Supplement: Meta-analysis of Bayesian analyses

Paul Blomstedt, Diego Mesquita and Samuel Kaski
Helsinki Institute for Information Technology HIIT,
Department of Computer Science, Aalto University
{paul.blomstedt, diego.mesquita, samuel.kaski}@aalto.fi

1 Meta-analysis as Bayesian inference with observed beliefs

In this section, we first develop a posterior update rule given observed beliefs, which is motivated by the problem of conducting meta-analysis for a set of related posterior distributions. The notion of relatedness is here characterized as the exchangeability of the quantities targeted by the posteriors. The proposed update rule is given in Equation (3). We then show in Section 2 that this rule retains some basic theoretical properties of standard Bayesian inference.

Let us first assume that $\theta_1, \dots, \theta_J$ is a collection of observable and exchangeable random quantities. Following standard theory, de Finetti's representation theorem [e.g. 12] states that if $\theta_1, \theta_2, \dots$ is an infinitely exchangeable sequence of random quantities taking values in a Borel space $(\Theta, \mathcal{A})$, then there exists a probability measure Q such that the joint distribution $\mathbb{P}$ of the subsequence $\theta_1, \dots, \theta_J$, i.e. the *predictive distribution*, has the form

$$\mathbb{P}(\theta_1 \in A_1, \dots, \theta_J \in A_J) = \int_\Phi \prod_{j=1}^{J} \left[\int_{A_j} p(\theta_j|\varphi)\lambda(d\theta_j) \right] Q(d\varphi), \tag{1}$$

where $A_1, \dots, A_J \in \mathcal{A}$. Here, the set of probability measures on Θ is taken to be a family $\{P_\varphi | \varphi \in \Phi\}$, indexed by a parameter φ, such that Q is a probability measure on Φ. Furthermore, we define the density function $p(\cdot|\varphi) := dP_\varphi/d\lambda$ with respect to a reference measure λ (Lebesgue or counting measure).

In the above standard setting, the Bayesian learning process works through updating Q conditional on observed data. Following Equation (1), the posterior distribution of φ, given observed values $\theta_1 = t_1, \dots, \theta_J = t_J$, has the form

$$Q(B|t_1, \dots, t_J) = \frac{\int_B \prod_{j=1}^J p(t_j|\varphi)Q(d\varphi)}{\int_\Phi \prod_{j=1}^J p(t_j|\varphi)Q(d\varphi)}, \tag{2}$$

with $B \in \mathcal{B}$, the Borel σ-algebra on Φ. We will now build further on this setting, assuming that instead of directly observing the value of each θ_j, we have a set of distributions $\Pi_1, \dots, \Pi_J$, expressing our currently available beliefs about the values of θ_j. Note that, while in our current context, we assume that each of the beliefs is obtained as the posterior distribution from a previously conducted analysis, this assumption is not essential to our developments. Importantly, the observed distributions are assumed independent of the distribution we seek to update. In the absence of fixed likelihood contributions $p(t_j|\varphi)$ for each observation, we propose to compute the *expected likelihood contributions* $\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)$ with respect to the available beliefs.

The proposed modification of the likelihood now leads to an update of the form

$$Q^*(B|\Pi_1, \dots, \Pi_J) = \frac{\int_B \prod_{j=1}^J \left[\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}{\int_\Phi \prod_{j=1}^J \left[\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}, \tag{3}$$

where, with slight abuse of notation, we write $Q^*(\cdot|\Pi_1, \dots, \Pi_J)$ to denote conditioning on beliefs in analogy with conditioning on fully observed values; we give more context for this choice of notation

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

below in Section 2.2. Equation (3) further induces a joint distribution on $\Phi \times \Theta^J$, which can be marginalized with respect to Q, resulting in a predictive distribution as follows:

$$\mathbb{P}^*(\theta_1 \in A_1, \ldots, \theta_J \in A_J) = \frac{\int_\Phi \prod_{j=1}^J \left[\int_{A_j} p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}{\int_\Phi \prod_{j=1}^J \left[\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}. \tag{4}$$

It easy to see that standard Bayesian inference, Equation (2), emerges as a special case of Equation (3) by setting Π_j to be δ_{t_j}, the Dirac measure centered at t_j. This yields

$$\int_\Theta p(\theta_j|\varphi)\delta_{t_j}(d\theta_j) = p(t_j|\varphi),$$

such that $Q^*(\cdot|\delta_{t_1}, \ldots, \delta_{t_J}) = Q(\cdot|t_1, \ldots, t_J)$. Throughout this work, we assume that Π_j is a probability measure. However, it is interesting to note that if we make an exception and allow Π_j to be the Lebesgue (or counting) measure λ for all j, which corresponds to having a uniformly distributed—possibly improper—belief about the value of θ_j, then the updated measure Q^* in Equation (3) equals the prior probability measure Q. Moreover, with this choice of Π_j, Equation (4) reduces to the standard predictive distribution in Equation (1).

2 Theoretical properties

Two well-known properties of standard Bayesian inference, which are of practical relevance in our meta-analysis setting, are (i) order-invariance in successive posterior updates of exchangeable models and (ii) posterior concentration. The former ensures that inferences conditional on exchangeable data are coherent. The latter tells us that the posterior distribution becomes increasingly informative about the quantity of interest, as we accumulate more data. We will now briefly discuss these properties in the context of the framework introduced above.

2.1 Order-invariance in successive posterior updates

Under the assumption of exchangeability, standard Bayesian inference can be constructed as a sequence of successive updates, invariant to the order in which the data are processed. The following proposition establishes that the update rule defined in Equation (3) retains the same property.

Proposition 1. *The update rule defined in Equation (3) is invariant to permutations of the indices* $1, \ldots, J$.

Proof. It suffices for us to verify the claim for $J = 2$. Beginning with $J = 1$, we update the probability $Q(B)$ into $Q^*(B|\Pi_1)$ using Equation (3):

$$Q^*(B|\Pi_1) = \frac{\int_B \int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}{\int_\Phi \int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}.$$

Then, we reapply Equation (3) to update $Q^*(B|\Pi_1)$ into $Q^*(B|\Pi_1, \Pi_2)$:

$$\begin{aligned} Q^*(B|\Pi_1, \Pi_2) &= \frac{\int_B \int_\Theta p(\theta_2|\varphi)\Pi_2(d\theta_2)Q^*(d\varphi|\Pi_1)}{\int_\Phi \int_\Theta p(\theta_2|\varphi)\Pi_2(d\theta_2)Q^*(d\varphi|\Pi_1)} \\ &= \frac{\int_B \int_\Theta p(\theta_2|\varphi)\Pi_2(d\theta_2)}{\int_\Phi \int_\Theta p(\theta_2|\varphi)\Pi_2(d\theta_2)} \frac{\frac{\int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}{\cancel{\int_\Phi \int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}}}{\frac{\int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}{\cancel{\int_\Phi \int_\Theta p(\theta_1|\varphi)\Pi_1(d\theta_1)Q(d\varphi)}}} \\ &= \frac{\int_B \prod_{j=1}^2 \left[\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}{\int_\Phi \prod_{j=1}^2 \left[\int_\Theta p(\theta_j|\varphi)\Pi_j(d\theta_j)\right] Q(d\varphi)}, \end{aligned}$$

which is equivalent to a direct application of Equation (3) for $J = 2$, and independent of the order in which Π_1 and Π_2 are processed. □

As an alternative strategy to Equation (3), we could first attempt to formulate a posterior distribution $Q(\cdot|\theta_1, \ldots, \theta_J)$ according to Equation (2) and then, as a final step, integrate out the uncertainty in the conditioning set with respect to the observed beliefs. This is in essence the strategy of Jeffrey's rule of conditioning. It is, however, well known that Jeffrey's rule is in general not order-invariant [3].

2.2 Posterior concentration

Asymptotic theory states that if a consistent estimator of the true value (or an optimal one in terms KL-divergence) of the parameter φ exists, then the posterior distribution (2) concentrates in a neighborhood of this value, as $J \to \infty$ [e.g. 12]. Here we discuss conditions under which the same property holds for the measure $Q^*(\cdot|\Pi_1, \ldots, \Pi_J)$, defined in Equation (3). Considerations of asymptotic normality will not be discussed here.

Our strategy is to first formulate a generative hierarchical model for the observed distributions $\Pi_1, \ldots, \Pi_J$. Then, we show that $Q^*(\cdot|\Pi_1, \ldots, \Pi_J)$ can be expressed as the marginal posterior distribution of φ in this model. Finally, we show that under some further technical conditions, standard asymptotic theory can be applied to this distribution. To this end, consider the following hierarchical model:

$$\varphi \sim Q \tag{5a}$$

$$\theta_j \sim P_\varphi \tag{5b}$$

$$\Pi_j \sim G^{(j)}_{\theta_j}, \tag{5c}$$

where Π_j is treated as a soft observation of the unobserved value of θ_j, and $G^{(j)}_{\theta_j}$ is the inference mechanism that produces Π_j. Note that in the particular case of Π_j being the Dirac measure, $G^{(j)}_{\theta_j}$ simply generates a point mass at the true value of θ_j, such that the hierarchical model reduces to an ordinary, non-hierarchical Bayesian model. Since Π_j is an inference over the values of θ_j, produced by $G^{(j)}_{\theta_j}$, it is also a direct representation of the likelihood of θ_j under the model $G^{(j)}_{\theta_j}$, given the observation Π_j itself. We finally note, that the generating mechanism $G^{(j)}_{\theta_j}$ may in general be different for each j, which is highlighted in the notation by the superscript.

Assume now that the Radon-Nikodym derivative $g_j(\cdot|\theta_j) := dG^{(j)}_{\theta_j}/d\kappa$ with respect to a dominating measure $\kappa \gg G^{(j)}_{\theta_j}$ can be defined for all j. The marginal posterior distribution of φ is then

$$Q'(B|\Pi_1, \ldots, \Pi_J) = \frac{\int_B \prod_{j=1}^J \left[\int_\Theta g_j(\Pi_j|\theta_j) P_\varphi(d\theta_j)\right] Q(d\varphi)}{\int_\Phi \prod_{j=1}^J \left[\int_\Theta g_j(\Pi_j|\theta_j) P_\varphi(d\theta_j)\right] Q(d\varphi)},$$

where $g_j(\Pi_j|\theta_j)$, taken as a function of θ_j, is the likelihood of θ_j given Π_j. On the other hand, according to our previous assumption, the likelihood is directly encapsulated in Π_j itself. We will therefore assume, by construction, that $g_j(\Pi_j|\theta_j) = \pi_j(\theta_j)$, where $\pi_j := d\Pi_j/d\lambda$ is the density function corresponding to Π_j. Using this equivalence, we state the following lemma:

Lemma 2. *Let $g_j(\Pi_j|\theta_j) = \pi_j(\theta_j)$. Then the measures $Q'(\cdot|\Pi_1, \ldots, \Pi_J)$ and $Q^*(\cdot|\Pi_1, \ldots, \Pi_J)$ are equivalent.*

Proof. To prove the claim, we only need to verify that the integrals $\int_\Theta g_j(\Pi_j|\theta_j) P_\varphi(d\theta_j)$ and $\int_\Theta p_j(\theta_j|\varphi)\Pi_j(d\theta_j)$ are equivalent. Since Π_j and P_φ are both probability distributions on $(\Theta, \mathcal{A})$, and their densities are defined with respect to the same reference measure λ, we may swap the roles of the integrand and the measure that we integrate against:

$$\begin{aligned}\int_\Theta g_j(\Pi_j|\theta_j) P_\varphi(d\theta_j) &= \int_\Theta \pi_j(\theta_j) P_\varphi(d\theta_j) = \int_\Theta \frac{d}{d\lambda}\Pi_j(\theta_j) P_\varphi(d\theta_j) \\ &= \int_\Theta \frac{d}{d\lambda} P_\varphi(\theta_j)\Pi_j(d\theta_j) = \int_\Theta p_j(\theta_j|\varphi)\Pi_j(d\theta_j).\end{aligned}$$

□

Corollary 2.1. *In the hierarchical model (5a)–(5c), let $H^{(j)}_\varphi$ be the marginal conditional distribution of Π_j, given φ, and let $h_j(\Pi_j|\varphi) := \int_\Theta g_j(\Pi_j|\theta_j) P_\varphi(d\theta_j)$ be the corresponding marginal likelihood. Following Lemma 2, the probability $Q^*(B|\Pi_1, \ldots, \Pi_J)$ can be written as*

$$Q^*(B|\Pi_1, \ldots, \Pi_J) = \frac{\int_B \prod_{j=1}^J h_j(\Pi_j|\varphi) Q(d\varphi)}{\int_\Phi \prod_{j=1}^J h_j(\Pi_j|\varphi) Q(d\varphi)}.$$

The essential difference between the standard posterior distribution defined in Equation (2) and that of Corollary 2.1 above, is that in the former, the observations are conditionally i.i.d., while in the latter they are conditionally independent but *non-identically* distributed.

Assuming that there exists, for all j, a unique minimizer φ_0 of the KL-divergence from the true distribution of Π_j to the parametrized representation $H_\varphi^{(j)}$, a key step in proving posterior concentration is to establish a limit for the log-likelihood ratio $\sum_{j=1}^J \log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)}$. Since the summands are now non-identically distributed, standard forms of the strong law of large numbers cannot be applied to obtain this limit. However, with further conditions imposed on the second moment of each term in the sum, an alternative form can be used, which relaxes the requirement of the terms being identically distributed [13, Theorem 2.3.10]. The conditions are stated in the following theorem.

Theorem 3. *Assume that the log-likelihood ratio terms $\xi_j := \log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)}$ are independent, and that $\mathbb{E}(\xi_j) = \mu_j$ and $\mathrm{Var}(\xi_j) = \sigma_j^2$ exist for all $j \geq 1$. Let $\overline{\mu}_J = J^{-1} \sum_{j=1}^J \mu_j$, for $J \geq 1$. Then*

$$\sum_{j \geq 1} j^{-2} \sigma_j^2 < \infty \Rightarrow J^{-1} \sum_{j=1}^J \xi_j - \overline{\mu}_J \xrightarrow{a.s.} 0.$$

In conclusion, if the measure $Q^*(\cdot|\Pi_1, \ldots, \Pi_J)$ can be written in the form of Equation (2.1) and the conditions of Theorem 3 hold, then posterior concentration falls back to the standard case. A rigorous treatment is given in Schervish [12]. As an example, we provide a proof of posterior concentration of $Q^*(\cdot|\Pi_1, \ldots, \Pi_J)$ for discrete parameter spaces. The proof follows the basic structure found in many sources [e.g. 1, 4], with the essential difference that the observations are independent but non-identically distributed.

Theorem 4. *Let $\{\Pi_1, \ldots, \Pi_J\}$ be a set of observations from a corresponding set of distributions $\{R_1, \ldots, R_J\}$. Furthermore, let $\left\{\{H_\varphi^{(j)} | \varphi \in \Phi\}\right\}_{j=1}^J$ be a set of families, such that*

(i) Φ consists of (at most) a countable set of values,

(ii) $\varphi_0 = \arg\min_{\varphi \in \Phi} \mathrm{KL}\left(R_j || H_\varphi^{(j)}\right)$, for all j.

If the conditions of Theorem 3 hold, and furthermore if $\sum_{\varphi \in \Phi} Q(\varphi) = 1$ and $Q(\varphi_0) > 0$, then $Q^*(\varphi_0|\Pi_1 \ldots, \Pi_J) \to 1$, as $J \to \infty$.

Proof. For any $\varphi \neq \varphi_0$, the log posterior odds can written as

$$\log \frac{Q(\varphi|\Pi_1, \ldots, \Pi_J)}{Q(\varphi_0|\Pi_1, \ldots, \Pi_J)} = \log \frac{Q(\varphi)}{Q(\varphi_0)} + \sum_{j=1}^J \log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)}, \tag{6}$$

where the second term is a sum of J independent but non-identically distributed random variables. By Theorem 3, we have that

$$\frac{1}{J} \sum_{j=1}^J \log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)} \to \mathbb{E}\left(\log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)}\right),$$

with probability 1, as $J \to \infty$. Since φ_0 is the unique minimizer of the KL-divergence $\mathrm{KL}\left(R_j || H_\varphi^{(j)}\right)$, by definition

$$\mathbb{E}\left(\log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)}\right) = \mathrm{KL}\left(R^j || H_{\varphi_0}^j\right) - \mathrm{KL}\left(R^j || H_\varphi^j\right) < 0,$$

and consequently,

$$\sum_{j=1}^J \log \frac{h_j(\Pi_j|\varphi)}{h_j(\Pi_j|\varphi_0)} \to -\infty.$$

Since $Q(\varphi_0) > 0$, the entire expression (6) approaches $-\infty$ as $J \to \infty$, which implies that $Q(\varphi|\Pi_1, \ldots, \Pi_J) \to 0$ and $Q(\varphi_0|\Pi_1, \ldots, \Pi_J) \to 1$. □

3 Meta-analysis of Bayesian analyses in practice

We now turn to a practical view of the framework developed in the previous section. To this end, it is convenient to work with densities instead of measures. We are motivated by the problem of conducting meta-analysis for Bayesian analyses summarized as posterior distributions, and refer to our framework as *meta-analysis of Bayesian analyses* (MBA). The central belief updates of the framework are given in Equations (8) and (9), which update beliefs regarding global and local effects, respectively. Figures 1a and 1b visualize the updates by interpreting them as message passing in probabilistic graphical models.

Assume that a set of posterior density functions $\{\pi_1, \ldots, \pi_J\}$ is available, each expressing a belief about the value of a corresponding quantity of interest in a set $\{\theta_1, \ldots, \theta_J\}$. While the density functions can be thought of as resulting from previously conducted Bayesian analyses, it is worth pointing out that from a methodological point of view, we are agnostic to *how* they have been formed; instead of posteriors, some (or all) of the π_j's could be purely subjective prior beliefs, or as previously discussed, even directly observed values.

Judging the quantities θ_j to be exchangeable, the *meta-analyst* now formulates a model

$$\prod_{j=1}^{J} p(\theta_j|\varphi)q(\varphi), \tag{7}$$

with an appropriate prior q placed on the parameter φ. Note that this model initially makes no reference to the π_j's, and it is formulated *as if* the θ_j's were fully observable quantities. Then, to update q based on the observed density functions, we apply Equation (3) in density form to have

$$q^*(\varphi) \propto \prod_{j=1}^{J} \left[\int_{\Theta} p(\theta_j|\varphi)\pi_j(\theta_j)d\theta_j \right] q(\varphi), \tag{8}$$

where for brevity, we denote $q^* := q^*(\cdot|\Pi_1, \ldots, \Pi_J)$.

In a meta-analysis context, the parameter φ often has an interpretation as the central tendency of some shared property of $\theta_1, \ldots, \theta_J$, such as the mean or the covariance (or both jointly). As such, inference on φ is often of primary interest in providing a 'consensus' over a number of studies. As a secondary goal, we may also be interested in updating a (possibly weakly informative) belief about any individual quantity θ_j, subject to the information provided by observations on the remaining quantities. To do so, we first write Equation (4) in density form:

$$p^*(\theta_1, \ldots, \theta_J) \propto \int_{\Phi} \prod_{j=1}^{J} \left[p(\theta_j|\varphi)\pi_j(\theta_j)d\theta_j \right] q(\varphi)d\varphi,$$

and then marginalize over all quantities but the one to be updated. Let $\mathcal{J} := \{1, \ldots, J\}$ be a set of indices and let $j' \in \mathcal{J}$ be an arbitrary index in this set. The density function $\pi_{j'}$ is then updated as follows:

$$\pi^*_{j'}(\theta_{j'}) \propto \int_{\Phi} p(\theta_{j'}|\varphi)\pi_{j'}(\theta_{j'}) \prod_{j \in \mathcal{J} \setminus j'}^{J} \left[\int_{\Theta} p(\theta_j|\varphi)\pi_j(\theta_j)d\theta_j \right] q(\varphi)d\varphi. \tag{9}$$

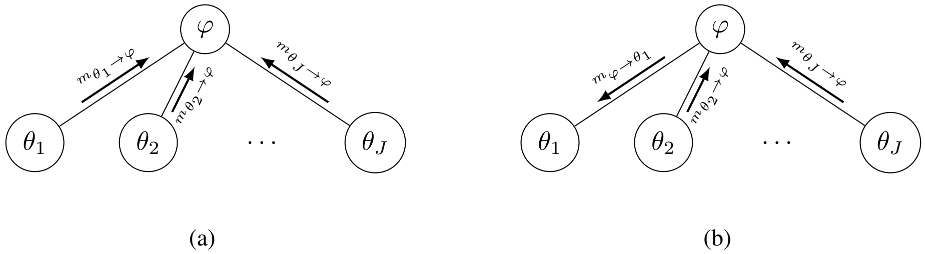
Remark 1. *According to Section 2.2, the density $q^*(\varphi)$, defined in Equation (8), will under suitable conditions become increasingly peaked around some point φ_0, as $J \to \infty$. That $\pi^*_{j'}(\theta_j)$ does not behave similarly, becomes clear by the following considerations. First, we note that Equation (9) is equivalent to*

$$\pi^*_{j'}(\theta_{j'}) = Z^{-1}_{j'}\pi_{j'}(\theta_{j'}) \int_{\Phi} p(\theta_{j'}|\varphi)q^*(\varphi|\Pi_1, \ldots, \Pi_{j'-1}, \Pi_{j'+1}, \ldots, \Pi_J)\, d\varphi,$$

where Z_j is a normalizing constant. As $q^*(\varphi|\Pi_1, \ldots, \Pi_{j'-1}, \Pi_{j'+1}, \ldots, \Pi_J)$ becomes increasingly peaked around φ_0, the integral in the above equation converges to $p(\theta_{j'}|\varphi_0)$. Consequently,

$$\pi^*_{j'}(\theta_{j'}) \to Z^{-1}_{j'}\pi_{j'}(\theta_{j'})p(\theta_{j'}|\varphi_0),$$

which can only be degenerate if either $\pi_{j'}(\theta_{j'})$ or $p(\theta_{j'}|\varphi_0)$ is degenerate by design. Instead of degeneracy, $\pi^*_{j'}(\theta_{j'})$ exhibits shrinkage *with respect to φ_0.*



(a) (b)

Figure 1: Propagating beliefs by messages-passing (a) from each of the leaf nodes to the root node and (b) from each of the nodes $\theta_2, \dots, \theta_J$ to φ, and then finally, from φ to θ_1. The updated beliefs over φ and θ_1 are $q^*(\varphi) \propto q(\varphi) \prod_{j=1}^{J} m_{\theta_j \to \varphi}(\varphi)$ and $\pi_1^*(\theta_1) \propto \pi_1(\theta_1)\, m_{\varphi \to \theta_1}(\theta_1)$, respectively.

3.1 Interpretation as message passing

The formulation of the above meta-analysis framework, constructed as an extension of standard Bayesian inference, can also be viewed within the formalism of probabilistic graphical models. This provides both an intuitive interpretation and a visualization of Equations (8) and (9), and gives a straightforward way of extending the framework to more complex model structures. To elaborate further on this, consider a tree-structured undirected graphical model with J leaf nodes and a root. This is a special case of a pairwise Markov random network [5], where all factors, or *clique potentials*, are over single variables or pairs of variables, referred to as node and edge potentials, respectively. Note that the potential functions are simply non-negative functions, which may not integrate to 1. Choosing, for $j = 1, \dots, J$, the node potentials as $\pi_j(\theta_j)$ and $q(\varphi)$, and the edge potentials as $\psi_j(\theta_j, \varphi) := p(\theta_j|\varphi)$, the model has the joint density

$$\frac{1}{Z} q(\varphi) \prod_{j=1}^{J} \psi_j(\theta_j, \varphi) \pi_j(\theta_j),$$

where Z is a normalizing constant. Finding the marginal density of φ can then be interpreted as propagating beliefs from each of the leaf nodes up to the root node in the form of messages, a process known as *message passing* or belief propagation [e.g. 19]. To that end, we specify the following messages to be sent from the jth leaf node to the root:

$$m_{\theta_j \to \varphi}(\varphi) \propto \int \psi_j(\theta_j, \varphi) \pi_j(\theta_j) d\theta_j. \tag{10}$$

The initial belief $q(\varphi)$ on φ is then updated according to

$$q^*(\varphi) \propto q(\varphi) \prod_{j=1}^{J} m_{\theta_j \to \varphi}(\varphi), \tag{11}$$

which is exactly equal to Equation (8), and illustrated in Figure 1a.

In a similar way, we may pass information to any single leaf node from the remaining leaf nodes. We now specify two kinds of messages: from leaf nodes indexed by $j \in \mathcal{J} \setminus j'$ to the root node, as given by Equation (10), and from the root node to the j'th leaf node,

$$m_{\varphi \to \theta_{j'}}(\theta_{j'}) \propto \int_{\Phi} \psi_{j'}(\varphi, \theta_{j'}) q(\varphi) \prod_{j \in \mathcal{J} \setminus j'} m_{\theta_j \to \varphi}(\varphi)\, d\varphi.$$

The updated belief over $\theta_{j'}$ is then

$$\pi_{j'}^*(\theta_{j'}) \propto \pi_{j'}(\theta_{j'}) m_{\varphi \to \theta_{j'}}(\theta_{j'}), \tag{12}$$

which is exactly equal to Equation (9), and illustrated in Figure 1b.

Although not directly utilized in this work, the graphical model view may also be useful in devising efficient computational strategies. Especially with more complex model structures, making use of the conditional independencies made explicit by the graphical model may bring considerable computational gains.

3.2 Bayesian meta-analysis as a special case

It is straightforward to show that Bayesian random-effects and fixed-effects meta-analyses can be recovered as special cases of the proposed framework. In its traditional formulation [e.g. 10], random-effects meta-analysis (REMA) assumes that for each of J studies, a summary statistic, D_j, $j = 1, \ldots, J$, has been observed, drawn from a distribution with study-specific mean $\mathbb{E}(D_j) = \theta_j$ and variance $\mathrm{Var}(D_j) = \sigma_j^2$:

$$D_j \sim \mathcal{N}(\theta_j, \sigma_j^2), \tag{13}$$

where the approximation of the distribution of D_j by a normal distribution is justified by the asymptotic normality of maximum likelihood estimates. The variances σ_j^2 are directly estimated from the data, while the means θ_j, are assumed to be drawn from some common distribution, typically

$$\theta_j \sim \mathcal{N}(\mu, \sigma_0^2),$$

where the parameters μ and σ_0^2 represent the average treatment effect and inter-study variation, respectively. Fixed-effects meta-analysis is a special case of REMA, where $\sigma_0^2 = 0$, such that $\theta_1 = \theta_2 = \cdots = \theta_J$.

The posterior density for the parameters (μ, σ_0^2) in REMA can be written as

$$\begin{aligned} q(\mu, \sigma_0^2 | D_1, \ldots, D_J) &\propto q(\mu, \sigma_0^2) \prod_{j=1}^{J} \int_{\Theta} N(D_j | \theta_j, \hat{\sigma}_j^2) N(\theta_j | \mu, \sigma_0^2) d\theta_j \\ &\propto q(\mu, \sigma_0^2) \prod_{j=1}^{J} \int_{\Theta} l(\theta_j; D_j) N(\theta_j | \mu, \sigma_0^2) d\theta_j, \end{aligned}$$

where $N(\cdot|\cdot, \cdot)$ denotes a Gaussian density function, $l(\theta_j; D_j)$ is the likelihood function of θ_j given D_j, and $\hat{\sigma}_j^2$ is the empirical variance of D_j. To study the connection between the above posterior density and Equation (8), assume that instead of a summary statistic D_j, each study has been summarized using a posterior distribution with density $\pi_j(\theta_j)$ over its study-specific effect parameter θ_j. If the distribution has been computed under the data model given by equation (13), and using an improper uniform prior $\nu_j(\theta_j) \propto 1$, the density is

$$\pi_j(\theta_j) = N(\theta_j | D_j, \hat{\sigma}_j^2) \propto \exp\left\{ -\frac{(D_j - \theta_j)^2}{2\hat{\sigma}_j^2} \right\} = l(\theta_j; D_j),$$

resulting in the posterior density of (μ, σ_0^2) being equivalent in both cases.

4 Computational strategy

Here we describe a simple computational strategy, which is used in our numerical examples. Some further alternatives are briefly discussed at the end of this section. Recall now that the density of the joint distribution of the parameters $\theta_1, \ldots, \theta_J, \varphi$ can be written as

$$\frac{1}{Z} q(\varphi) \prod_{j=1}^{J} p_j(\theta_j | \varphi) \pi_j(\theta_j). \tag{14}$$

Our goal is to produce joint samples from the above model, enabling any desired marginals to be extracted from them. Probabilistic programming languages [e.g. 2, 11] allow sampling from an arbitrary model, provided that the components of the (unnormalized) model can be specified in terms of probability distributions of some standard form. In the illustrations of this section, we use Hamiltonian Monte Carlo implemented in the Stan probabilistic programming language [2].

We first note that in the above joint model (14), the part specified by the meta-analyst, i.e. $q(\varphi) \prod_{j=1}^{J} p_j(\theta_j | \varphi)$, can by design be composed using standard parametric distributions. The observed part of the model $\prod_{j=1}^{J} \pi_j(\theta_j)$, however, is in general analytically intractable, and instead of having direct access to posterior density functions of standard parametric form, we typically have a sets of posterior samples $\left\{ \theta_j^{(1)}, \ldots, \theta_j^{(L_j)} \right\}$, with $\theta_j^{(l)} \sim \Pi_j$. Our strategy is then to first find an intermediate parametric approximation $\hat{\pi}_j$ for π_j, which enables us to sample from an approximate joint distribution. Assuming that the true densities $\pi_j(\theta_j)$ can be evaluated using e.g. kernel density estimation, and that $\hat{\pi}_j(\theta_j) = 0 \Rightarrow \pi_j(\theta_j) = 0$, the joint samples can be further refined using sampling/importance resampling [SIR; 15]. The steps of the computational scheme are summarized below:

1. For $j = 1, \ldots, J$, fit a parametric density function $\hat{\pi}_j$ to the samples $\left\{\theta_j^{(1)}, \ldots, \theta_j^{(L_j)}\right\}$.
2. Draw M samples $\mathcal{S} = \left\{\theta_1^{*\,(m)}, \ldots, \theta_J^{*\,(m)}, \varphi^{*\,(m)}\right\}_{m=1}^{M}$ from the approximate joint model $\frac{1}{Z'} q(\varphi) \prod_{j=1}^{J} \psi_j(\theta_j, \varphi) \hat{\pi}_j(\theta_j)$.
3. Compute importance weights $w_m = \tilde{w}_m / \sum_{m=1}^{M} \tilde{w}_m$, where

$$\tilde{w}_m = \frac{Z^{-1}\, q\left(\varphi^{*\,(m)}\right) \prod_{j=1}^{J} \psi_j\left(\theta_j^{*\,(m)}, \varphi^{*\,(m)}\right) \pi_j\left(\theta_j^{*\,(m)}\right)}{(Z')^{-1}\, q\left(\varphi^{*\,(m)}\right) \prod_{j=1}^{J} \psi_j\left(\theta_j^{*\,(m)}, \varphi^{*\,(m)}\right) \hat{\pi}_j\left(\theta_j^{*\,(m)}\right)} = \frac{Z' \prod_{j=1}^{J} \pi_j\left(\theta_j^{*\,(m)}\right)}{Z \prod_{j=1}^{J} \hat{\pi}_j\left(\theta_j^{*\,(m)}\right)}.$$

   Note that the constant Z'/Z cancels in the computation of the normalized weights w_m.
4. Resample $\mathcal{S}$ with weights $\{w_1, \ldots, w_M\}$.

For problems with a very large number of studies or high dimensional local parameters, or if the imposed parametric densities approximate the actual posteriors poorly, the computation of importance weights may become numerically unstable. The issue could possibly be mitigated using more advanced importance sampling schemes, such as Pareto-smoothed importance sampling [18]. If we are only interested in sampling from the density of the global parameter, as given by Equation (8), then an obvious alternative strategy would be to implement a Metropolis-Hastings algorithm, using the samples $\theta_j^{(l)} \sim \Pi_j$ to compute Monte Carlo estimates of the integrals $\int_\Theta p(\theta_j|\varphi)\pi_j(\theta_j)d\theta_j$. However, this would lead to expensive MCMC updates as the integrals need to be re-estimated at every iteration of the algorithm. Finally, instead of directly sampling from the full joint distribution, we could try to utilize the induced graphical model structure (Section 3.1) to do localized inference.

5 Numerical illustration: Tuberculosis outbreak dynamics

We now apply MBA to conduct meta-analysis of parameters regulating a stochastic birth-death (SBD) model proposed by Lintusaari *et al.* [8], who used their model in a single-study setting to analyze tuberculosis outbreak data from the San Francisco Bay area, initially reported by Small *et al.* [14]. The goal of the analysis was to estimate disease transmission parameters from genotype data which, in contrast to outbreak models relying on count data,renders the likelihood-function intractable and necessitates the use of likelihood-free inference [16]. Furthermore, such models are often complex in relation to the available data, which may result in poor identifiability, as discussed by Lintusaari *et al.* [7]. To alleviate the problem, they formulated their model as a mixture of stochastic processes, taking into account the individual transmission dynamics of different subpopulations. In our analyses, we focus on two key parameters of the model, R_1 and R_2, which are the reproductive numbers for two subpopulations: those that are compliant and non-compliant to treatment, respectively[1].

For our current experiment, we analyzed three additional data sets using the model of Lintusaari *et al.* [8]. These data sets reported tuberculosis outbreaks in Estonia [6], London [9] and the Netherlands [17]. For each data set, we independently conducted likelihood-free inference, generating 1000 samples from the posteriors. Following Lintusaari *et al.* [8], we used the following six summary statistics for the ABC simulations: the number of observations, the total number of genotype clusters, the size of the largest cluster, the proportion of clusters of size two, the proportion of singleton clusters, and finally, the average successive difference in size among the four largest clusters. The original publication additionally used two summary statistics on the observation times of the largest cluster. While these were found to improve model identifiability, such information was not available for the additional data sets analyzed in our current experiment. We used a weighted Euclidean distance as dissimilarity function, with the same weights as in Lintusaari *et al.* [8].

[1] Note that Lintusaari *et al.* [8] used the notation R_0 for the parameter R_2.

Figure 2 shows the joint posterior distributions for the parameters R_1 and R_2, obtained individually for all four geographical locations using ABC. Compared to the San Francisco data, the posteriors computed on the remaining data sets, in particular London and the Netherlands, show severe problems with identifiability. This can at least partly be attributed to these data sets being less informative than the San Francisco data set. A key question for our experiment then is whether we could borrow strength across the studies to improve the identifiability of the models computed on the remaining data sets. Additionally, it will be of interest to obtain an overall analysis of the central tendency of the reproductive numbers.

As in the MA(q) experiment in the main text, we first define a model for the local effects $\boldsymbol{\theta}_j = (R_{1j}, R_{2j})$, and assign priors for the global mean effect $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and the covariance matrix Σ_0, resulting in the model:

$$
\begin{aligned}
\boldsymbol{\theta}_j &\sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma_0), \\
\mu_1 &\sim \text{Gamma}(a_1, b_1), \\
\mu_2 &\sim \text{Gamma}(a_2, b_2), \\
\Sigma_0 &\sim \mathcal{W}^{-1}(\nu, \Psi).
\end{aligned}
$$

The hyper-parameters were set as follows:

$$
a_1 = 0.12,\ b_1 = 0.36 \quad a_2 = .030,\ b_2 = 0.05, \quad \text{and} \quad \nu = 4, \quad \Psi = \begin{bmatrix} 4 & -0.1 \\ -0.1 & 0.01 \end{bmatrix}.
$$

We then incorporate the observed beliefs by fitting a bivariate Gaussian distribution to each of the $J = 4$ study-specific ABC posteriors. Inference is performed using Stan and the obtained posterior is improved using SIR. Note that due to the indirect nature in the relationship between the infection data and the parameters of interest, using the data to directly construct an estimator for the parameters of interest would be difficult. While this does not pose a challenge in our framework, it renders the application of traditional meta-analysis approaches infeasible.

Figure 3 shows the updated beliefs for the local effects $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$, after borrowing strength across the individual studies. While the updated beliefs clearly retain some of the individual characteristics of their original counterparts (e.g. a similar covariance structure), they exhibit a much more identifiable behavior. The posterior of the overall mean of the reproductive numbers is shown in Figure 4.

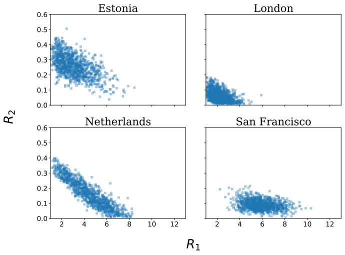


Figure 2: Posteriors for the reproductive numbers (R_1, R_2), individually obtained using ABC in four different studies on tuberculosis outbreak dynamics.

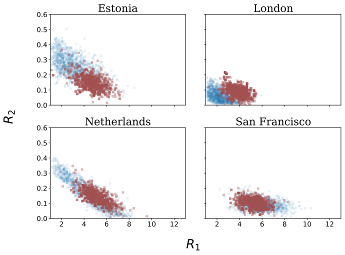


Figure 3: Posteriors for the reproductive numbers (R_1, R_2) updated using MBA (red), plotted on top of the original, individually obtained posteriors (blue).

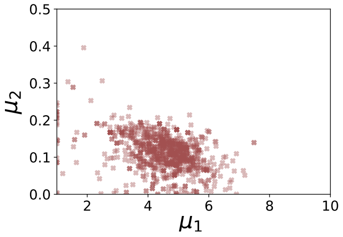


Figure 4: MBA joint posterior for the overall mean effect $\boldsymbol{\mu} = (\mu_1, \mu_2)$.

References

[1] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, Inc., Chichester.

[2] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *J. Statist. Soft.*, **76**(1).

[3] Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability. *J. Am. Statist. Ass.*, **77**(380), 822–830.

[4] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, London, third edition.

[5] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

[6] Krüüner, A., Hoffner, S. E., Sillastu, H., Danilovits, M., Levina, K., Svenson, S. B., Ghebremichael, S., Koivula, T., and Källenius, G. (2001). Spread of drug-resistant pulmonary tuberculosis in estonia. *J. of Clincl Microb.*, **39**(9), 3339–3345.

[7] Lintusaari, J., Gutmann, M. U., Kaski, S., and Corander, J. (2016). On the identifiability of transmission dynamic models for infectious diseases. *Genetics*, **202**(3), 911–918.

[8] Lintusaari, J., Blomstedt, P., Sivula, T., Gutmann, M., Kaski, S., and Corander, J. (2019). Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth-death models [version 1; referees: awaiting peer review]. *Wellcome Open Research*, **4**(14).

[9] Maguire, H., Dale, J. W., McHugh, T. D., Butcher, P. D., Gillespie, S. H., Costetsos, A., Al-Ghusein, H., Holland, R., Dickens, A., Marston, L., Wilson, P., Pitman, R., Strachan, D., Drobniewski, F. A., and Banerjee, D. K. (2002). Molecular epidemiology of tuberculosis in london 1995–7 showing low rate of active transmission. *Thorax*, **57**(7), 617–622.

[10] Normand, S.-L. T. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statist. Medcn*, **18**, 321–359.

[11] Salvatier, J., Wiecki, T., and Fonnesbeck, C. (2015). Probabilistic programming in python using PyMC. *Preprint arXiv:1507.08050*.

[12] Schervish, M. J. (1995). *Theory of statistics*. Springer, New York.

[13] Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall/CRC, Boca Raton.

[14] Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L., and Schoolnik, G. K. (1994). The epidemiology of tuberculosis in San Francisco – a population-based study using conventional and molecular methods. *New Eng. J. Med.*, **330**(24), 1703–1709.

[15] Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, **46**(2), 84–88.

[16] Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. A. (2006). Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, **173**(3), 1511–1520.

[17] van Soolingen, D., Borgdorff, M. W., de Haas, P. E. W., Sebek, M. M. G. G., Veen, J., Dessens, M., Kremer, K., and van Embden, J. D. A. (1999). Molecular epidemiology of tuberculosis in the netherlands: A nationwide study from 1993 through 1997. *J Infect. Dis.*, **180**(3), 726–736.

[18] Vehtari, A., Gelman, A., and Gabry, J. (2015). Pareto smoothed importance sampling. *Preprint arXiv:1507.02646*.

[19] Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *Advances in Neural Information Processing Systems 13*, pages 689–695. MIT Press.