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

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

June 30, 2019

Abstract

Key-words: logarithmic pooling; expert opinion; hierarchical modelling; Bayesian melding.

1 Introduction

Combining probability distributions is a topic of general interest, both in the statistical (West, 1984; Genest et al., 1986; Genest and Zidek, 1986) and decision theory literatures (Genest et al., 1984; French, 1985; 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 et al., 1986) (see **Background**). In a practical setting, logarithmic pooling finds use in a range of fields, from engineering (Lind and Nowak, 1988; Savchuk and Martz, 1994) to wildlife conservation (Poole and Raftery, 2000) and infectious disease modelling (Coelho and Codeço, 2009).

A common situation of interest is combining expert opinions about a quantity of interest $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ when they 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). 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 et al., 2012a,b).

These approaches, while moving away from the problem of arbitrarily assigning the weights, arrive at single point solutions, similar to point estimates in Statistical theory. Albeit acknowledging that these approaches have merit, we argue that in many settings, where one has substantial prior information on the relative reliabilities of the information sources (experts), it would be desirable to incorporate this information into the pooling procedure while accommodating uncertainty about the weights. 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 and Raftery, 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. In Section 3 we present different approaches to choose 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 4.2), meta-analysis (Section 4.3) and Bayesian melding (Section 4.4).

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 $\mathbf{F}_\theta := \{f_0(\theta), f_1(\theta), f_2(\theta), \ldots, f_K(\theta)\}$ be a set of (densities of) prior distributions representing the opinions of $K+1$ experts and let $\boldsymbol{\alpha} := \{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_K\}$ be the vector of weights, such that $\alpha_i > 0 \, \forall i$ and $\sum_{i=0}^K \alpha_i = 1$. The log-pooled prior 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 $t(\boldsymbol{\alpha}) = \int_{\boldsymbol{\Theta}} \prod_{i=0}^K f_i(\theta)^{\alpha_i} d\theta$.

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

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

A novel, simple proof using Hölder's inequality is given 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, 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. The proof of this fact is trivial and is given here for completeness. 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 $\sum_{i=0}^K \alpha_i = 1$. □

Much less trivially, 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, 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. 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.

Remark 3. ***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.*

We give a proof in the Appendix.

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. Suppose we are interested in a random variable, Y, from an 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:

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

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 the Appendix).

2.1.1 Conjugate priors to the exponential family

A conjugate prior family for $f(y|\theta)$, given in (5), is the following

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

where $K(a,b)$ is a normalising constant. Similar to the above, let $\mathbf{G}_\theta$ be a set of 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 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 and Raftery (2000). Deterministic simulation models are widespread in Science and Engineering (see Poole and Raftery (2000) and references therein). One is often interested in a deterministic model M with inputs $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and outputs $\phi \in \boldsymbol{\Phi} \subseteq \mathbb{R}^q$, 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 and Raftery (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 acknowledgement of uncertainty is crucial.

In Poole and Raftery (2000) (Section 6.2), 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 the case of their Bayesian melding analysis one is combining distributions based on different bodies of evidence, but assessed by the same expert. Since this need not be the case in all applications, in this paper we relax this restriction, modelling α through a hyperprior instead.

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 and Raftery (2000); Rufo et al. (2012a,b) 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

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 uninformative. Such approach would ensure that, given the constraints imposed by $\mathbf{F}_\theta$, the pooled distribution is the one which best represents the current state of knowledge (Jaynes, 1957; Savchuk and Martz, 1994). In order to choose $\boldsymbol{\alpha}$ so as to maximise prior diffuseness, one can maximise the entropy of the log-pooled prior, i.e.

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

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 4 for intuition as to why.

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 $\mathbf{F}_\theta$. We can define a loss function such that

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

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.

Remark 4. ***Uniqueness*** *(Rufo et al., 2012a). The distribution obtained following (17) is unique, i.e., there is only one aggregated prior $\pi(\theta \mid \boldsymbol{\alpha})$ that minimizes $L(\boldsymbol{\alpha})$.*

One can get some intuition into the proof by noting that minimising (16) is equivalent to maximising $\ln t(\boldsymbol{\alpha}) = \ln \int_{\boldsymbol{\Theta}} \prod_{i=0}^{K} f_i(\theta)^{\alpha_i} d\theta$. Rufo et al. (2012a) show that $t(\boldsymbol{\alpha})$ is concave, therefore the problem in (17) has a unique solution. By contrast, the problem in (15) requires one to minimise $\ln t(\boldsymbol{\alpha})$, hence lacking a sufficient condition for the existence of a unique solution. See Section 3.3.1 for implementation details.

3.2 Hierarchical modelling of the weights

As discussed by Poole and Raftery (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 and Shen, 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 and Shen, 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,

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

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).

3.3 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 (Stan Development Team, 2014) 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`.

3.3.1 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.

3.3.2 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 and Gelman, 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 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 and Rubin (1992)) were below 1.01, indicating no convergence problems.

3.3.3 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 and Raftery, 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, construct $\boldsymbol{\alpha}_k$ 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

$$w_i = t\left(\boldsymbol{\alpha}^{(i)}\right)\left(\frac{q_2(M(\theta^{(i)}))}{q_1^\star(M(\theta^{(i)}))}\right)^{1-\boldsymbol{\alpha}^{(i)}} L_1(\theta^{(i)})L_2(M(\theta^{(i)})), \tag{24}$$

   where $t\left(\boldsymbol{\alpha}^{(i)}\right) = \int_\Phi q_1^*(\phi)^{\alpha^{(i)}} q_2(\phi)^{1-\alpha^{(i)}} d\phi$ is computed using standard quadrature methods;
5. (Re)Sample l values from $\boldsymbol{\theta}_k$ according to the weights $\boldsymbol{w}_k$.

4 Applications

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 and Martz (1994) (also discussed in Rufo et al. (2012b)) 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 and Martz (1994) then use to construct prior distributions with maximum variance given the restriction on the means. From the vector of prior means $\mathbf{m} = \{m_0 = 0.95, m_1 = 0.80, m_2 = 0.90, m_3 = 0.70\}$, the authors obtain the parameters of the Beta distributions for each expert, $\mathbf{a} = \{a_0 = 18.10, a_1 = 3.44, a_2 = 8.32, a_3 = 1.98\}$ and $\mathbf{b} = \{b_0 = 0.955, b_1 = 0.860, b_2 = 0.924, b_3 = 0.848\}$. 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 4.2 for a justification).

The probability density of the survival probability for each expert is given by

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

where $\mathcal{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the Beta function. The log-pooled distribution for θ is then

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

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

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

Table 1 lists the weights proposed by the methods discussed above, as well as the solution found by Rufo et al. (2012b) (Section 5.2). We note that the weights from maximum entropy, minimum KL divergence and using the method by Rufo et al. (2012b) use only information contained in the collection of expert distributions $\boldsymbol{F}_\theta$, while the results for the hierarchical priors are posterior means obtained after observing data. Maximising the entropy of the pooled distribution lead to the degenerate solution $\boldsymbol{\alpha} = \{0, 0, 0, 1\}$, which gives all the weight to the most diffuse prior distribution – 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 et al. (2012b) 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 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

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

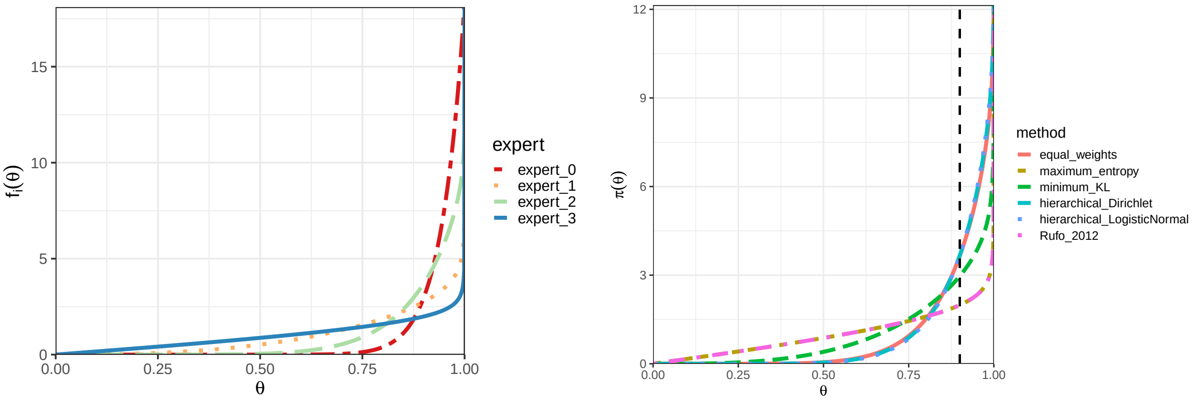
Method	α_0	α_1	α_2	α_3
Maximum entropy	0.00	0.00	0.00	1.00
Minimum KL[1] divergence	0.04	0.96	0.00	0.00
Rufo et al. (2012a)	0.00	0.00	0.00	1.00
Dirichlet prior[2]	0.26	0.24	0.27	0.23
Logistic-normal prior[2]	0.27	0.24	0.31	0.18

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 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. If one were to normalise the marginal likelihoods in order to form the expert weights, one would find $\boldsymbol{\alpha}' = \{0.27, 0.24, 0.30, 0.19\}$, which is not very different from equal weights. The posterior distribution for the weights estimated under both priors favours expert 1, the expert with the highest marginal likelihood. The logistic-normal gives expert 1 a higher weight when compared with the Dirichlet. This is connected to the increased flexibility of the logistic-normal (see Section 4.2).

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 that 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, for instance.



(a) Expert priors

(b) Pooled priors

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

Table 2: **Prior and posterior mean and credibility intervals for each method for assigning the weight, survival probability example (Savchuk and Martz, 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 et al. (2012b)	0.70 (0.17–0.99)	0.86 (0.63–0.98)
Dirichlet prior	0.86 (0.40–1.00)	0.89 (0.70–0.99)
Logistic-normal prior	0.88 (0.35–1.00)	0.89 (0.71–0.99)

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

4.2 Posterior distribution of the weights: interpretability and prior sensitivity

The analysis of the survival probabilities example 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 using simulated data. Throughout 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 Coelho and Carvalho (2015). We employ four hyperpriors in our analysis: 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.

The setting we will investigate is when one expert provides a distribution that is significantly more compatible with the data 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 support 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 $\boldsymbol{\mu} = \{0.1, 0.2, 0.5, 0.8, 0.9\}$ and $\boldsymbol{c} = \{0.1, 0.1, c_2, 0.1, 0.1\}$, where we will vary c_2 between 0.001 and 0.6 in order to study point (b) above.

In Figure 2, 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 2a) 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 2b. 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 appreciable 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 2b, 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 one 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. 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 "correct" expert.

In conclusion, the results of this section make clear that interpreting the posterior distribution of the weights, in particular the posterior means, is not 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 we let $c_2 = 0.2$ and $c_j = 0.1$ for all $j \neq 2$, with $\boldsymbol{\mu}$ 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 this data for this configuration of the expert opinions. If the data were, say, $x = 50, n = 100$, then one would obtain a pooled prior for which $\mathbb{E}_\pi[p] = 0.51$.

4.3 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 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.

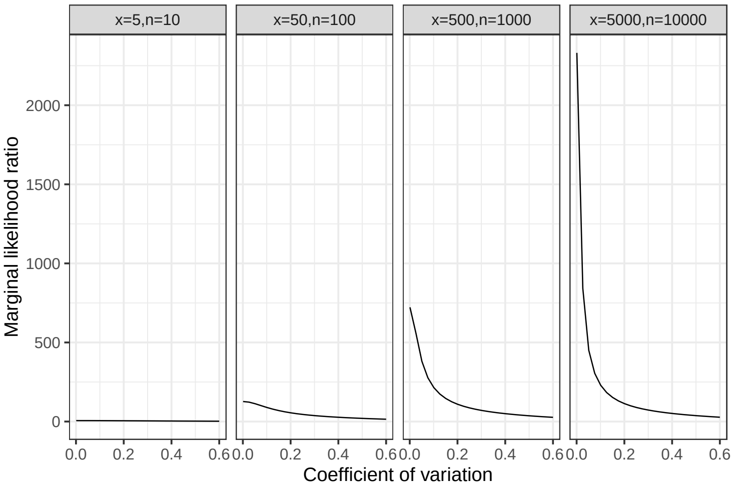
Assuming a uniform prior for the HIV prevalence among MSM, denoted by φ, and a binomial model for each study, i.e. $Y_i \sim \text{Binomial}(n_i, \varphi)$. The posterior distribution for the HIV prevalence conditional on each study is then a Beta distribution with parameters $a_i = y_i + 1$ and $b_i = n_i - y_i + 1$, 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

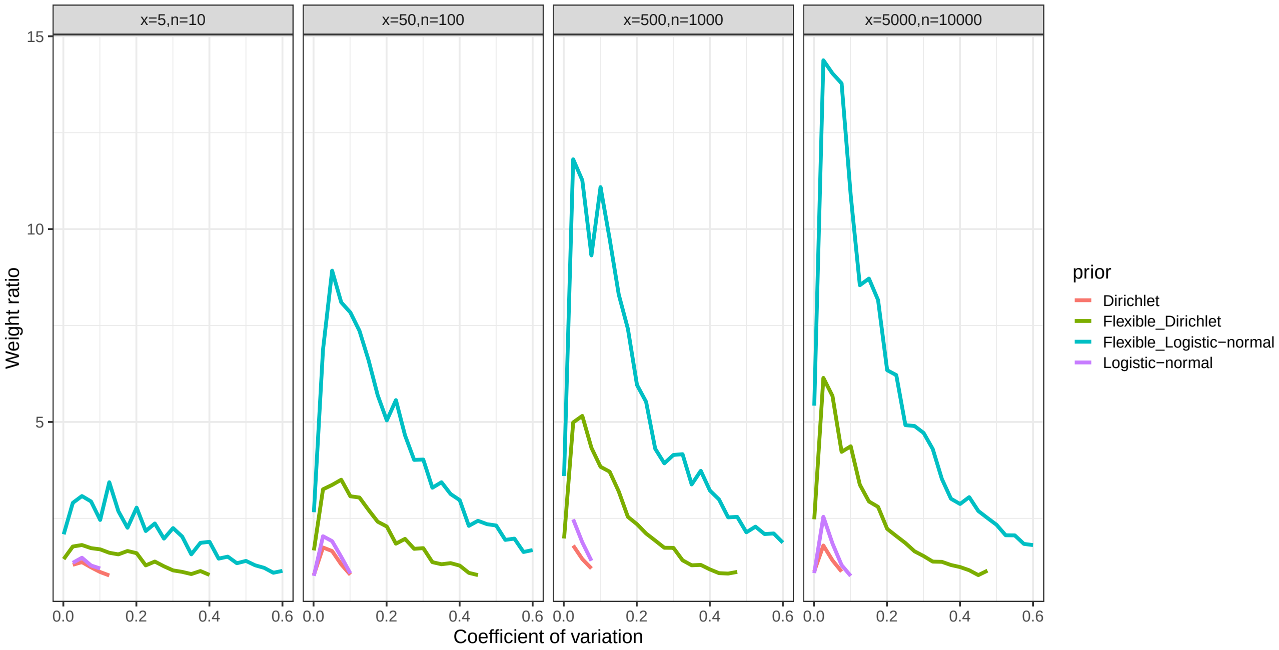
$$f_i(\varphi; m_i, v_i) = \frac{1}{\sqrt{2\pi v_i}} \exp\left(\frac{-(\varphi - m_i)^2}{2v_i}\right),$$

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

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



(a) Marginal likelihoods



(b) Weights

Figure 2: **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).

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

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

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

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

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).

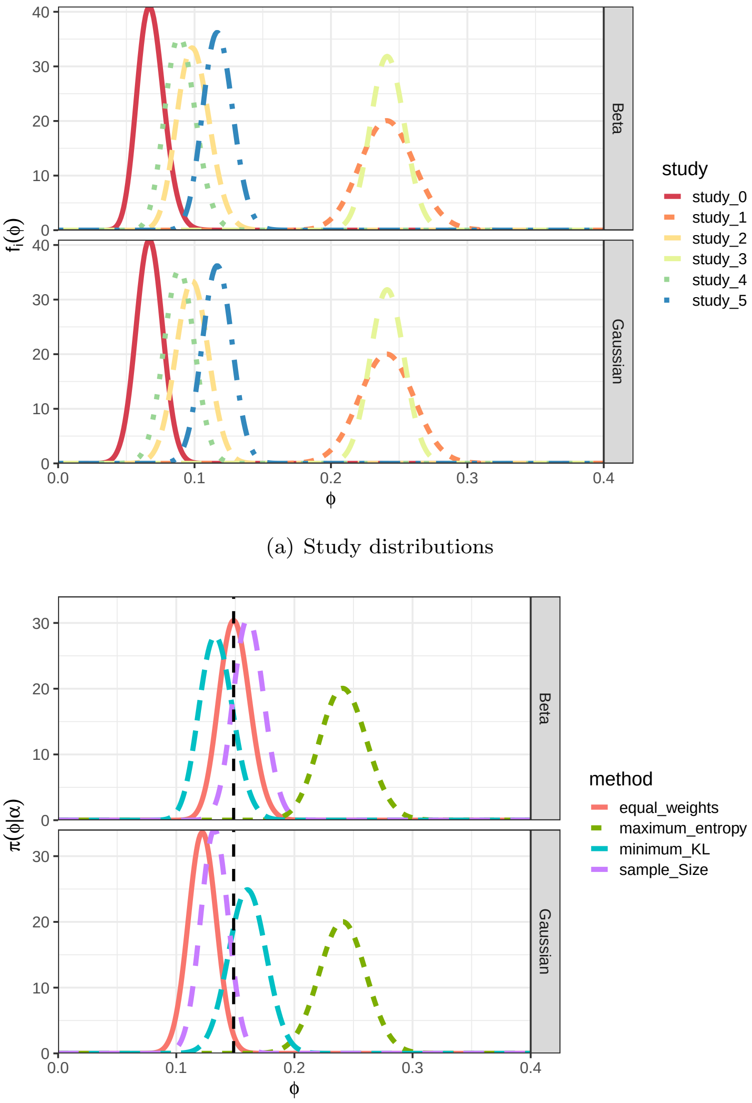
				Estimated prevalence, φ (95% CI)	
Study	Reference	n	y	Beta	Gaussian
0	Tun et al. (2008)	658	44	0.068 (0.050–0.089)	0.067 (0.048–0.086)
1	Barcellos et al. (2003)	461	111	0.242 (0.204–0.282)	0.241 (0.202–0.280)
2	Carneiro et al. (2003)	621	61	0.100 (0.077–0.124)	0.098 (0.075–0.122)
3	Sutmoller et al. (2002)	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 3 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.



(a) Study distributions

(b) Pooled distributions

Figure 3: **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, $\hat{\varphi} = \sum_{i=0}^{K} y_i / \sum_{i=0}^{K} n_i$.

4.4 Bayesian melding with varying weights

In their seminal paper, Poole and Raftery (2000) lay out the Bayesian melding as way to achieve full Bayesian inference for deterministic models – see also Section 2.2 above. In this section we explore

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)

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-aged-structured population deterministic model (PDM) population model for bowhead whales originally carried out by Poole and Raftery (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 and Raftery (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 and Raftery (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 Section 3.3.3. 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 an HMC implementation in Stan. 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 the Appendix. 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 and Raftery (2000)) for a few values of α (including the endpoints 0 and 1) and verifying very little difference in the resulting posteriors (Figure S2).

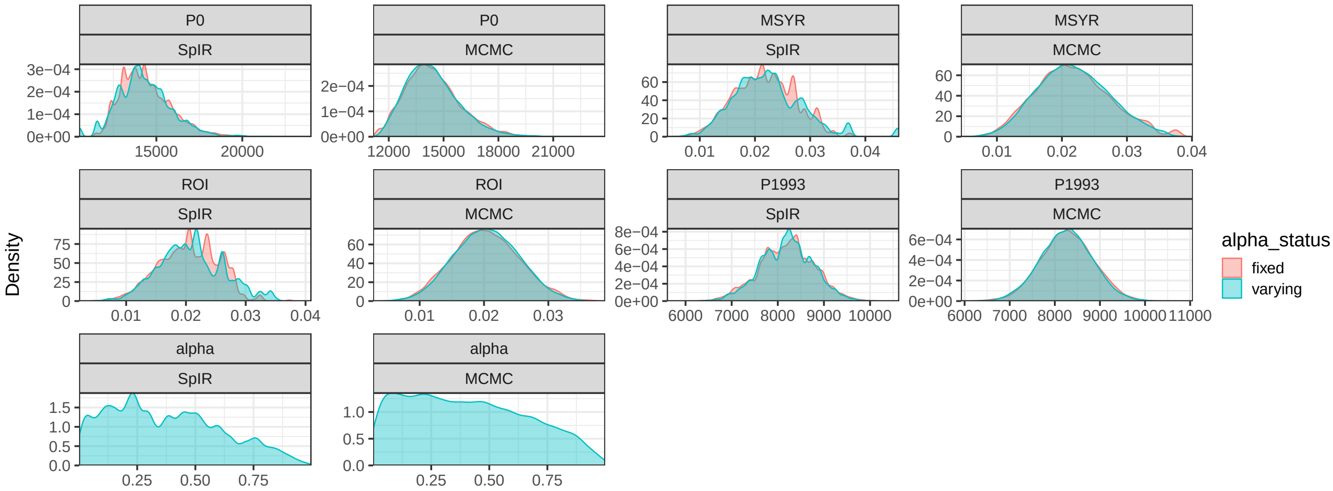


Figure 4: **Marginal posterior distributions for various quantities of interest in the bowhead population model**. We show the posterior distributions obtained by using sampling importance-resampling (SpIR) and Markov chain Monte Carlo (HMC-MCMC), for fixed $\alpha = 1/2$ and placing a prior π_A on α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

4.4.2 Influenza in a boarding school

Now, we will consider 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

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

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

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

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 $\sigma_\beta^2 = \sigma_\gamma^2 = 1$, which induces a log-normal distribution (q_1^*) on R_0 with parameters $\mu_1 = \mu_\beta - \mu_\gamma$ and $\sigma_1^2 = \sigma_\beta^2 + \sigma_\gamma^2$. 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.3), while the log-normal priors are flexible and useful in practice. We approximate the posterior in (37) using HMC as described above.

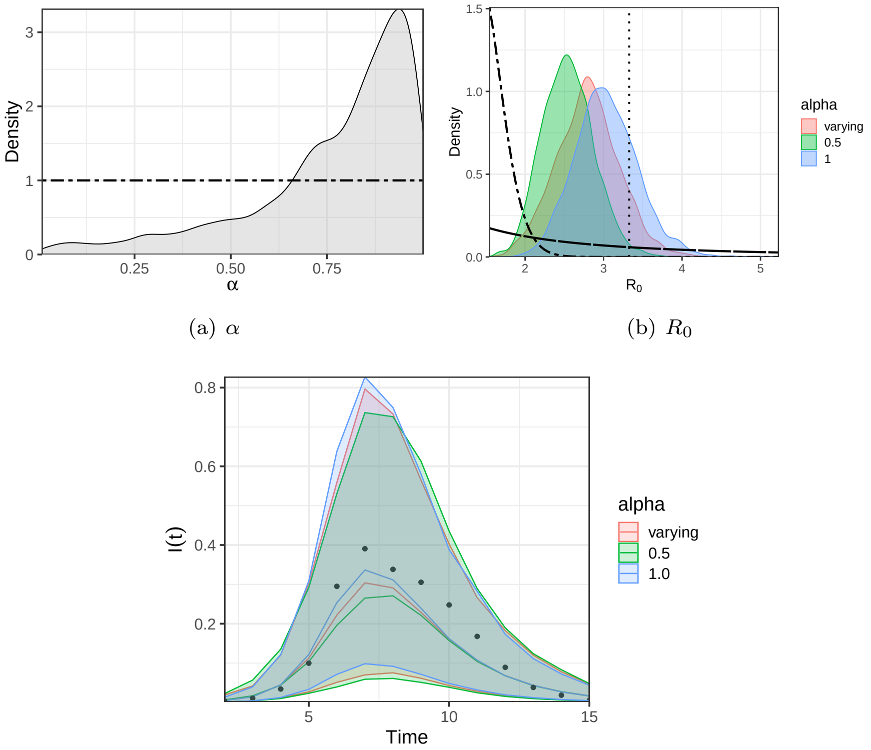
In Figure 5a we show the posterior distribution of the pooling weight α, which favours high values with a mean and 95% credibility interval of 0.77 (0.21–0.99). The posterior distribution for R_0 obtained by letting α vary and also the resulting distributions of fixing $\alpha = 1/2$ or $\alpha = 1$ 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 and Raftery (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 acknowledgement 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

First, we note that the properties in Section 2 hold in more generality and the main choice of this paper, namely the choice to represent opinions as probability distributions, is not inconsequential. As pointed out by French (1985), one needs not assume that all experts express their opinion on the same scale. This means one needs not assume that the expert opinions are encoded as probability distributions, let alone distributions in the same family. Since our main goal is to motivate the use of logarithmic pooling in a statistical context, we argue this provides justification for the use of probability distributions.

With regard to the construction of the pooling weights ($\boldsymbol{\alpha}$) based on optimality criteria of the pooled distribution, we show that these usually lead to degenerate solutions that exclude one or more experts/information sources. Considering the maximum entropy solutions in sections 4.1 and 4.3 one might wonder whether maximising entropy can lead non-degenerate solutions. A detailed characterisation of the conditions for the existence non-degenerate solutions is beyond the scope of the present paper, but we note that any configuration of the densities $\boldsymbol{F}_X$ can also lead to equally good (under the loss function) degenerate solutions, which is unavoidable due to the lack of a unique solution. The minimum KL



(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 longdash 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.

solutions (including the approach of (Rufo et al., 2012b)), which are unique, also lead to several experts being excluded.

Of course, 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. The weights are not probabilities, but 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 events that are logically impossible. So far we have discussed approaches that consider only the information contained in the set of probability distributions $\boldsymbol{F}_X$. But a hierarchical prior approach, which assigns a probability distribution to the weights, addresses both the issue of Cromwell's rule and the case where one has data x from which the relative merits of the experts can be discerned by analysing the posterior distribution of $\boldsymbol{\alpha}$. The results in Sections 4.1 and 4.2 show that it is possible to learn the weights from data, while achieving efficient inference about the quantity of interest (but see below).

The flexibility provided by the hierarchical prior deserves a bit more consideration. For the Bayesian melding method, it is usual to set the pooling weight at $\alpha = 1/2$. This choice leads to the so-called geometric pooling, since it is equivalent to taking the geometric mean of the densities q_1^* and q_2 and is the only choice of α which is invariant to the re-labelling of inputs and outputs (Poole and Raftery, 2000). Since there is usually clear scientific separation between inputs and outputs, we do not feel this is adequate justification for fixing the value of α. Indeed, Poole and Raftery (2000) (Section 5.2) do argue that estimating α would be a fruitful path to explore and our results corroborate that view. For the bowhead example (Section 4.4.1), while the large posterior uncertainty about α precludes definitive statements, support for $\alpha < 1/2$ makes sense because q_1^* is very diffuse. By placing a lower weight on q_1^*, we favour q_2 as a prior for P_{1993} and hence achieve a reduction in posterior uncertainty, albeit a small one. We obtained $\Pr(\alpha < 1/2) = 0.69$ and 0.64 for SpIR and MCMC, respectively. In contrast, the SIR example (Section 4.4.2) shows the situation in reverse, where the data are incompatible with the more informative prior. The posterior distribution for α in this example accommodates this incompatibility, favouring the induced prior which results from non-informative priors on the inputs α and β. For this example, $\Pr(\alpha < 1/2) = 0.11$.

The use of logarithmic pooling is not without its pitfalls, however. The meta-analysis in Section 4.3 indicates that it can be very sensitive to the choice of distribution to represent expert opinion. This is particularly true for the optimality-based methods, specially minimising KL distance. Representing the distribution of HIV prevalence by a Beta or a Gaussian distribution lead to different studies being favoured by the minimum KL solution. The (degenerate) maximum entropy solution was the same for both representations, but it is difficult to tell how dependent this is on the differences or lack thereof between Beta and Gaussian distributions due to the lack of uniqueness and propensity to achieve degenerate solutions. The results in Section 4.2 show that posterior inference for the weights only allows identifying the distribution most compatible with the data for some configurations of the set of opinions $\boldsymbol{F}_X$ (Figure 2b). More specifically, the ability to discern the "best" expert is linked to the separation between the expert opinions; if the "correct" expert provides a distribution that is too diffuse, we loose the ability to correctly identify it. Importantly, of the four priors we considered here, only the logistic-normal prior with the first two moments matched to those of a Dirichlet(1/10, ..., 1/10) was flexible enough to allow correct identification over a large range of diffuseness. Further research is needed in order to evaluate the possibility of more flexible priors.

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 multiple integral, which is unfeasible to do via quadrature and would likely necessitate specially-designed MCMC methods.

6 Final remarks

In this paper, our goal was to study how to one can assign the weights in logarithmic pooling such that it can be employed in statistical analysis. We have shown how it can be a useful tool in constructing prior distributions, combining posterior distributions in meta-analysis and learning about the parameters of deterministic models. We discussed how to approach the assigning of the weights based on optimality criteria such as maximum entropy and proposed a novel hierarchical prior approach that accommodates uncertainty about the weights and allows learning about them from data whilst providing more flexible prior modelling. Future research will explore further applications of logarithmic pooling in statistical learning like for instance combining several posterior predictive distributions obtained from cross-validation (Yao et al., 2018). 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/information sources be judged by how reasonable their distributions look in transformed space?

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

Acknowledgements

The authors would like to thank Professor Adrian Raftery (University of Washington) and David Poole 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. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001. LMC is supported by a postodoctoral fellowship from CAPES.

References

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

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

Alkema, L., Raftery, A. E., and Brown, T. (2008). Bayesian melding for estimating uncertainty in national hiv prevalence estimates. *Sexually transmitted infections*, 84(Suppl 1):i11–i16.

Anon. (1978). Influenza in a boarding school. *The British Medical Journal*, 1:587.

Barcellos, N. T., Fuchs, S. C., and Fuchs, F. D. (2003). Prevalence of and risk factors for HIV infection in individuals testing for HIV at counseling centers in Brazil. *Sexually transmitted diseases*, 30(2):166–173.

Betancourt, M. (2012). Cruising the simplex: Hamiltonian monte carlo and the dirichlet distribution. In *AIP Conference Proceedings 31st*, volume 1443, pages 157–164. AIP.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., and Finelli, L. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC infectious diseases*, 14(1):480.

Brazilian Ministry of Health (2000). Bela vista & horizon: Behavioral and serological studies among men who have sex with men. Technical report, Brasília: Brazilian Ministry of Health. In Portuguese.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Carneiro, M., Cardoso, F. A., Greco, M., Oliveira, E., Andrade, J., Greco, D. B., and Antunes, C. M. d. F. (2003). Determinants of human immunodeficiency virus (HIV) prevalence in homosexual and bisexual men screened for admission to a cohort study of HIV negatives in Belo Horizonte, Brazil: Project Horizonte. *Memórias do Instituto Oswaldo Cruz*, 98(3):325–329.

Coelho, F. C. and Carvalho, L. M. (2015). Estimating the attack ratio of dengue epidemics under time-varying force of infection using aggregated notification data. *Scientific reports*, 5:18455.

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

Evans, R. J. and Boersma, J. (1988). The entropy of a poisson distribution (c. robert appledorn). *SIAM Review*, 30(2):314–317.

French, S. (1985). Group consensus probability distributions: A critical survey in bayesian statistics. *Bayesian statistics*, 2.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

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

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

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

Guardoni, G. L. (2002). On irrelevance of alternatives and opinion pooling. *Brazilian Journal of Probability and Statistics*, pages 87–98.

Harrison, L. H., do Lago, R. F., Friedman, R. K., Rodrigues, J., Santos, E. M., de Melo, M. F., Moulton, L. H., Schechter, M., Group, P. P. O. S., et al. (1999). Incident hiv infection in a high-risk, homosexual, male cohort in rio de janeiro, brazil. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 21(5):408–412.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

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

Jombart, T., Frost, S., Nouvellet, P., Campbell, F., and Sudre, B. (2019). *outbreaks: A Collection of Disease Outbreak Data*. R package version 1.6.0.

Li, Z. S., Guo, J., Xiao, N.-C., and Huang, W. (2017). Multiple priors integration for reliability estimation using the bayesian melding method. In *Reliability and Maintainability Symposium (RAMS), 2017 Annual*, pages 1–6. IEEE.

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

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Malta, M., Magnanini, M. M., Mello, M. B., Pascom, A. R. P., Linhares, Y., and Bastos, F. I. (2010). HIV prevalence among female sex workers, drug users and men who have sex with men in Brazil: a systematic review and meta-analysis. *BMC Public Health*, 10(1):1.

Murray, J. D. (2002). *Mathematical Biology I. An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, 3 edition.

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

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566.

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, 3rd. edition.

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

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

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

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

Sutmoller, F., Penna, T. L., de Souza, C. T. V., Lambert, J., Group, O. C. F. S. P., et al. (2002). Human immunodeficiency virus incidence and risk behavior in the 'Projeto Rio': Results of the first 5 years of the Rio de Janeiro open cohort of homosexual and bisexual men, 1994–1998. *International journal of infectious diseases*, 6(4):259–265.

Tun, W., de Mello, M., Pinho, A., Chinaglia, M., and Diaz, J. (2008). Sexual risk behaviours and HIV seroprevalence among male sex workers who have sex with men and non-sex workers in Campinas, Brazil. *Sexually Transmitted Infections*, 84(6):455–457.

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

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003.

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

Zhong, M., Goddard, N., and Sutton, C. (2015). Latent bayesian melding for integrating individual and population models. In *Advances in Neural Information Processing Systems*, pages 3618–3626.

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 $\mathbb{E}_0[X_j]^{\alpha_j} = \left(\int_{\boldsymbol{\Theta}} f_0(\theta) \frac{f_j(\theta)}{f_0(\theta)} \right)^{\alpha_j} d\theta = 1^{\alpha_j} = 1$, Theorem 1 is proven. □

To establish Remark 3, we will need the following result from Genest et al. (1984).

Lemma 1. ***Representation of a pooling operator with RPC*** *(Genest et al., 1984, eq. 3.1). The only relative propensity consistent operator can always be represented by*

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

with $\boldsymbol{B}\left(\boldsymbol{F}_\theta\right) > 0$, $c(\theta) > 0$ and $w_0, w_1, \ldots, w_K \geq 0$ arbitrary.

Now we can state the proof of Remark 3.

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). Then, with the goal of obtaining a contradiction, suppose that there exists a pooling operator $\mathcal{T}$ that is log-concave but does not enjoy RPC. From Lemma 1, we know that $\mathcal{T}$ cannot be represented as $\boldsymbol{B}(\boldsymbol{F}_\theta)c(\theta)\prod_{i=0}^{K} f_i(\theta)^{w_i}$. But every non-negative log-concave function $g(x)$ can be represented as $g(x) = a \cdot c(x) \cdot h(x)$, with $a \geq 0$ and $c(x)$ and $h(x)$ non-negative and log-concave. Since under the assumptions on $\boldsymbol{F}_\theta$ we have that $h(\theta) := \prod_{i=0}^{K} f_i(\theta)^{w_i}$ is non-negative and log-concave, a contradiction is reached and the result follows. □

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 Bowhead population growth model: details

For convenience, here we will describe the priors and likelihoods used by Poole and Raftery (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 and Raftery (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 and Raftery (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 3.3.3 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.3 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 4.3) distributions as the expert opinions. In this section we give further results for the pooling of commonly used distributions.

A.3.1 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

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

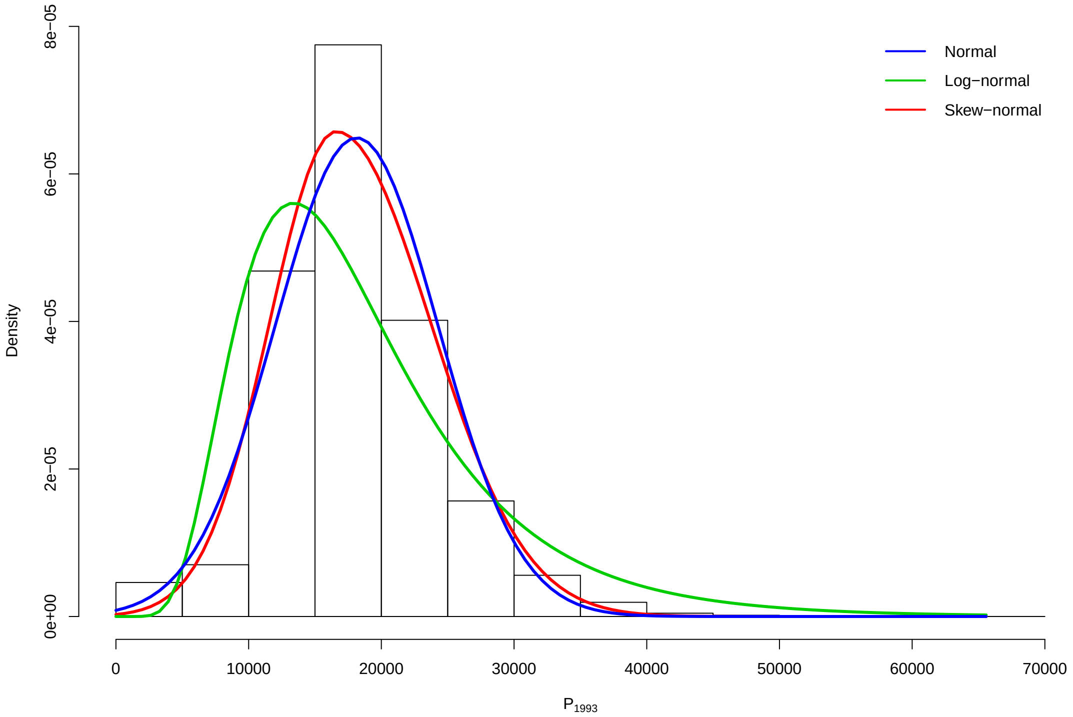


Figure S1: **Induced distribution on P_1993 (q_1^*) and approximating distributions**. We present the histogram of $100,000$ simulations from the prior. Lines show the densities of the three distributions considered.

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

$$\begin{aligned}\pi(\lambda) &= t(\boldsymbol{\alpha})\prod_{i=0}^{K} f_i(\lambda; a_i, b_i)^{\alpha_i},\\ &\propto \prod_{i=0}^{K}\left(\lambda^{a_i-1}e^{-b_i\lambda}\right)^{\alpha_i},\\ &\propto \lambda^{a^*-1}e^{-b^*\lambda}, \end{aligned} \tag{40}$$

where $a^* = \sum_{i=0}^{K}\alpha_i a_i$ and $b^* = \sum_{i=0}^{K}\alpha_i b_i$. Noticing (40) is the kernel of a gamma distribution with parameters a^* and b^*, $H_\pi(\lambda)$ becomes

$$H_\pi(\lambda; \boldsymbol{\alpha}) = a^* - \log b^* + \log\Gamma(a^*) + (1-a^*)\psi(a^*), \tag{41}$$

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

$$\text{KL}(\pi||f_i) = (a_i - a^*)\psi(a_i) - \log\Gamma(a_i) + \log\Gamma(a^*) + a^*\left(\log\frac{b_i}{b^*}\right) + \frac{a_i}{b_i}(b^* - b_i). \tag{42}$$

A.3.2 Log-normal

Another popular choice for modelling a quantity $\eta \in \mathbb{R}^+$ is the log-normal family. Following the results given in Section 4.3, 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 $w_i = \alpha_i/\sigma_i^2$.

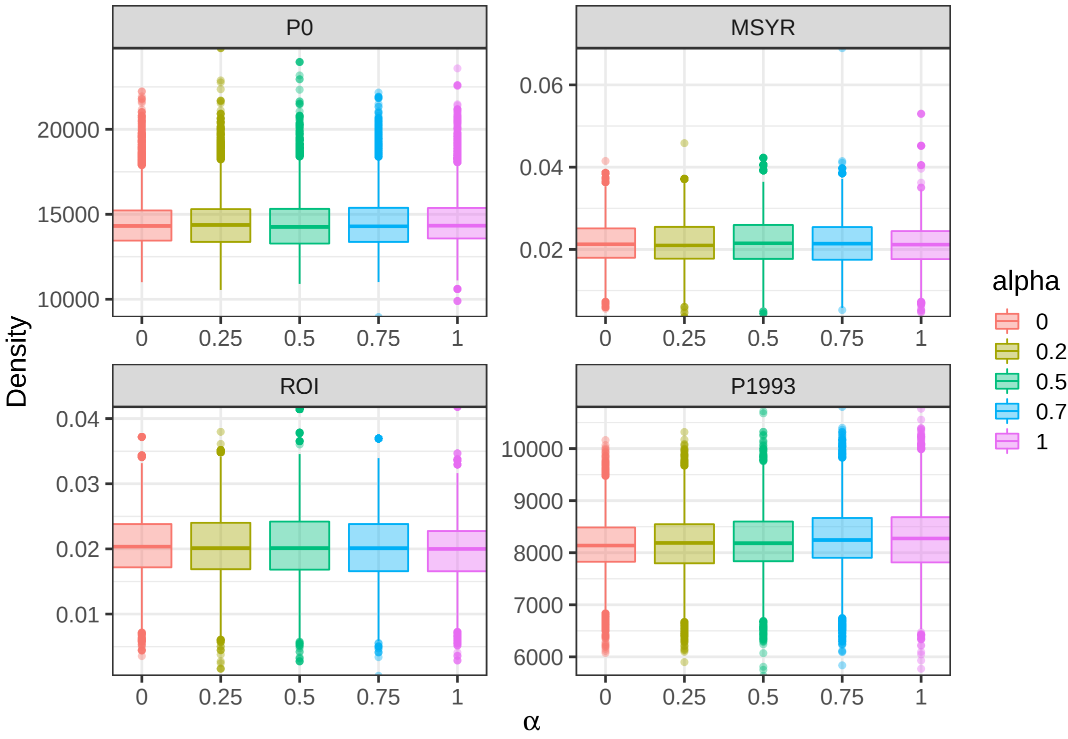


Figure S2: **Sensitivity of posterior inferences to varying the value of α, bowhead population model**.
We show the posterior distributions obtained by SpIR for P_0, MSYR, ROI and P_1993 as we fix α to different values.

The entropy function is then:

$$H_\pi(\eta; \boldsymbol{\alpha}) = \log_2(e) \log\left(\sigma^\star \exp\left(\mu^\star + \frac{1}{2}\right)\sqrt{2\pi}\right),$$
$$= \log_2(e)\left[\log\left(\sigma^\star\right) + \mu^\star + \frac{1}{2} + \log(\sqrt{2\pi})\right]. \tag{43}$$

The KL divergence evaluates to

$$\mathrm{KL}(\pi||f_i) = \frac{1}{2\sigma_i^2}\left[(\mu^\star - \mu_i)^2 + \sigma^{\star 2} - \sigma_i^2\right] + \log\left(\frac{\sigma_i^2}{\sigma^{\star 2}}\right). \tag{44}$$

A.3.3 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

$$\pi(y) \propto \prod_{i=0}^{K}\left(\frac{\lambda_i^y}{y!}\right)^{\alpha_i},$$
$$\pi(y) = \frac{\exp(-\lambda^\star)\lambda^{\star y}}{y!}, \text{ with } \lambda^\star = \prod_{i=0}^{K}\lambda_i^{\alpha_i}. \tag{45}$$

The entropy of the pooled distribution is

$$H_\pi(y; \boldsymbol{\alpha}) = -\lambda^\star \log\left(\frac{\lambda^\star}{e}\right) + E_\pi\left[\log(k!)\right], \tag{46}$$

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

The KL divergence is

$$\begin{aligned}\mathrm{KL}(\pi||f_i) &= \lambda^\star \log\left(\frac{\lambda^\star}{\lambda_i}\right) + \lambda_i - \lambda^\star, \\ &= \lambda^\star \left[\sum_{k=0}^{K} \alpha_k \log(\lambda_k) - \log(\lambda_i)\right] + \lambda_i - \lambda^\star. \qquad (47)\end{aligned}$$