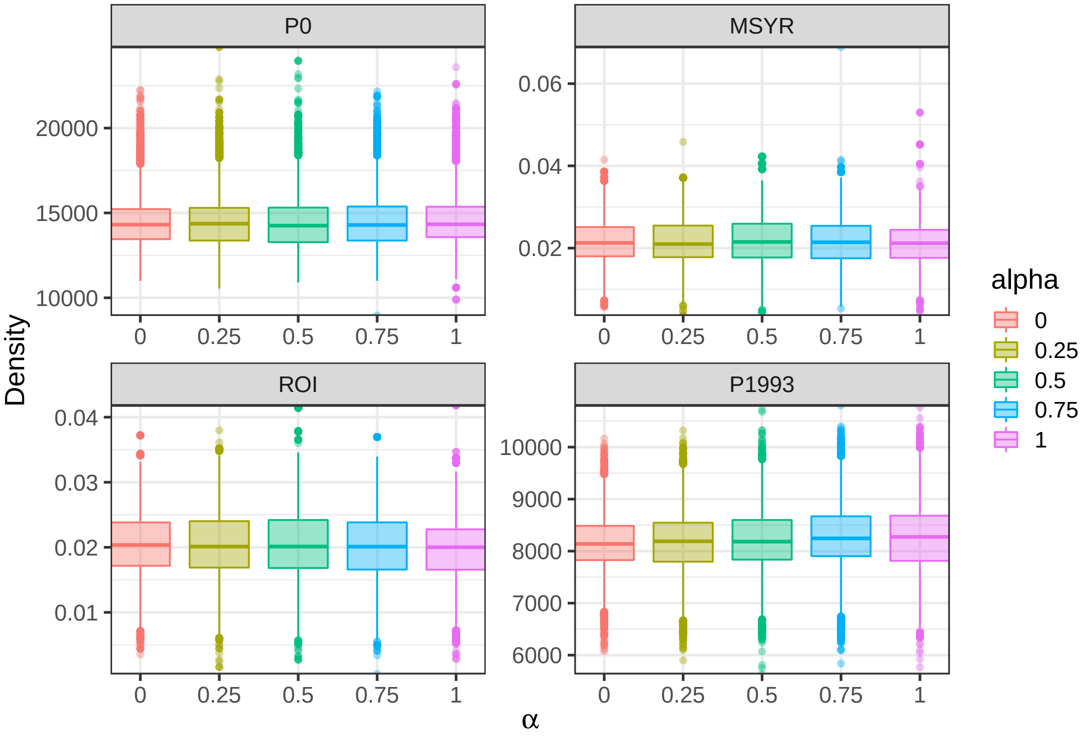
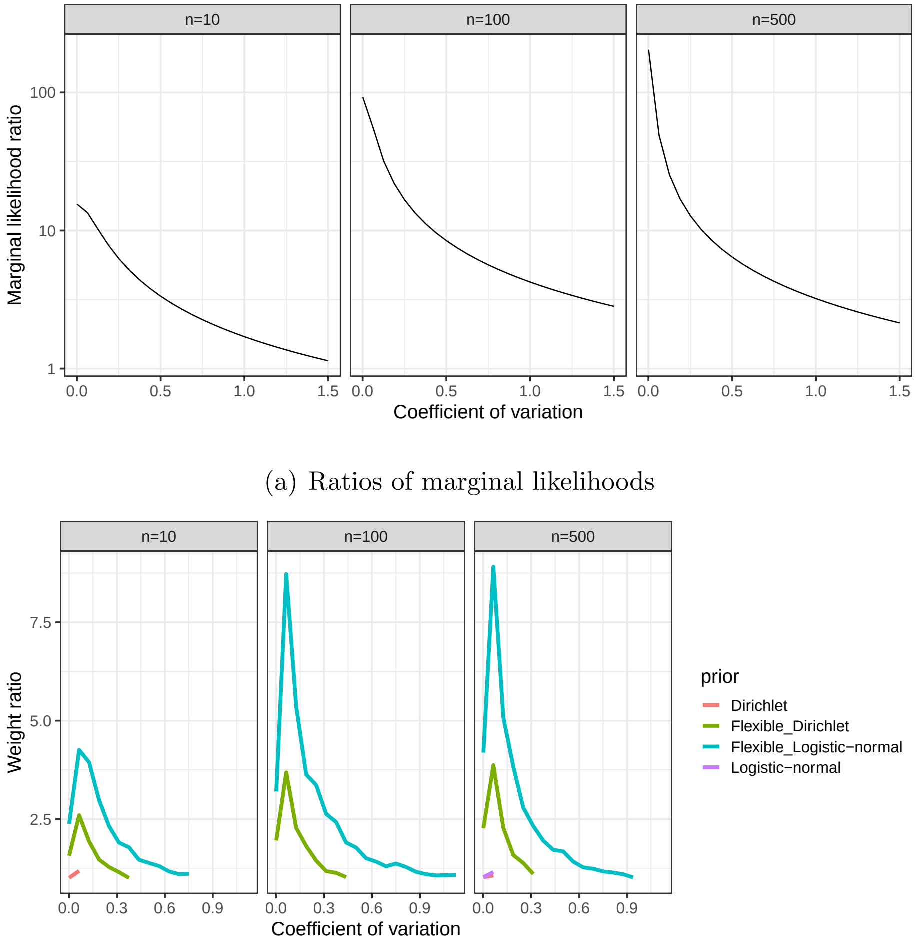


Figure S2: **Sensitivity of posterior inferences to varying the value of α, bowhead population model**.

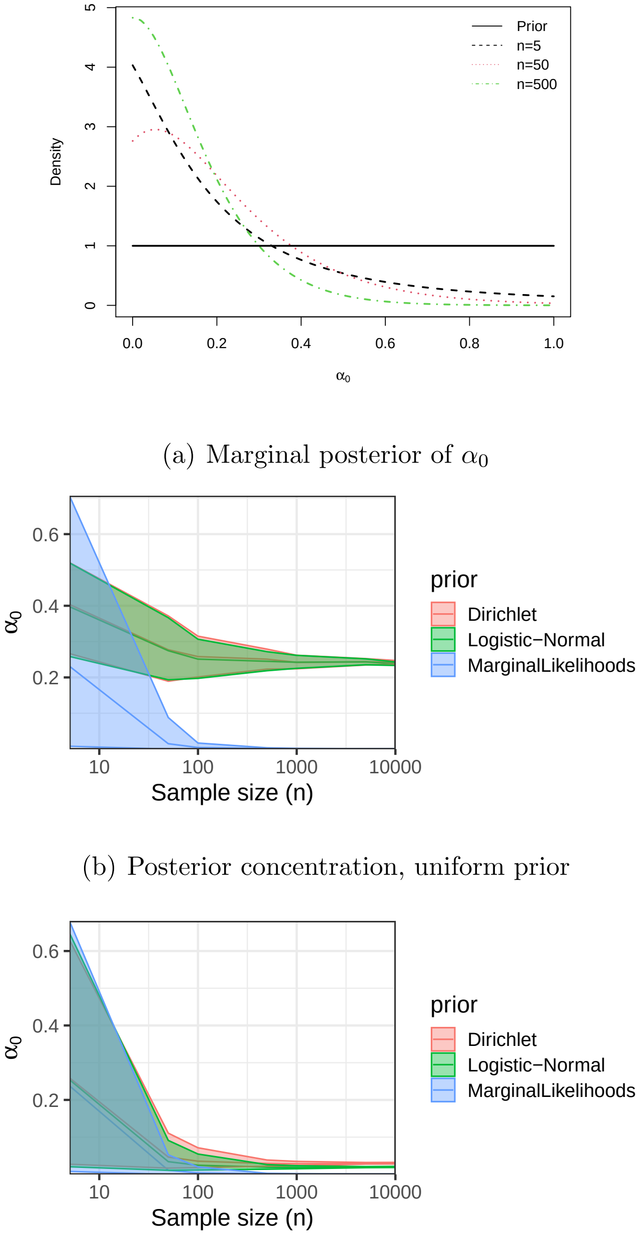
We show the posterior distributions obtained by SpIR for P_0, MSYR, ROI and P_1993 as we fix α to different values.



(a) Ratios of marginal likelihoods

(b) Ratios of posterior weights

Figure S3: **Marginal likelihood and weight ratios for the simulated situation with one correct expert, various strengths of evidence, Gaussian example**. Panel (a) shows the ratio between the largest and second largest marginal likelihoods (r_l) as the correct expert's coefficient of variation (c_2) changes, while panel (b) shows the ratio between the largest and second largest posterior mean weights (r_w) in the same settings. Vertical tiles show the observed data and colours in panel (b) show the hyperprior on $\boldsymbol{\alpha}$. "Flexible" priors are a Dirichlet(1/10, 1/10, 1/10, 1/10, 1/10) and the corresponding moment-matching logistic-normal. We interrupt the lines for values of c_2 for which expert 2, the correct expert, does not attain the largest posterior weight (see Section 4.3.2).



(a) Marginal posterior of α_0

(b) Posterior concentration, uniform prior

(c) Posterior concentration, Beta$(1/10, 1/10)$ prior

Figure S4: **Posterior concentration for the weights in a two-expert setting, Gaussian conjugate analysis with known variance** (σ^2). In this experiment we set the expert hyperparameters to $m_0 = 1$ $v_0 = (1/4)^2$, $m_1 = 2$, $v_1 = (1/2)^2$ and the true data-generating parameters at $\mu = 2$ and $\sigma^2 = 1^2$. In panel (a) we show the marginal posterior of α_0 for various sample sizes under a Beta prior for α_0 with parameters $a = b = 1$. Panel (b) shows the posterior mean of α_0 versus sample size under repeated sampling, using 100 simulated data sets per sample size. Bands correspond to the 95% quantiles of the sampling distribution of the posterior mean. In this experiment the prior on α_0 is a uniform prior over $(0, 1)$. For comparison we show the BMA weight for expert 0, computed from the