University of Warwick

Markov Melding

Author:
Adam Howes (*1864575*)

Supervisor:
Dr. Murray Pollock

Submitted in partial fulfilment of the requirements
for the degree of *Master of Statistics*

September 5, 2019

Contents

Abstract

Markov melding is a Bayesian generic approach to combining evidence from multiple sources. Each evidence source is modelled by separate submodels, which are then joined by a process called melding. This process, its motivations and methods of inference on the resulting melded model are the primary topics of study in this dissertation. We discuss one simple example of melding, as well as one more involved example involving the synthesis of a Bayesian random-effects meta-analysis with a Bayesian fixed-effects meta-analysis. For the later example, Monte Carlo computational techniques are used to perform inference.

Acknowledgements

I would like to thank Dr Murray Pollock, for his continued support, guidance and encouragement throughout this project. Thanks also go to Dr Robert Goudie for very kindly providing access to the code and data for the paper Goudie et al. (2019).

This dissertation was typeset using `bookdown` (Xie 2016), plots produced using `ggplot2` (Wickham 2016) and written in `R` version 3.5.1 (R Core Team 2018). The code for this dissertation and the algorithms used can be found at the GitHub repo `github.com/athowes/meld`.

Chapter 1

Introduction

It is rare that societally relevant questions can be addressed entirely by a single study (Royal Society, Academy of Medical Sciences 2018). Well-founded policy and informed public discourse therefore depend upon taking into account all available evidence, evenhandedly. More and more, this evidence is in terms of data; with vast amounts now becoming available in all sectors of society. The sources of this data are varied and diverse, making it challenging to analyse simultaneously and often computationally intractable. As a result, there is a need for statistical methodology able to synthesise disparate evidence across multiple sources.

Using all available data has a range of benefits. From a statistical point of view evidence synthesis is typically associated with more precise estimates, just as is an increase in sample size. Any single given study or model typically harbors biases. Integrating multiple models may help to mitigate the effect of bias in any particular model. Furthermore, taking into account multiple sources of evidence gives a fuller and more realistic understanding of the uncertainty involved.

Evidence synthesis is particularly relevant when the required quantity is difficult to measure directly, or otherwise a challenge to obtain complete, reliable, unbiased information about. There are many situations in which it is impossible to carry out a randomised controlled trial or other highly informative study and only partial data sources are available. For instance, in the monitoring of pandemics, only fragmented, biased surveillance data is available (Presanis et al. 2014). Effective public health response to the pandemic depends on up-to-date severity information. Evidence synthesis provides a feasible solution using only the available data.

Meta-analysis, also known as systematic review, is one quantitative approach to combining evidence from multiple studies. Classically, it has been frequentist in nature and typically focused on combining sources of evidence which are in some sense alike, such as multiple studies of the same kind.

Multiparameter Evidence Synthesis (MPES) (Welton et al. 2012) provides a more general approach, set within a Bayesian framework. The only requirement of MPES is that each study is informative about a shared parameter. Bayesian statistics provides a flexible and systematic approach to combining prior knowledge with data from multiple, potentially disparate, sources. The Bayesian paradigm also allows uncertainty to be quantified and propagated through the models such that information is shared between models.

Computational tools such as Markov chain Monte Carlo (MCMC) (Gilks, Richardson, and Spiegelhalter 1995) facilitate fitting complex Bayesian models. Using MCMC, Ades and Cliffe (2002) demonstrate the application of MPES to HIV data in the context of prenatal screening policy. MPES is now the key method (Hickman et al. 2013) used by the UK government to estimate HIV prevalence; accomplished using "a collection of census, surveillance and survey-type prevalence data" (Public Health England 2008).

Markov melding (Goudie et al. 2019) is a recent approach to MPES which introduces a generic framework for combining modular evidence sources. In Chapter 2 we explain the route taken by Goudie et al. (2019) to constructing a joint model over all of the considered sources of evidence. In Chapter 3 we discuss computational methods for conducting inference on this model, applying these methods to synthesising multiple, different meta-analysis motivated by a practical example in clinical trials. Finally, we conclude and discuss directions for further research in Chapter 4.

Chapter 2

Model specification

Markov melding is a Bayesian procedure based on the construction and integration of a collection of graphical models; one for each of the separate evidence components. We begin this chapter by introducing Bayesian statistics in Section 2.1 and outlining how models might be constructed using a Bayesian networks, a type of graphical model, in Section 2.2.

2.1 Bayesian statistics

Consider observable data $Y \in \mathcal{Y}$. Statistical models simplify reality by using parameters $\theta \in \Theta$ to explain the processes governing Y. In Bayesian statistics, the unknown quantities such as Y and θ are treated as random variables. A statistical model $\mathcal{M}$ is then defined entirely by the joint probability distribution $p(\theta, Y)$, encompassing all beliefs about the parameters, data and their interdependence.

This distribution $p(\theta, Y)$ can be decomposed into a product of the likelihood $p(Y \mid \theta)$ and the prior $p(\theta)$ distributions. Having observed the data $Y = y$, the likelihood function $\mathcal{L}(\theta \mid y) = p(y \mid \theta)$ expresses how probable the generation of data y is under parameter setting θ. The prior distribution $p(\theta)$ can be interpreted as embodying subjective judgments about the plausibility of the different parameter settings in advance of seeing the data. Alternatively, from an objective Bayesian point of view, a prior which is in some sense non-informative about the parameters should be chosen. In either setting there is significant freedom in prior-specification which is left to the practitioner.

Statistical inference is the process of learning about the underlying parameters once the data

$Y = y$ has been observed. In Bayesian statistics, this is achieved using probability theory via the eponymous Bayes' rule

$$p(\theta \,|\, y) = \frac{p(y \,|\, \theta)p(\theta)}{p(y)}. \tag{2.1}$$

The posterior distribution $p(\theta \,|\, y)$ represents updated beliefs about the parameters conditional on the observed data. The marginal likelihood $p(y)$ is a constant which normalises the product $p(y \,|\, \theta)p(\theta)$ to a valid probability distribution which integrates to one. Calculating the marginal likelihood is often difficult, preventing exact Bayesian inference from being tractable. However, during this chapter we will concentrate on statistical modelling, mostly ignoring potential computational difficulties for the time being.

Bayes' rule (2.1) can also be written in the form

$$p(\theta \,|\, y) \propto p(y \,|\, \theta)p(\theta) = p(y, \theta), \tag{2.2}$$

omitting the dependence on the marginal likelihood. Indeed, in practice Equation (2.2) is the form of Bayes' rule most widely useful. To demonstrate the application of Bayesian inference to a statistical model we present the following example.

Example 2.1: Poisson likelihood and conjugate Gamma prior

Consider the Poisson-Gamma statistical model $p(\lambda, Y_1, \ldots, Y_n)$ defined as

$$\begin{aligned} &(\textit{Prior}) && \lambda \sim \text{Gamma}(a, b), && (2.3) \\ &(\textit{Likelihood}) && Y_i \sim \text{Pois}(\lambda), \quad i = 1, \ldots, n. && (2.4) \end{aligned}$$

The parameter of interest is the rate $\lambda > 0$, the observable data is $Y = (Y_1, \ldots, Y_n)$ and $a, b > 0$ are fixed hyperparameters (parameters of the prior distribution). Having observed data $y_1, \ldots, y_n$, in terms of probability density functions, the prior and likelihood are

$$\begin{aligned} &(\textit{Prior}) && p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \propto \lambda^{a-1} e^{-b\lambda}, && (2.5) \\ &(\textit{Likelihood}) && \mathcal{L}(\lambda) = \prod_{i=1}^{n} p(y_i \,|\, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{\lambda!} \propto \lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda}. && (2.6) \end{aligned}$$

This particular prior has been chosen as it has functional form matching that of Equation (2.6), making it relatively simple to calculate the posterior distribution using Equation (2.2), as follows

$$\begin{aligned} p(\lambda \,|\, y_1, \dots, y_n) &\propto \mathcal{L}(\lambda)p(\lambda) \propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda} \cdot \lambda^{a-1} e^{-b\lambda} \\ &= \lambda^{a-1+\sum_{i=1}^n y_i} e^{-(b+n)\lambda} \propto \text{Gamma}(a + \sum_{i=1}^n y_i, b + n). \end{aligned} \tag{2.7}$$

After updating conditional on the observed data, the posterior distribution (2.7) is of the same family, a Gamma distribution, as the prior distribution (2.5). This property is called conjugacy and allows exact Bayesian inference to be performed in some situations.

Having observed the data $\{1, 1, 2, 4, 1, 4\}$, Figure (2.1) shows the posterior distribution contracting as beliefs about the rate parameter become stronger. □

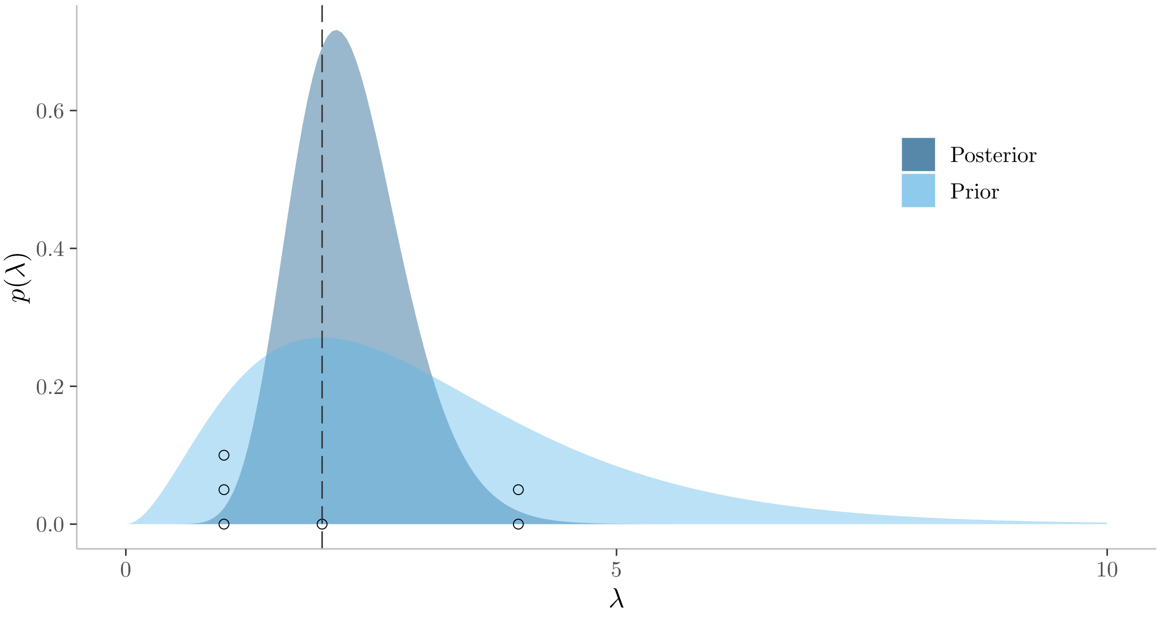


Figure 2.1: Particular example of Bayesian inference for the Poisson-Gamma model. The prior hyperparameters are $a = 3$ and $b = 1$. Data, plotted as open circles, is simulated from a Poisson distribution with $\lambda = 2$ shown as the dashed line

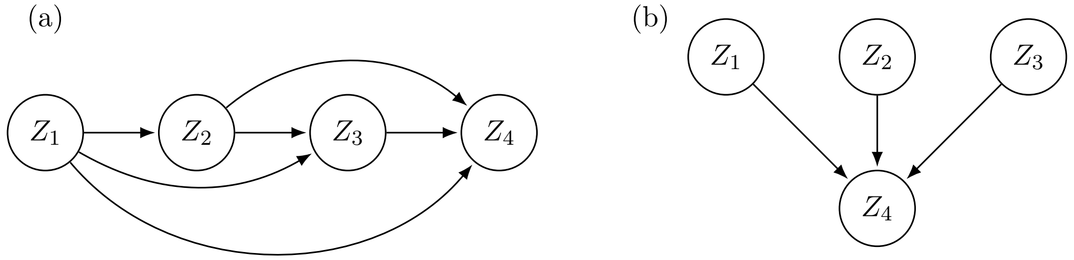


Figure 2.2: Directed acyclic graphs for the four variables $Z_1, \dots, Z_4$.

2.2 Graphical models

The particular features of the joint distribution $p(\theta, Y)$ can be constructed in many ways. Typically both θ and Y are multivariate, so specifying $p(\theta, Y)$ may be no small task. Probabilistic graphical models (PGMs) (Bishop 2006) are one popular and useful tool for this purpose.

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a set of vertices $\mathcal{V}$ together with a set of edges $\mathcal{E}$, where each edge connects a pair of vertices. In a PGM, the vertices represent random variables and the edges represent probabilistic relationships between the random variables. Bayesian networks are a particular type of PGM in which conditional dependencies (defined below) are the probabilistic relationships represented by the edges (Bishop 2006).

Consider random variables $Z_k \in \mathcal{Z}_k$, $k = 1, \dots, K$ with joint distribution $p(Z_1, \dots, Z_K)$. Z_k is conditionally independent from Z_l given Z_m, written $Z_k \perp Z_l \,|\, Z_m$, if and only if

$$p(Z_k, Z_l \,|\, Z_m) = p(Z_k \,|\, Z_m) p(Z_l \,|\, Z_m). \tag{2.8}$$

By repeated conditioning, it is always true that the joint distribution $p(Z_1, \dots, Z_K)$ can be decomposed according to

$$\begin{aligned} p(Z_1, \dots, Z_K) &= p(Z_1) p(Z_2, \dots, Z_K \,|\, Z_1) = p(Z_1) p(Z_2 \,|\, Z_1) p(Z_3, \dots, Z_K \,|\, Z_1, Z_2) \\ &= \cdots = p(Z_1) p(Z_2 \,|\, Z_1) \cdots p(Z_K \,|\, Z_1, \dots, Z_{K-1}). \end{aligned} \tag{2.9}$$

This factorisation can be represented by a particular kind of graph where the edges are directional, shown using arrows, and there are no directed cycles between the nodes in what is known as a directed acyclic graph (DAG). A Bayesian network is a DAG where there is a directed edge from variable Z_k to Z_l if the conditional distribution of Z_l depends on Z_k. For example, the DAG corresponding to Equation (2.9) with $K = 4$ is shown in part (a) of Figure 2.2.

If Equation (2.8) holds then conditional densities of the form $p(Z_k \,|\, Z_l, Z_k)$ can be rewritten to remove the dependence on Z_l as follows

$$p(Z_k \,|\, Z_l, Z_m) = \frac{p(Z_k, Z_l \,|\, Z_m)}{p(Z_l \,|\, Z_m)} = \frac{p(Z_k \,|\, Z_m) p(Z_l \,|\, Z_m)}{p(Z_l \,|\, Z_m)} = p(Z_k \,|\, Z_m). \tag{2.10}$$

Intuitively speaking, once the value of Z_m is known then Z_l is irrelevant to the prediction of Z_k. Equation (2.10) shows that conditional independence statements allow some of the dependencies, and hence directed edges in DAGs, to be removed. Suppose that $K = 4$ as before and each Z_1, Z_2 and Z_3 are conditionally independent from each other given Z_4. Then, Equation (2.9) may be rewritten as

$$p(Z_1, \ldots, Z_K) = p(Z_1) p(Z_2) p(Z_3) p(Z_4 \,|\, Z_1, Z_2, Z_3), \tag{2.11}$$

and the corresponding DAG redrawn taking into account this structure, as shown in part (b) of Figure 2.2. It is visually clear to see that conditional independence statements result in sparser graphs with less interactions, thereby reducing model complexity. More generally, given a set of conditional independence statements, joint probability distributions may be rewritten as

$$p(Z_1, \ldots, Z_K) = \prod_{k=1}^{K} p(Z_k \,|\, \mathrm{pa}(Z_k)), \tag{2.12}$$

where Z_k is only conditionally dependent only on the set of nodes $\mathrm{pa}(Z_k)$, called its parents. In the corresponding DAG, there is a directed edge from each $Z_l \in \mathrm{pa}(Z_k)$ to Z_k. For example, in Figure 2.2, $\mathrm{pa}(Z_4) = \{Z_1, Z_2, Z_3\}$.

Bayesian networks allow complex models $p(Z_1, \ldots, Z_k)$ to be built up from the simpler building blocks of $p(Z_k \,|\, \mathrm{pa}(Z_k))$. These models possess the flexibility required to model the varied data sources necessary for evidence synthesis. Furthermore, domain experts can often efficaciously express their beliefs via Bayesian networks due to their ease of interpretability.

In the context of Bayesian statistics, supposing the prior parameter is multivariate $\theta = (\theta_1, \ldots, \theta_K)$ then the statistical model $p(\theta, Y)$ may then be written as a product of the prior $\prod_{k=1}^{K} p(\theta_k \,|\, \mathrm{pa}(\theta_k))$ with the likelihood $p(Y \,|\, \mathrm{pa}(Y))$. Bayesian models which are built up in stages this way are also known as hierarchical models.

Having introduced the necessary prerequisites, we may now discuss how they might be used to synthesise evidence by combining modular models.

2.3 Modular models

Consider multiple sources of observable data $Y_m \in \mathcal{Y}_m$ with $m = 1, \dots M$. Each Y_m is, at least to some extent, directly informative about a shared parameter ϕ which is common to all of the models. Therefore, for each evidence source, define a statistical model $p_m(\phi, \psi_m, Y_m)$, which will be henceforth called a submodel. For each submodel the parameters are (ϕ, ψ_m): the link parameter $\phi \in \Phi$ together with model specific parameters $\psi_m \in \Psi_m$ unique to each submodel.

It is essential that each of the submodels agree conceptually on what the link parameter ϕ represents: conversations in which the participants do not agree on the basic tenants are rarely productive. It is worth mentioning this fact because in some models the interpretation of parameters can be subtle. For example, in a regression the regression coefficient of a covariate implicitly takes into account the inclusion and exclusion of all other potential covariates; thereby making it a different parameter in different models and rarely an appropriate link parameter. It is therefore safest if ϕ is clearly interpretable, with a conceptual link to the real world.

Given the collection of submodels $p_m(\phi, \psi_m, Y_m)$, the aim of the joining operation in Markov melding (Goudie et al. 2019) is to produce a joint distribution $p(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M)$ over all parameters and observable data. Synthesising a joint distribution is attractive as it facilitates the propagation of uncertainty between the submodels. Additionally, the joint distribution defines a statistical model to which general methodological tools can be applied. Higher quality inference about ϕ is typically the primary aim of evidence synthesis, although due to the sharing of information between the submodels another benefit may be that inference for ψ_m $m = 1, \dots, M$ is also improved.

The joining of submodels is best suited to situations where the model specific observed quantities Y_m are relatively distinct. The more unique information there is dispersed across the different submodels, the greater the potential benefits of evidence synthesis. On the other extreme, the observed data $Y_m = Y$ could be identical for each of the M submodels (with different likelihood components say). This is in some sense unappealing as each of the submodels then offers a competing explanation for the data generating mechanism. Such a collection of beliefs cannot be reasonably held confidently by a single individual simultaneously.

Rather, the individual may have to assign prior probabilities to each of the models being true. Then, if the link parameter plays different roles in the different conflicting submodels, although Markov melding may be possible it might not be advisable. In machine learning, multiple models may be trained on the same data and aggregated for the purposes of prediction (Bishop 2006). However, in this setting the focus is not on model building, interpretation or uncertainty quantification.

Markov melding extends and unifies previous work, particularly Markov combination (Dawid and Lauritzen 1993) and Bayesian melding (Poole and Raftery 2000) which we review here. In particular, Markov combination considers joining submodels where an added simplifying assumption is made; whilst Bayesian melding provides the methodological inspiration for overcoming this assumption.

2.4 Markov combination

As above, consider the collection of submodels $p_m(\phi, \psi_m, Y_m)$, $m = 1, \ldots, M$. Each submodel features a prior distribution on the link parameter ϕ which in the simplest case is specified directly with $p_m(\phi, \psi_m, Y_m) = p_m(\psi_m, Y_m \,|\, \phi) p_m(\phi)$ such that $\mathrm{pa}(\phi) = \emptyset$. Alternatively, the prior can be accessed by marginalising out both ψ_m and Y_m as follows

$$p_m(\phi) = \iint p_m(\phi, \psi_m, Y_m) d\psi_m dY_m. \tag{2.13}$$

An important property of the prior distributions is consistency; where the marginals $p_m(\phi)$ of the link parameter ϕ are called consistent if

$$\forall m \; p_m(\phi) = p(\phi). \tag{2.14}$$

If the prior marginals are consistent then the submodels each have the same beliefs about ϕ in advance of seeing the data. As a result, all of the submodels and their associated beliefs could theoretically reasonably be held by a single individual without any self-contradiction. (That being said, in practice a single practitioner may justifiably place different priors on the same parameter depending on the particular likelihood component. This might be the case if there exists a conjugate prior, as in Example 2.1, for example.)

Supposing that assumption (2.14) holds, Dawid and Lauritzen (1993) define a joint model $p_{\text{comb}}(\phi, \psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M)$ called the Markov combination of the submodels $p_1, \ldots, p_M$.

In advance of justification, this joint model p_{comb} is

$$p_{\text{comb}}(\phi, \psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M) = \frac{\prod_{m=1}^{M} p_m(\phi, \psi_m, Y_m)}{p(\phi)^{M-1}}. \tag{2.15}$$

The construction of p_{comb} is based on prior consistency together with an additional assumption. This assumption is that, conditional on the link parameter, the models are independent. To be exact

$$\forall m \neq \ell \; (\psi_m, Y_m) \perp (\psi_\ell, Y_\ell) \,|\, \phi, \tag{2.16}$$

where conditional independence is defined as in Equation (2.8). Figure (2.3) illustrates this assumption using a DAG for $M = 2$ submodels.

In some sense, the truth of this assumption justifies the use of separate submodels, rather than a single monolithic model which is specified at the outset. Just as with DAGs, this conditional independence statement simplifies the process of modelling: it is easier to specify a collection of submodels than it is to additionally consider their interactions. This is particularly pertinent if the observable data Y_m are from different background areas. In this case, the modelling process naturally decomposes via a consultation of different domain experts, who then specify the submodels for each data source. In this situation, the conditional independence assumption seems, in general, quite reasonable. In general, it is worth noting that conditional independence statements are hard to test (Shah and Peters 2018) and therefore often their validity will rest on more informal "common sense" arguments.

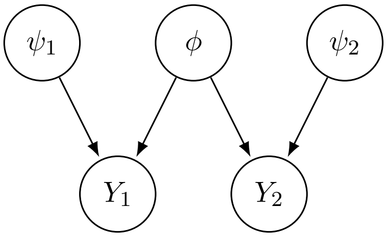


Figure 2.3: Directed acyclic graph corresponding to Equation (2.16) for $M = 2$

The structure of the Markov combination, given by Equation (2.15), is explained by applying these two assumptions - conditional independence and prior consistency - to a hypothetical joint model

$$\begin{aligned}
p(\phi, \psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M) &= p(\phi)\, p(\psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M \mid \phi) \\
&\overset{(2.16)}{=} p(\phi) \prod_{m=1}^{M} p_m(\psi_m, Y_m \mid \phi) \\
&= p(\phi) \prod_{m=1}^{M} \frac{p_m(\phi, \psi_m, Y_m)}{p_m(\phi)} \qquad (2.17) \\
&\overset{(2.14)}{=} \frac{\prod_{m=1}^{M} p_m(\phi, \psi_m, Y_m)}{p(\phi)^{M-1}}, \qquad (2.18)
\end{aligned}$$

where the result is exactly that of Equation (2.15).

Markov combination has the attractive property that submodel marginals and submodel-specific conditional distributions are preserved, that is $p_{\text{comb}}(\phi, \psi_m, Y_m) = p_m(\phi, \psi_m, Y_m)$ and $p_{\text{comb}}(\psi_m, Y_m \mid \phi) = p_m(\psi_m, Y_m \mid \phi)$ for all m (Goudie et al. 2019). This can be shown via

$$\begin{aligned}
p_{\text{comb}}(\phi, \psi_m, Y_m) &= \int p_{\text{comb}}(\phi, \psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M) d\psi_{-m} dY_{-m} \\
&= \int p(\phi) \prod_{m=1}^{M} p_m(\psi_m, Y_m \mid \phi) d\psi_{-m} dY_{-m} \\
&= p(\phi) p_m(\psi_m, Y_m \mid \phi) = p_m(\phi, \psi_m, Y_m), \qquad (2.19)
\end{aligned}$$

where the notation $d\psi_{-m} dY_{-m}$ refers to integration with respect to each of the model specific parameters and observed quantities apart from ψ_m and Y_m. Similarly

$$p_{\text{comb}}(\psi_m, Y_m \mid \phi) = \frac{p_{\text{comb}}(\phi, \psi_m, Y_m)}{p(\phi)} = \frac{p(\phi, \psi_m, Y_m)}{p(\phi)} = p_m(\psi_m, Y_m \mid \phi). \qquad (2.20)$$

In its form, the Markov melded joint model is similar to that of Markov combination. This is because both methods assume the conditional independence assumption, given by Equation (2.16). The primary difference is that Markov melding does not assume that the prior marginals of the link parameter are consistent (as noted in Section 2.1, there is a great deal of flexibility in prior specification - so in general prior consistency can not be expected). Instead, these marginals are "melded" by a process similar to that developed in Bayesian melding, which we will now detail.

2.5 Bayesian melding

The presence of multiple priors on a given quantity also arises in the study of simulation models. Bayesian melding (Poole and Raftery 2000) is motivated by this challenge and aims to perform inference on deterministic simulation models $M : \theta \to \phi$ (the notation in this section in part follows Poole and Raftery (2000), in particular ϕ is not a link parameter). The outputs $\phi \in \Phi$ are a function of the inputs $\theta \in \Theta$ only, such that $\phi = M(\theta)$. There may exist data relating to one or both of θ and ϕ which we denote by Y_θ and Y_ϕ respectively. Irrespective of the simulation model M, a statistical model $p(\theta, \phi, Y_\theta, Y_\phi)$ can be specified which Poole and Raftery (2000) additionally assume can be decomposed into a product of independent prior and likelihood components

$$p(\theta, \phi, Y_\theta, Y_\phi) = p(Y_\theta \mid \theta) p(\theta) p(Y_\phi \mid \phi) p(\phi). \tag{2.21}$$

We call this distribution the joint premodel prior distribution. It may be marginalised to find the premodel prior distributions $p(\theta)$ and $p(\phi)$ respectively.

Bayesian melding builds upon a previous approach called Bayesian synthesis (Raftery, Givens, and Zeh 1995). In this approach, the model M is incorporated into the joint premodel distribution to create a joint postmodel distribution $\pi(\theta, \phi, Y_\theta, Y_\phi)$. Poole and Raftery (2000) describe that this is done by restricting $p(\theta, \phi, Y_\theta, Y_\phi)$ to the submanifold $\{(\theta, \phi, Y_\theta, Y_\phi) : \phi = M(\theta)\}$, such that

$$\pi(\theta, \phi, Y_\theta, Y_\phi) \propto \begin{cases} p(\theta, M(\theta), Y_\theta, Y_\phi), & \text{if } \phi = M(\theta) \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

The marginal postmodel prior distribution for θ is then

$$\pi(\theta) = p(\theta \mid \phi = M(\theta)) \tag{2.23}$$

However, in a discussion of the paper, Wolpert (1995) call attention to the fact that conditional distributions of the form (2.23) are ill-defined and therefore subject to what is known as the Borel-Kolmogorov paradox. A key repercussion of this fact is that the postmodel distribution depends on how M is parametrised. Furthermore, Schweder and Hjort (1996) show that any postmodel distribution can be obtained by arbitrary parametrisation of M. These observations place serious concern on the usefulness of Bayesian synthesis, motivating the development of Bayesian melding.

The problem that Wolpert (1995) find arises because in Bayesian synthesis, as well as the premodel marginal prior distribution $p(\phi)$, there is an additional prior on the output ϕ. In particular, applying the deterministic model M to the premodel marginal prior distribution on inputs $p(\theta)$ results in a model-induced prior on outputs ϕ denoted by $p^\star(\phi)$. If M^{-1} exists then this prior is given by

$$p^\star(\phi) = p(M^{-1}(\phi))|J(\phi)|, \tag{2.24}$$

where $|J(\phi)|$ is the Jacobian of the transformation. The two priors are typically inconsistent as they are based on different information.

A simple approach to rectifying this problem would be not to attempt to place a separate prior on outputs ϕ, instead relying on the prior induced by applying M to $p(\theta)$ However, in hopes of combining both information from the premodel prior and that of the model induced prior, Poole and Raftery (2000) take a different approach: Bayesian melding instead replaces $p(\phi)$ and $p^\star(\phi)$ by a single melded output prior $\tilde{p}(\phi)$. For the particular motivating application, relating to a population dynamics model for bowhead whales, this is done by taking the (normalised) geometric mean of the two priors as follows

$$\tilde{p}(\phi) \propto p(\phi)^{0.5} p^\star(\phi)^{0.5}. \tag{2.25}$$

Equation (2.25) is an example of a broader class of pooling operations which mathematically combine probability distributions. More examples will be discussed further in the next section.

Poole and Raftery (2000) propose then inverting the prior $\tilde{p}(\phi)$ to the input space to define a melded input prior $\tilde{p}(\theta)$ as (omitting some details)

$$\begin{aligned} \tilde{p}(\theta) &= \tilde{p}(M(\theta)) \left(\frac{p(\theta)}{p^\star(M(\theta))} \right) \\ &\propto p(\theta) \left(\frac{p(M(\theta))}{p^\star(M(\theta))} \right)^{1-\alpha}. \end{aligned} \tag{2.26}$$

Equation (2.26) corresponds to the original input prior $p(\theta)$ weighted, according to α, by a ratio of the output prior and the model induced prior evaluated at $\phi = M(\theta)$ for given value of $\theta \in \Theta$.

To conclude discussion of Bayesian melding, having observed $Y_\theta = y_\theta$ and $Y_\phi = y_\phi$ a standard Bayesian posterior for θ can be then defined as follows

$$\pi(\theta \mid y_\phi, y_\theta) \propto p(y_\theta \mid \theta) p(y_\phi \mid M(\theta)) \tilde{p}(\theta), \tag{2.27}$$

allowing inference to proceed as usual.

2.6 Combining expert opinion

The prior pooling step, for which Equation (2.25) is one instance, in Bayesian melding has methodological similarities with previous work on combining the opinions of multiple experts (O'Hagan et al. 2006). Most relevant from this literature is mathematical aggregation, where distributions are elicited independently from each expert individually and then combined by some mathematical rule. The primary alternative to mathematical aggregation is behavioural aggregation, where the group of experts interact and a single distribution is elicited from the group after discussion. In the case of Markov melding, in general the practitioners who originally developed each of the submodels may not be willing and able to justify their choices as would be required in behavioural aggregation.

Just as it is important that the submodels in Markov melding have a shared concept of the link parameter, Clemen and Winkler (1999) caution that "the mathematical and behavioural approaches … assume that the experts have ironed out differences in definitions and that they agree on exactly what is to be forecast or assessed"; also noting that"practising risk analysts know these are strong assumptions“.

In order to outline some of the proposed (O'Hagan et al. 2006) approaches, suppose each of a group of n experts are independently asked their beliefs about an unknown quantity θ, resulting in elicited distributions $f_i(\theta), i = 1, \ldots, n$. The simplest and most widely used technique within mathematical aggregation is opinion pooling, where a consensus distribution $f(\theta)$ is obtained as some function of the individual distributions $\{f_1(\theta), \ldots, f_n(\theta)\}$. Two of the most common examples are the linear opinion pool (2.28) and logarithmic opinion pool (2.29) defined respectively as

$$f(\theta) \propto \sum_{i=1}^{n} w_i f_i(\theta), \tag{2.28}$$

$$f(\theta) \propto \prod_{i=1}^{n} f_i(\theta)^{w_i}, \tag{2.29}$$

where $f(\theta)$ is normalised to a probability density function. The weights w_i can be chosen freely, either giving more weight to some experts than others or choosing to weight all of the experts equally.

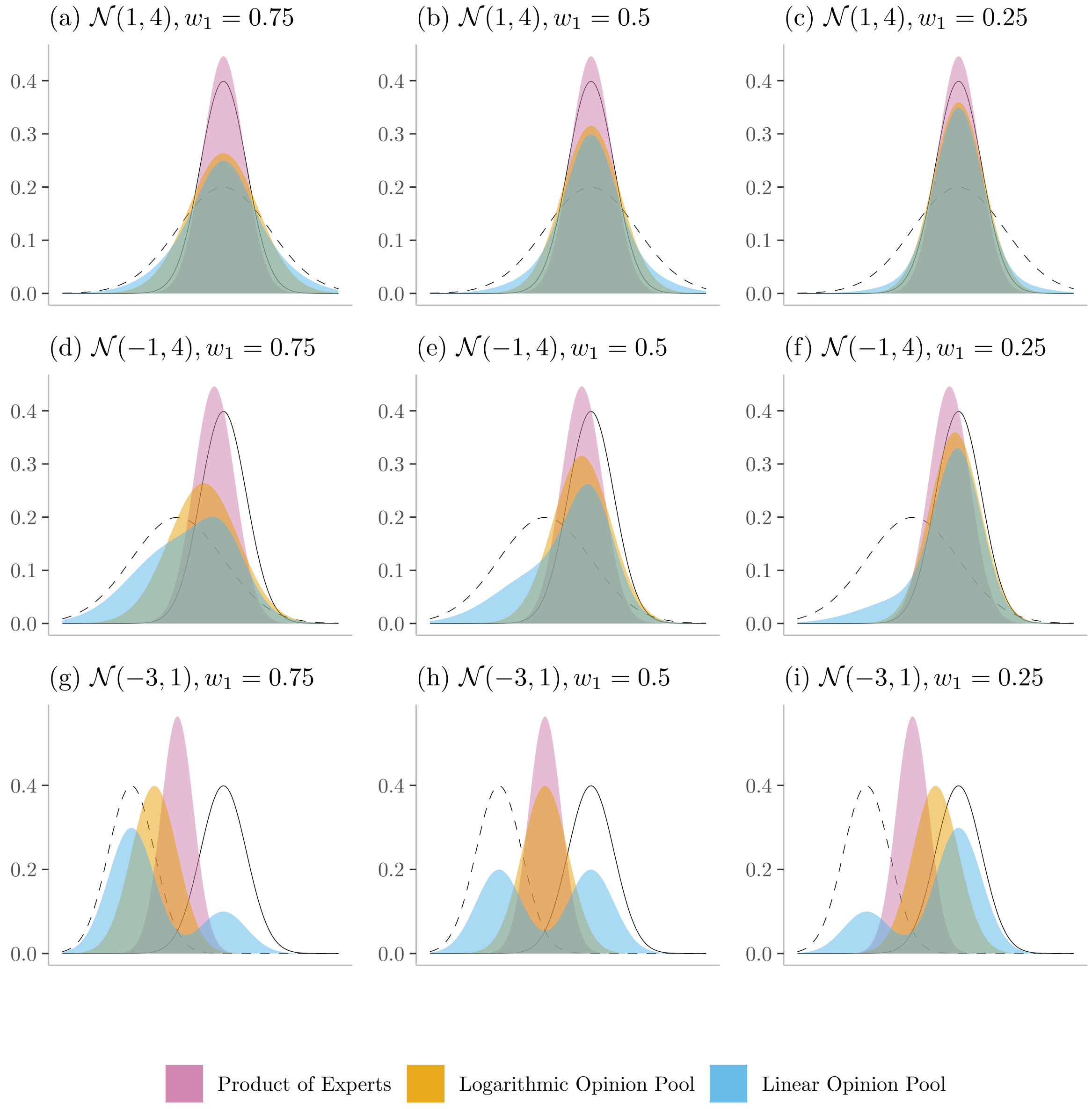
Dictatorial pooling, in which one experts opinion is chosen as the consensus distribution, is special case of linear opinion pooling. To see this, set $w_j = 0$ for all $j \neq i$ in Equation (2.28) such that $f(\theta) = f_i(\theta)$. In Section 2.5, the suggestion to set $\tilde{p}(\phi) = p^{\star}(\phi)$ is an example of dictatorial pooling. The method used by Poole and Raftery (2000), geometric pooling given by Equation (2.25), is an example of logarithmic pooling with $w_i = 1/M$. Another special case of logarithmic opinion pooling is the product of experts (PoE) pooling (Hinton 2002) which sets $w_i = 1$ for all i such that

$$f(\theta) \propto \prod_{i=1}^{n} f_i(\theta). \tag{2.30}$$

Figure (2.4) shows the PoE, logarithmic and linear opinion pooling applied to various Gaussian distributions. Note that in the PoE pooling, there are no weights. Or, another way of looking at it, the weights are fixed to be $w_1 = w_2 = 1$.

The PoE is named as such because it gives each expert the benefit of the doubt in assuming that their beliefs are correct. The result, as most clearly illustrated by row (g, h, i) of Figure (2.4), is that the regions where the experts overlap has to be the truth. Logarithmic pooling is slightly more doubtful of the experts but still produces relatively contracted pooled beliefs in comparison to the more cautious linear pooling. O'Hagan et al. (2006) find that "while the linear opinion pool has been quite widely used in practice, the logarithmic opinion pool has been largely ignored, perhaps because it is perceived to lead to unrealistically strong aggregated beliefs".

Row (a, b, c) of Figure (2.4) prompt another consideration: to what extend should expert agreement lead to higher confidence. To use an analogy, suppose you and a group of friends want to decide whether or not to eat at a certain restaurant. If each friend expresses a positive opinion about the restaurant then may seem reasonable to be more confident than any given



(a) $\mathcal{N}(1,4), w_1 = 0.75$

(b) $\mathcal{N}(1,4), w_1 = 0.5$

(c) $\mathcal{N}(1,4), w_1 = 0.25$

(d) $\mathcal{N}(-1,4), w_1 = 0.75$

(e) $\mathcal{N}(-1,4), w_1 = 0.5$

(f) $\mathcal{N}(-1,4), w_1 = 0.25$

(g) $\mathcal{N}(-3,1), w_1 = 0.75$

(h) $\mathcal{N}(-3,1), w_1 = 0.5$

(i) $\mathcal{N}(-3,1), w_1 = 0.25$

Figure 2.4: Opinion pooling of solid line $\mathcal{N}(1,1)$ with dashed line Gaussians, specified by plot text. Dashed line Gaussians are weighted w_1 with $w_2 = 1 - w_1$. Reproduced with alternations from Goudie et al. (2019)

friend about the restaurant. However, if each friend bases their opinion solely on a positive review of the restaurant in that week's paper then this effectively would be double counting; in reality there is only one "expert" - the journalist. As such, the more that the experts opinions contain novel, independent information the more suitable PoE pooling is. On the other hand, if the experts opinions are similar to those of our friends in the restaurant example then perhaps linear opinion pooling may more fitting.

2.7 Markov melding

In Section 2.4 we detailed the approach of Markov combination, which requires prior consistency. By using ideas from Bayesian melding to overcome inconsistent priors, in the following section we will show that Markov melding naturally extends Markov combination.

If the marginals are inconsistent, that is condition (2.14) fails to hold, then the approach of Markov combination requires some modification. Just as pooling is used to reconcile disagreement between the premodel prior and induced prior in Bayesian melding, the inconsistent prior marginals $p_m(\phi)$ can be also pooled to allow Bayesian inference to proceed. The resulting pooled prior $p_{\text{pool}}(\phi)$ is such that

$$p_{\text{pool}}(\phi) = g(p_1(\phi), \dots, p_M(\phi)), \tag{2.31}$$

for some choice of pooling function g. As discussed in Section 2.6, many possible pooling functions are possible and could be reasonably justified depending on the situation. That being said, some choices of pooling function may be more computational efficient and so attractive from a practical point of view.

The Markov melded joint model p_{meld} can be achieved by replacing $p(\phi)$ in Equation (2.17) with the pooled prior (2.31) such that

$$p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) = p_{\text{pool}}(\phi) \prod_{m=1}^{M} \frac{p_m(\phi, \psi_m, Y_m)}{p_m(\phi)} \tag{2.32}$$

This construction is equivalent to altering each of the submodels by a procedure Goudie et al. (2019) call marginal replacement and then applying Markov combination. In particular, the marginal $p_m(\phi)$ of the original submodel is replaced by $p_{\text{pool}}(\phi)$ resulting in the replacement submodel

$$p_{\text{repl,m}}(\phi, \psi_m, Y_m) = p_m(\psi_m, Y_m \mid \phi) p_{\text{pool}}(\phi). \tag{2.33}$$

After marginal replacement has occurred it is possible to apply Markov combination to the collection of replacement submodels $\{p_{\text{repl,m}}\}_{m=1,\dots,M}$ since they have consistent marginals, such that

$$\begin{aligned}
p_{\text{comb}}(\phi, \psi_1, \dots, \psi_M, Y_1, \dots, Y_M) &= \frac{\prod_{m=1}^M p_{\text{repl,m}}(\phi, \psi_m, Y_m)}{p_{\text{pool}}(\phi)^{M-1}} \\
= \frac{\prod_{m=1}^M p_m(\psi_m, Y_m \mid \phi) p_{\text{pool}}(\phi)}{p_{\text{pool}}(\phi)^{M-1}} &= p_{\text{pool}}(\phi) \prod_{m=1}^M p_m(\psi_m, Y_m \mid \phi) \\
= p_{\text{pool}}(\phi) \prod_{m=1}^M \frac{p_m(\psi_m, Y_m, \phi)}{p_m(\phi)}. &
\end{aligned} \tag{2.34}$$

The result is exactly equal to that of the Markov melded joint model, Equation (2.32).

Since a joint model has been obtained, Bayesian inference can proceed as usual. Given observed data $Y_m = y_m$ for $m = 1, \dots, M$, under the Markov melded model (2.32) the joint posterior distribution is proportional to the joint distribution

$$p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M \mid y_1, \dots, y_M) \propto p_{\text{pool}}(\phi) \prod_{m=1}^M \frac{p_m(\phi, \psi_m, y_m)}{p_m(\phi)}, \tag{2.35}$$

which is called the Markov melded posterior.

We now present the following pedagogical example of Markov melding with $M = 2$ submodels which, as with Example 2.1, uses conjugacy to enable exact Bayesian computation.

Example 2.2: Binomial and Geometric submodels with conjugate Beta priors

Consider the following two submodels, illustrated as DAGs in Figure 2.5.

Submodel 1 $p_1(\theta, Y_1)$ is defined as

$$\theta \sim \text{Beta}(a, b), \tag{2.36}$$

$$Y_1^i \sim \text{Bin}(m, \theta), \quad i = 1, \dots, n_1, \tag{2.37}$$

where $a = 2$, $b = 3$ and $m = 10$ are known and $Y_1 = (Y_1^1, \dots, Y_1^{n_1})$.

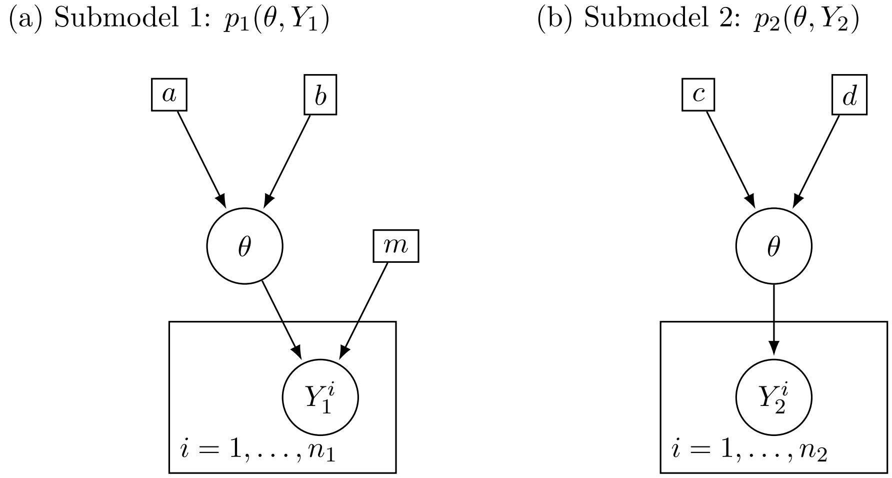


Figure 2.5: DAGs representing submodels one and two. Further to the notation introduced in Section 2.2, constants may be included in DAGs with square boxes, whereas random variables have circular boxes. Rectangular boxes (sometimes called plates) are drawn around sections of the diagram which are repeated across the indices indicated

Suppose the true value of the parameter θ is 0.4. We simulate observed data $Y_1 = y_1$ of size $n_1 = 3$ from this submodel, resulting in $y_1 = \{3, 5, 4\}$. Conditioning on the observed data, inference can be done exactly as the prior is conjugate. The posterior distribution is then given by

$$\begin{aligned} p_1(\theta \,|\, y_1) \propto p_1(y_1 \,|\, \theta) p_1(\theta) &= \prod_{i=1}^{n_1} \binom{m}{y_1^i} \theta^{y_1^i} (1-\theta)^{m-y_1^i} \cdot \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \theta^{a-1} (1-\theta)^{b-1} \\ &\propto \theta^{\tau_1 + a - 1} (1-\theta)^{n_1 m - \tau_1 + b - 1} \propto \text{Beta}(\tau_1 + a, n_1 m - \tau_1 + b), \end{aligned} \tag{2.38}$$

where $\sum_{i=1}^{n_1} y_1^i = \tau_1$. Using the specified values of the hyperparameters and observed data the posterior distribution can be found by this equation. Figure 2.6 part (a) shows both the prior and posterior plotted in blue: essentially this submodel is entirely accurate about θ.

Submodel 2 The second submodel $p_2(\theta, Y_2)$ is such that

$$\theta \sim \text{Beta}(c, d) \tag{2.39}$$

$$Y_2 \sim \text{Geo}(\theta), \quad i = 1, \ldots, n_2, \tag{2.40}$$

where $c = 5$ and $d = 4$ are known and $Y_1 = (Y_2^1, \ldots, Y_2^{n_2})$.

We again simulate observed data $Y_2 = y_2$, this time of size $n_2 = 6$ resulting in $y_2 = \{2, 0, 1, 0, 0, 1\}$. This submodel, like the one before, is conjugate and the posterior distribution is simply

$$\begin{aligned} p_2(\theta \,|\, y_2) \propto p_2(y_2 \,|\, \theta) p_2(\theta) &= \prod_{i=1}^{n_2} (1-\theta)^{y_2^i} \theta \cdot \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)} \theta^{c-1} (1-\theta)^{d-1} \\ &\propto \theta^{n_2+c-1} (1-\theta)^{\tau_2+d-1} \propto \text{Beta}(n_2 + c, \tau_2 + d), \end{aligned} \tag{2.41}$$

where $\sum_{i=1}^{n_2} y_2^i = \tau_2$. Figure 2.6 part (b) shows that in this submodel, both the prior and posterior, are somewhat misguided with respect to θ.

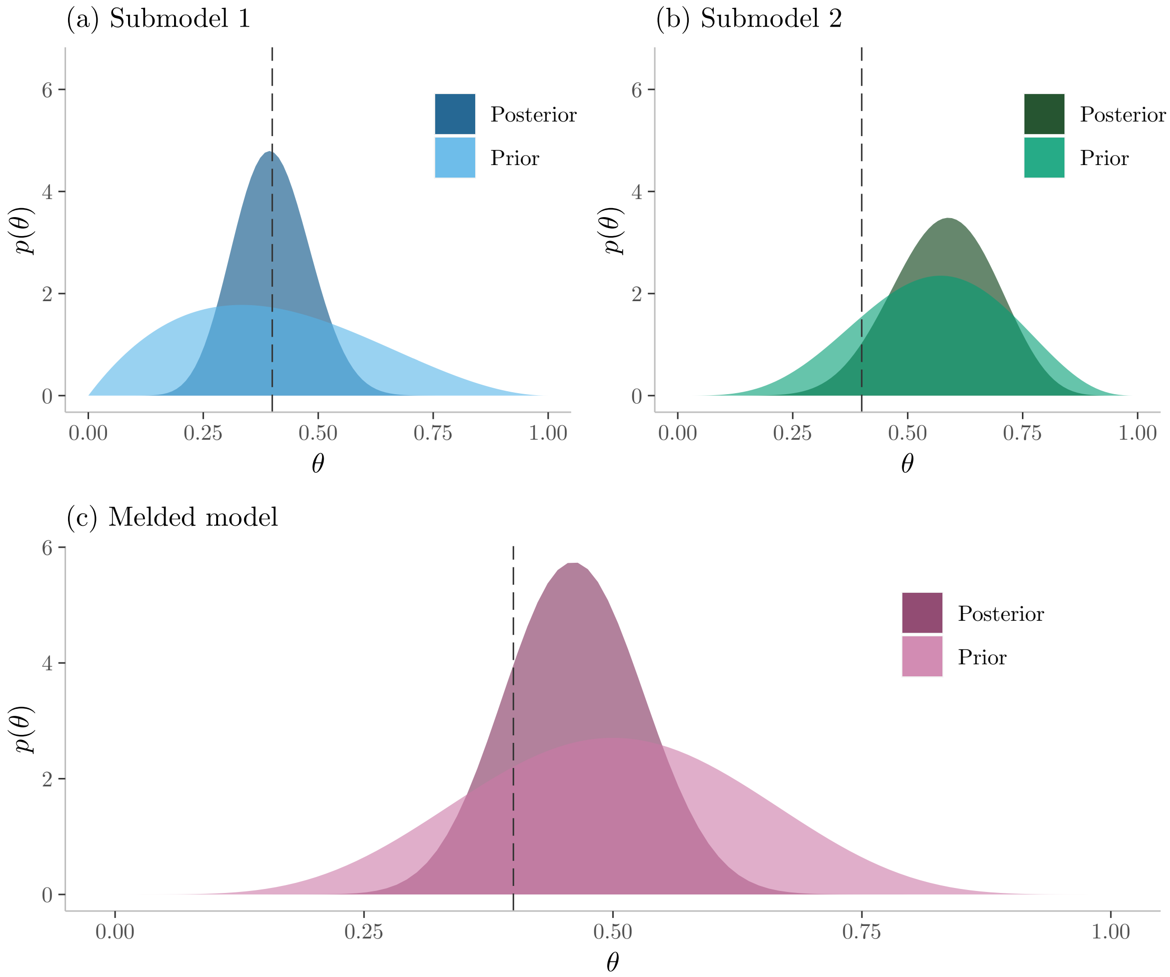


Figure 2.6: Parts (a) and (b) show the standard Bayesian updating of the prior for submodels one and two respectively. Part (c) shows the PoE pooled prior and Markov melded posterior

Melded model This is a simple situation: the link parameter is $\phi = \theta$, and there are no model specific parameters $\psi_1 = \psi_2 = \emptyset$. Notice with this hyperparameter setting, there is not consistency between the two Beta priors, given by Equations (2.36) and (2.39), on the link parameter. For this reason, Markov melding is an appropriate methodological choice to perform inference.

To facilitate Markov melding, we pool the inconsistent prior distributions $p_1(\theta)$ and $p_2(\theta)$. Using the PoE pooling (2.30) gives the pooled prior to be simply another Beta distribution

$$\begin{aligned} p_{\text{pool}}(\theta) &\propto \text{Beta}(a, b) \cdot \text{Beta}(c, d) \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^{c-1}(1-\theta)^{d-1} \\ &= \theta^{a+c-2}(1-\theta)^{b+d-2} \propto \text{Beta}(a+c-1, b+d-1). \end{aligned} \tag{2.42}$$

Using this pooled prior, by Equation (2.35) the relevant Markov melded posterior is

$$\begin{aligned} p_{\text{meld}}(\theta \mid y_1, y_2) &\propto p_{\text{pool}}(\theta) \prod_{m=1}^{2} \frac{p_m(\theta \mid y_m)}{p_m(\theta)} \propto p_1(\theta) p_2(\theta) \prod_{m=1}^{2} \frac{p_m(\theta \mid y_m)}{p_m(\theta)} \\ &\propto p_1(\theta \mid y_1) p_2(\theta \mid y_2) \\ &\propto \text{Beta}(\tau_1 + a + n_2 + c - 1, n_1 m - \tau_1 + b + \tau_2 + d - 1). \end{aligned} \tag{2.43}$$

Notice that in this instance the pooled prior cancels with the prior marginals to simplify the expression. In fact, in general under PoE pooling, the melded posterior is proportional to the product of the submodel posteriors

$$\begin{aligned} p_{\text{meld}}(\phi, \psi_1, \ldots, \psi_M \mid y_1, \cdots, y_M) \propto p_{\text{pool}}(\phi) \prod_{m=1}^{M} \frac{p_m(\phi, \psi_m, y_m)}{p_m(\phi)} \\ = \prod_{m=1}^{M} p_m(\phi, \psi_m, y_m) \propto \prod_{m=1}^{M} p_m(\phi, \psi_m \mid y_m). \end{aligned} \tag{2.44}$$

Figure 2.6 shows that both the melded prior (2.42) and Markov melded posterior (2.43) occupy a centrist position between the two submodels.

Suppose rather than PoE pooling we use logarithmic pooling with weights w_m for submodel $m = 1, 2$ such that $w_1 + w_2 = 1$. The pooled prior is then

$$\begin{aligned} p_{\text{pool}}(\theta) &\propto \text{Beta}(a, b)^{w_1} \cdot \text{Beta}(c, d)^{w_2} \propto \theta^{w_1(a-1)}(1-\theta)^{w_1(b-1)} \cdot \theta^{w_2(c-1)}(1-\theta)^{w_2(d-1)} \\ &\propto \text{Beta}(w_1 a + w_2 c - w_1 - w_2 + 1, w_1 b + w_2 d - w_1 - w_2 + 1) \propto \text{Beta}(e, f), \quad (2.45) \end{aligned}$$

where $e = w_1 a + w_2 c - w_1 - w_2 + 1$ and $f = w_1 b + w_2 d - w_1 - w_2 + 1$. Using this pooled prior, the Markov melded posterior is given by

$$\begin{aligned} p_{\text{meld}}(\theta \mid y_1, y_2) &\propto p_{\text{pool}}(\theta) \prod_{m=1}^{2} \frac{p_m(\theta \mid y_m)}{p_m(\theta)} \\ &\propto \text{Beta}(e, f) \frac{\text{Beta}(\tau_1 + a, n_1 m - \tau_1 + b)}{\text{Beta}(a, b)} \frac{\text{Beta}(n_2 + c, \tau_2 + d)}{\text{Beta}(c, d)} \\ &\propto \theta^{e + \tau_1 + n_2 - 1}(1-\theta)^{f + n_1 m - \tau_1 + \tau_2 - 1} \\ &\propto \text{Beta}(e + \tau_1 + n_2, f + n_1 m - \tau_1 + \tau_2) \quad (2.46) \end{aligned}$$

Note that setting $w_1 = w_2 = 1$ gives, as in Equation (2.43), a $\text{Beta}(a + c + \tau_1 + n_2 - 1, b + d + n_1 m - \tau_1 + \tau_2 - 1)$ distribution. Figure 2.7 shows the results of Markov melding for three weight settings. For the prior, in parts (a) and (c) the weighting amounts to dictatorial pooling to submodels one and two respectively. The prior in part (b), in contrast, is a geometric mean between the two submodel priors. As it happens, under these circumstances of the particular pooling is of little (although not no) consequence to the Markov melded posterior. □

In this chapter we have motivated the form of the Markov melded posterior distribution and shown via Example 2.2 that, for a simple tractable case, it behaves as one would hope. To extend our discussion into more realistic situations, in the following chapter we will introduce statistical simulation techniques that allow inference to be conducted on generic Markov melded posterior distributions.

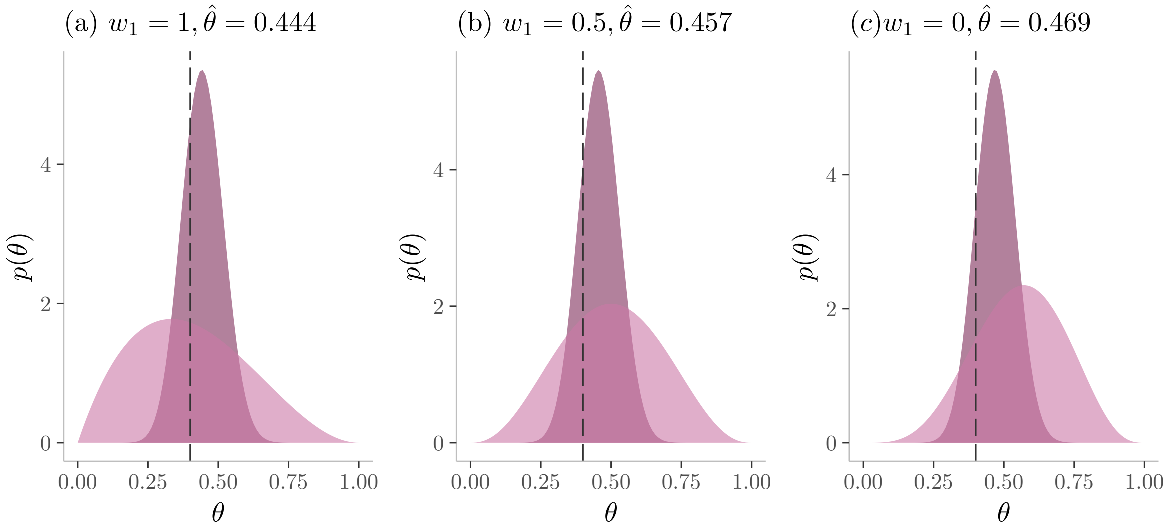


Figure 2.7: Markov melding using logarithmically pooled priors with various weights. The weighting w_1, with $w_2 = 1 - w_1$, as well as the posterior mean $\hat{\theta} = \mathbb{E}[\theta \mid y_1, y_2]$ is given above each plot

Chapter 3

Computation

The primary goal in Bayesian computation is to find expectations of functions with respect to the posterior distribution, or in our case the Markov melded posterior given by Equation (2.35). Many relevant quantities, such as probabilities or risks, can be cast as integrals as the following example demonstrates.

Example 3.1: Bayesian decision theory

In assessing developments in evidence-based medicine, Ashby and Smith (2000) argue that a Bayesian approach is the natural framework for decision making under uncertainty. Suppose $\mathcal{A}$ is the space of possible decisions with elements $a \in \mathcal{A}$ and there exists a utility function $U : \mathcal{A} \to \mathbb{R}^+$ weighing up the drawbacks and benefits of each decision. Then, it is clear to see that it is optimal to select the decision $a^\star$ which maximises this utility. Leaving aside the problem of specifying U, most (if not all) of the time the utility also depends on additional variables which we are uncertain about. If these additional variables are (leadingly) denoted by $\theta \in \Theta$ then our definition of U may be expanded so that $U : \mathcal{A} \times \Theta \to \mathbb{R}^+$. Now, in order to make well-informed decisions, we must take into account our uncertainty about θ. Having observed data $y \in \mathcal{Y}$ informative about θ, the distribution which represents this uncertainty is the posterior $p(\theta \,|\, y)$. Rather than computing utilities, we now must calculate expected utilities. This is done by integrating out our uncertainty about θ as follows

$$\mathbb{E}_{p(\theta \,|\, y)}[U(a, \theta)] = \int U(a, \theta) p(\theta \,|\, y) d\theta. \tag{3.1}$$

The optimal decision now is to select the action maximising this expected utility. The upshot is that Equation (3.1) amounts to computing expectations with respect to the posterior. □

3.1 Monte Carlo methods

Monte Carlo, a general class of simulation-based methods, are a prevailing technique for scalable computation of integrals like that of Equation (3.1). For ease of notation rather than the posterior, we consider a general probability density function $\pi(x)$ where $x \in \mathcal{X}$. The relevant integral is then

$$\mathbb{E}_{\pi}[\varphi(x)] = \int \varphi(x)\pi(x)dx \tag{3.2}$$

where $\varphi : \mathcal{X} \to \mathbb{R}$ is an arbitrary test function.

By sampling points independently from the probability density function $X_i \sim \pi(x)$ for $i = 1, \ldots, N$, $\pi(x)$ may be approximated by the empirical distribution

$$\widehat{\pi}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}(x). \tag{3.3}$$

This facilitates the approximation of expectations by using $\widehat{\pi}(x)$ in place of $\pi(x)$ such that

$$\mathbb{E}_{\pi}[\varphi(x)] = \int \varphi(x)\pi(x)dx \approx \int \varphi(x)\widehat{\pi}(x)dx = \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i) := \bar{\varphi}_N. \tag{3.4}$$

By the strong law of large numbers, granted the expectation exists, then the Monte Carlo estimator will converge

$$\lim_{N\to\infty} \bar{\varphi}_N = \mathbb{E}_{\pi}[\varphi(x)] \tag{3.5}$$

almost surely. In addition, the rate of convergence can be monitored by estimating the variance of the Monte Carlo estimator $\bar{\varphi}_N$ as

$$\sigma_N^2 = \frac{1}{N^2}\sum_{i=1}^{N} \left(\varphi(X_i) - \bar{\varphi}_N\right)^2, \tag{3.6}$$

and then applying the central limit theorem such that

$$\frac{\bar{\varphi}_N - \mathbb{E}_{\pi}[\varphi(x)]}{\sigma_N} \sim \mathcal{N}(0, 1). \tag{3.7}$$

These principles, both the approximation of the empirical distribution via Equation (3.3) and the convergence of the Monte Carlo estimator, are illustrated in Figure (3.1) for a simple exponential distribution with rate parameter $\lambda = 1$.

3.2 Markov chain Monte Carlo

It is typically not possible to generate independent samples from the relevant distribution directly. This is especially true for complex models and in high dimensions, which often will be the case for Markov melding.

Markov chain Monte Carlo (MCMC) is a flexible and broad class of algorithms for sampling from such probability distributions which otherwise cannot be accessed. In MCMC, rather than drawing independent samples, the sample at iteration i depends on that of the previous iteration $i - 1$; the result being a a Markov chain of samples. By careful specification, the stationary distribution of this Markov chain can be set to correspond to the distribution π - here referred to as the target distribution. As a result, if the Markov chain is run until it reaches convergence then it can, more-or-less, be used in place of direct samples from π. There are many diagnostic tools which can be used to check if the chain appears to have converged, the simplest being heuristic examination of traceplots which show the positions of the chain over time. However, convergence is not something which is possible to prove so discretion must be taken.

Due to the dependence on the previous value, samples from a Markov chain are typically positively correlated, thereby reducing the amount of information each sample contains about π. This autocorrelation should be taken into account when calculating the standard error of Monte carlo estimates, see for example Gelman and Rubin (1992). One popular metric of this sort is the effective sample size (ESS) (C. Robert and Casella 2013) which can be interpreted as the number of independent samples the correlated samples would be equivalent to.

In the following sections we discuss a range of MCMC algorithms and how they might be applied to Markov melding.

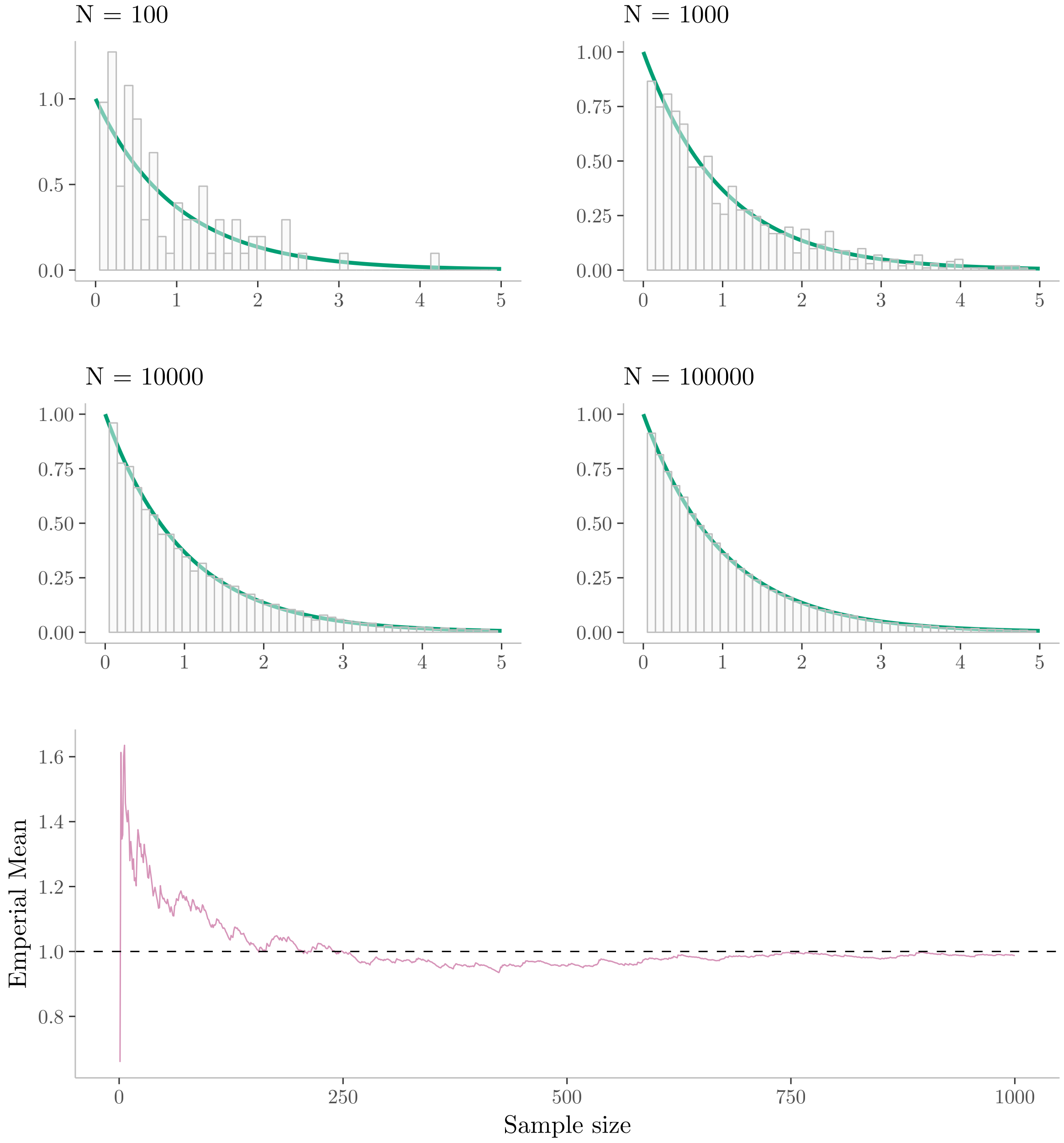


Figure 3.1: Above: Emperical distributions generated by samples $X_i \sim \text{Exp}(1)$ for $i = 1, \ldots, N$ for N of varying order of magnitude with exact Exp(1) distribution plotted in green. Below: Convergence of the empirical mean, in pink, to the true mean plotted as dashed grey line

3.2.1 Metropolis-Hastings

Algorithm 1: Metropolis-Hastings sampler

```
Target density $\pi$, initialise $x^{(0)}$;
for $i = 1, 2, \ldots, N$ do
    Set current value $x = x^{(i-1)}$;
    Draw candidate value $x^\star \sim q(\cdot \mid x)$;
    Draw $u \sim \mathcal{U}(0,1)$;
    Let $\alpha = \min(1, r)$, where $r = \frac{\pi(x^\star)/q(x^\star \mid x)}{\pi(x)/q(x \mid x^\star)}$;
    if $\alpha > u$ then
        $x^{(i)} = x^\star$;
    else
        $x^{(i)} = x$;
end for
```

The Metropolis-Hastings (MH) algorithm dates back to foundational work of Metropolis et al. (1953) and Hastings (1970). At each iteration i, the next step $x^\star$ is suggested according to a user-specified proposal distribution $q(\cdot \mid x)$ conditional on the current location $x = x^{(i-1)}$. Candidate values $x^\star$ are accepted with probability α, else the Markov chain remains at the current location x. The value of the acceptance probability α is calculated by

$$\begin{aligned}\alpha &= \mathbb{P}[x^\star \text{ accepted}] = \min(1, r) \\ &= \min\left(1, \frac{R(x^\star, x)}{R(x, x^\star)}\right) = \min\left(1, \frac{\pi(x^\star)/q(x^\star \mid x)}{\pi(x)/q(x \mid x^\star)}\right), \end{aligned} \tag{3.8}$$

where r is the ratio of target-to-proposal density ratios R at the candidate $x^\star$ and current x values.

In Bayesian statistics typically the target distribution π is such that

$$\pi(x) = \frac{\gamma(x)}{Z}, \tag{3.9}$$

where the normalising constant $Z = \int \gamma(x)dx$ is computationally intractable. To be specific, π is the product of the likelihood and prior and Z is the marginal likelihood. Therefore, it is crucial that Equation (3.8) only requires the point-wise evaluation of γ and not π since $\pi(x)/\pi(z) = \gamma(x)/\gamma(z)$.

The efficiency of MH depends greatly on the user's choice of proposal distribution $q(\cdot \mid x)$. A common option is to consider proposal distributions which are symmetric about the current value x, thereby simplifying the acceptance probability $\alpha = \min\{1, \pi(x^\star)/\pi(x)\}$ since $q(x \mid x^\star) = q(x^\star \mid x)$. This algorithm is called random walk Metropolis (RWM). Roberts, Gelman, and Gilks (1997) show that, under certain conditions, the acceptance rate for RWM which minimises autocorrelation is 0.44 for one-dimensional proposals tending asymptotically to 0.234 as the dimension increases.

By setting $\pi(x) = p_{\text{meld}}(\phi, \psi_1, \ldots, \psi_M \mid y_1, \ldots, y_M)$ the general MH algorithm can be applied to target the Markov melded posterior. The proposal distribution is then of the form $q(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star \mid \phi, \psi_1, \ldots, \psi_M)$, where $(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star)$ are the candidate values and $(\phi, \psi_1, \ldots, \psi_M)$ are the current values of the chain.

The acceptance probability $\alpha = \min(1, r)$ for moves $(\phi, \psi_1, \ldots, \psi_M) \rightarrow (\phi^\star, \psi_1^\star, \ldots, \psi_M^\star)$ can be calculated via

$$r = \frac{R(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star, \phi, \psi_1, \ldots, \psi_M)}{R(\phi, \psi_1, \ldots, \psi_M, \phi^\star, \psi_1^\star, \ldots, \psi_M^\star)}, \tag{3.10}$$

where $R(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star, \phi, \psi_1, \ldots, \psi_M)$ is the target-to-proposal density ratio

$$R(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star, \phi, \psi_1, \ldots, \psi_M) = \frac{p_{\text{pool}}(\phi^\star) \prod_{m=1}^{M} \frac{p_m(\phi^\star, \psi_m^\star, y_m)}{p_m(\phi^\star)}}{q(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star \mid \phi, \psi_1, \ldots, \psi_M)} \tag{3.11}$$

For Markov melding, due to the possible quantity of parameters $(\phi, \psi_1, \ldots, \psi_M)$, it is likely difficult to find a suitable choice of proposal distribution $q(\phi^\star, \psi_1^\star, \ldots, \psi_M^\star \mid \phi, \psi_1, \ldots, \psi_M)$ such that moves are accepted with reasonable probability.

3.2.2 Metropolis-within-Gibbs

The general MH algorithm ignores the modular structure of the submodels. As a result both of this structure and the difficulty of full parameter moves, it may be more appealing to update submodel components separately.

Gibbs sampling (S. Geman and Geman 1987) is an MCMC algorithm which updates componentwise. In fact, Gibbs sampling is a special case of MH where the acceptance rate can be calculated to be identically one. Returning to the more general notation, suppose that x can be written as a p-dimensional vector such that $x = (x_1, \ldots, x_p)$. In Gibbs sampling at each iteration i only one of the components x_j where $j \in 1, \ldots p$ is updated. In particular, x_j is sampled from its full conditional distribution, defined as

$$\pi_{X_j \,|\, X_{-j}}(x_j | x_{-j}) = \frac{\pi(x)}{\int \pi(x) dx_j}, \tag{3.12}$$

where the notation x_{-j} refers to the subvector of x excluding the jth element. There are multiple ways of choosing the component j to update at each iteration: it can be chosen uniformly using random-scan Gibbs sampling, or iterated over deterministically using systematic-scan Gibbs sampling. Having selected a component to update x_j, informally speaking Gibbs sampling moves perpendicular to the directions spanned by each of the other components x_{-j}. If components x_j and x_k are say positively correlated then increases in x_j are typically accompanied by increases in $x_k, k \neq j$. In this situation, Gibbs sampling may not be efficient as it struggles to find the relevant region of probability mass using perpendicular moves.

From a computational point of view, requiring that the full conditionals of π are available and can be sampled from is often restrictive. For Markov melding, the full conditionals of p_{meld} may not be available and as a result generic Gibbs sampling is not generally applicable. Instead, Goudie et al. (2019) propose the use of the Metropolis-within-Gibbs (Muller 1991) (MWG) algorithm.

Algorithm 2: (Random-scan) Metropolis-within-Gibbs sampler

```
Target density $\pi$, initialise $x^{(0)} := (x_1^{(0)}, \dots, x_p^{(0)})$;
for $i = 1, 2, \dots, N$ do
    Set current value $x = x^{(i-1)}$;
    Draw $j \sim \mathcal{U}\{1, \dots, p\}$;
    Draw $x_j^\star \sim q_j(\cdot \mid x)$;
    Set $x^\star = (x_1, \dots, x_j^\star, \dots, x_p)$;
    Draw $u \sim \mathcal{U}(0, 1)$;
    Let $\alpha = \min(1, r)$, where $r = \frac{\gamma(x^\star)/q(x^\star \mid x)}{\gamma(x)/q(x \mid x^\star).}$;
    if $\alpha > u$ then
        $x^{(i)} = x^\star$;
    else
        $x^{(i)} = x$;
end for
```

MWG unifies MH and Gibbs sampling by utilising componentwise MH proposals rather than sampling from the full conditionals exactly. As such, MWG does not require access to the full conditionals and as a result is more generally applicable. Of course, if for some index j the full conditional is available then it is a valid MH update and can be used.

As with MH, the user must specify the proposal distributions, in this case $q_j(\cdot \mid x_j^{(i-1)})$ for each of the p components. For random walk type proposals the resulting algorithm known as random walk Metropolis-within-Gibbs (RWM-within-Gibbs). This algorithm is studied in the context of optimal scaling by Neal and Roberts (2006) who show that, as with RWM, the optimal acceptance rate is 0.44 for one-dimensional proposals again tending to 0.234 for higher dimension as with Roberts, Gelman, and Gilks (1997).

Now, to target Equation (2.35) using MWG, define the components of $(\phi, \psi_1, \dots, \psi_M)$ be the link parameter ϕ and each of the model specific parameters ψ_m, $m = 1, \dots, M$ respectively as would be expected. Suppose the initial values $(\phi^{(0)}, \psi_1^{(0)}, \dots, \psi_M^{(0)})$ are given.

Model specific parameter updates In Markov melding, it is assumed (2.16) that the ψ_m are conditionally independent given ϕ. Supposing this assumption is reasonable, it would therefore be expected that Gibbs and MWG sampling would be relatively efficient for model specific parameters. For each of ψ_m for $m = 1, \dots, M$ define a proposal distribution $q_m(\cdot \mid \psi_m)$. The target-to-proposal density ratio is then

$$\begin{aligned}
&R(\phi, \psi_1, \dots, \psi_m^\star, \dots, \psi_M, \phi, \psi_1, \dots, \psi_m, \dots, \psi_M) \\
&= p_{\text{pool}}(\phi) \prod_{j=1}^{M} \frac{p_j(\phi, \psi_j, y_j)}{p_j(\phi)} \frac{1}{q_m(\psi_m^\star \mid \psi_m)} \\
&= p_{\text{pool}}(\phi) \prod_{j \neq m} \frac{p_j(\phi, \psi_j, y_j)}{p_j(\phi)} \cdot \frac{p_m(\phi, \psi_m^\star, y_m)}{p_m(\phi)} \frac{1}{q_m(\psi_m^\star \mid \psi_m)}.
\end{aligned} \tag{3.13}$$

In calculating r via (3.10) factors not dependent on submodel m, as well as the pooled prior, cancel leaving

$$r = \frac{p_m(\phi, \psi_m^\star, y_m) \cdot \frac{1}{q_m(\psi_m^\star \mid \psi_m)}}{p_m(\phi, \psi_m, y_m) \cdot \frac{1}{q_m(\psi_m \mid \psi_m^\star)}}. \tag{3.14}$$

Equation (3.14) precisely corresponds to updating ψ_m conditional on the link parameter ϕ for inference performed on the m^{th} alone. With each submodel having been specified as it is, it's reasonable to expect that inference for each submodel individually is tractable via MH. As such, it is possible to calculate the ratio in Equation (3.14) for each $m = 1, \dots, M$.

Link parameter updates In contrast to the above, one would expect the link parameter ϕ to have some covariance structure with each of the model specific parameters: potentially making link parameter updates relatively inefficient. For the link parameter ϕ define a proposal distribution $q(\cdot \mid \phi)$. The target-to-proposal density ratio is

$$R(\phi, \psi_1, \dots, \psi_M, \phi^\star, \psi_1, \dots, \psi_M) = p_{\text{pool}}(\phi^\star) \prod_{m=1}^{M} \frac{p_m(\phi^\star, \psi_m, y_m)}{p_m(\phi^\star)} \cdot \frac{1}{q(\phi^\star \mid \phi)}. \tag{3.15}$$

If PoE pooling is used then the marginal distributions of the link parameter cancel, leaving

$$\begin{aligned}
R(\phi, \psi_1, \dots, \psi_M, \phi^\star, \psi_1, \dots, \psi_M) &\propto \prod_{m=1}^{M} p_m(\phi^\star) \prod_{m=1}^{M} \frac{p_m(\phi^\star, \psi_m, y_m)}{p_m(\phi^\star)} \cdot \frac{1}{q(\phi^\star \mid \phi)} \\
&= \prod_{m=1}^{M} p_m(\phi^\star, \psi_m, y_m) \cdot \frac{1}{q(\phi^\star \mid \phi)}.
\end{aligned} \tag{3.16}$$

However, in general, in order to calculate Equation (3.15) the prior marginal distributions $p_m(\phi)$ and the closely related p_{pool} must be evaluated. If $p_m(\phi)$ is not tractable, Goudie et al. (2019) suggest using an approximation $\hat{p}_m(\phi)$ found using kernel density estimation with samples drawn from the prior marginal given by Equation (2.13). The samples can be produced by forward Monte Carlo: simulating the statistical relationships (which typically are standard distributions) top-down in the DAG representation of the submodel until the node corresponding to ϕ is reached.

In the following example, we demonstrate the use of MCMC for Markov melding two different types of meta-analysis. For full details of the code used, see the dissertation repository.

Example 3.2: Metropolis-within-Gibbs for Markov melding

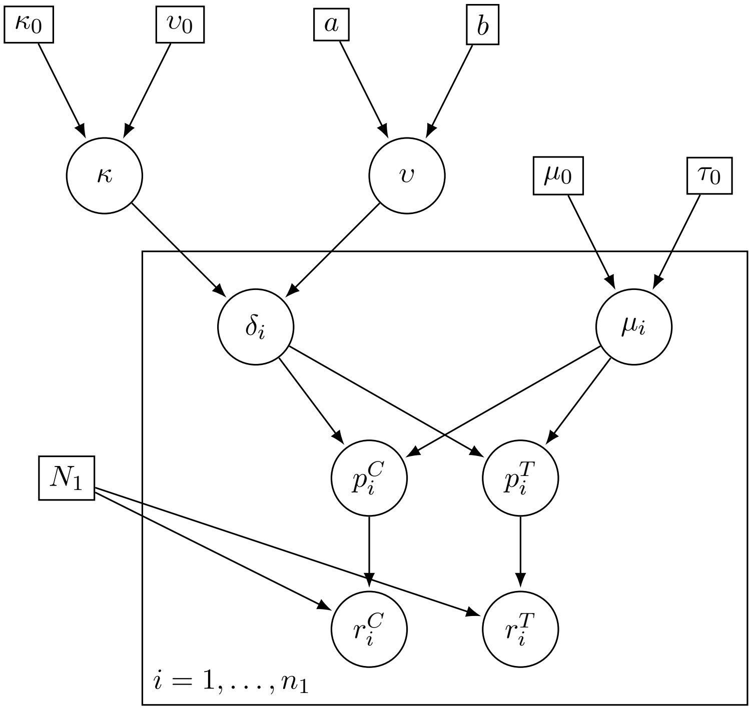


Figure 3.2: DAG representing Submodel 1

Submodel 1 Standard meta-analysis usually assumes that either effects are common across studies (fixed-effects) or drawn from a probability distribution (random-effects). T. Smith, Spiegelhalter, and Thomas (1995) present a fully Bayesian approach to random-effects meta-analysis. They use this model to study the effectiveness of a treatment (selective decontamination of the digestive tract) which aims to prevent patients acquiring infections whilst in intensive care units (ICUs). We present an adapted version of the model as follows.

Consider $n_1 = 5$ trials $i = 1, \ldots, n_1$ each with both a control (C) and treatment (T) group of size $N_1 = 10$. In trial i, individuals within group $j \in \{C, T\}$ are assumed to have independent probability p_i^j of developing an infection. The numbers of infections in the control and treatment groups for trial i are r_i^C and r_i^T respectively. Assume that μ_i is the average infection rate for the i^{th} trial and δ_i is the true treatment effect, both on the log-odds scale. Following T. Smith, Spiegelhalter, and Thomas (1995), we place an uninformative Gaussian prior on each μ_i, with fixed prior mean $\mu_0 = 0$ and precision $\tau_0 = 0.25$ hyperparameters. The prior on each δ_i is also Gaussian with mean κ and precision τ. Uninformative hyperpriors (3.17) and (3.18) are placed on both κ and τ with $\kappa_0 = 0$, $\tau = 0.1$ and $a = 3$, $b = 1$. Clinical trials aim to assess treatment effectiveness and therefore the mean (log-odds scale) treatment effect κ is the parameter of interest. The full submodel $p_1(\kappa, \upsilon, \delta, \mu, p^C, p^T, r^C, r^T)$ is then

$$\begin{align}
\kappa &\sim \mathcal{N}(\kappa_0, \upsilon_0^{-1}), \tag{3.17}\\
\upsilon &\sim \text{Gamma}(a, b), \tag{3.18}\\
\delta_i &\sim \mathcal{N}(\kappa, \upsilon^{-1}), \tag{3.19}\\
\mu_i &\sim \mathcal{N}(\mu_0, \tau_0^{-1}), \tag{3.20}\\
p_i^C &= \frac{e^{\mu_i - \delta_i/2}}{1 + e^{\mu_i - \delta_i/2}},\ p_i^T = \frac{e^{\mu_i + \delta_i/2}}{1 + e^{\mu_i + \delta_i/2}}, \tag{3.21}\\
r_i^C &\sim \text{Bin}(N_1, p_i^C),\ r_i^T \sim \text{Bin}(N_1, p_i^T), \quad i = 1, \ldots, n_1, \tag{3.22}
\end{align}$$

where the vectors $\delta, \mu, p^C, p^T, r^C, r^T$ are each of length n_1, e.g. $\delta = (\delta_1, \ldots, \delta_{n_1})$. Observed data for Submodel 1 is $Y_1 = (r^C, r^T)$. For this example, we forward simulate the observed data $Y_1 = y_1$ (with $N_1 = 10$ and $n_1 = 5$) from $\delta_i \sim \mathcal{N}(-0.25, 4^{-1}),\ \mu_i \sim \mathcal{N}(-1, 2^{-1})$, resulting in $r^C = \{8, 8, 11, 6, 5\},\ r^T = \{2, 2, 5, 11, 4\}$.

The probabilities $p = (p_i^C, p_i^T)$ are an (invertible) deterministic transformation of (μ_i, δ_i). Therefore, each probability has an induced distribution via Equation (2.24). Had a prior been specified on p then the situation would have been identical to that of Poole and Raftery (2000) where there exist multiple priors on parameters in a deterministic simulation model. To resolve issues of this sort Goudie et al. (2019) propose using Bayesian melding (introduced in Section 2.5) however in this instance we choose not to place a prior on p; implicitly accepting the distribution induced by μ_i and δ_i. In particular, for p_i^C (a similar statement is true for p_i^T) we have the inverse transformation

$$\mu_i - \delta_i/2 = \text{logit}(p_i^C) = \log \frac{p_i^C}{1 - p_i^C}, \tag{3.23}$$

with corresponding Jacobian

$$\left|\frac{\delta(\mu_i - \delta_i/2)}{\delta p_i}\right| = \frac{1}{p_i^C(1-p_i^C)}. \tag{3.24}$$

As linear transformations of Gaussian distributions, $\mu_i \pm \delta_i/2$ are Gaussian with mean and variance given by

$$\mathbb{E}[\mu_i - \delta_i/2|\mu_0, \tau_0, \kappa, \upsilon] = \mu_0 - \kappa/2, \tag{3.25}$$

$$\mathbb{E}[\mu_i + \delta_i/2|\mu_0, \tau_0, \kappa, \upsilon] = \mu_0 + \kappa/2, \tag{3.26}$$

$$\mathbb{V}[\mu_i \pm \delta_i/2|\mu_0, \tau_0, \kappa, \upsilon] = \tau_0^{-1} + \upsilon^{-1}/4. \tag{3.27}$$

Therefore, defining $(\tau_0^{-1} + \upsilon^{-1}/4)^{-1} = \omega$, the distributions of p_i^C and p_i^T are given by

$$p(p_i^C \mid \mu_0, \tau_0, \kappa, \upsilon) = \frac{\omega}{\sqrt{2\pi}} \exp\left(-\frac{\omega}{2}\left(\log\frac{p_i^C}{1-p_i^C} - (\mu_0 - \kappa/2)\right)^2\right)\frac{1}{p_i^C(1-p_i^C)}, \tag{3.28}$$

$$p(p_i^T \mid \mu_0, \tau_0, \kappa, \upsilon) = \frac{\omega}{\sqrt{2\pi}} \exp\left(-\frac{\omega}{2}\left(\log\frac{p_i^T}{1-p_i^T} - (\mu_0 + \kappa/2)\right)^2\right)\frac{1}{p_i^T(1-p_i^T)}. \tag{3.29}$$

Having calculated distributions (3.28) and (3.29) dependence on (μ, δ) can be omitted. Given (r^C, r^T) (and additionally omitting dependence on fixed hyperparameters) the posterior distribution is then

$$
\begin{aligned}
& p_1(\kappa, \upsilon, p^C, p^T \mid r^C, r^T) \\
\propto & \prod_{i=1}^{n_1} \binom{N_1}{r_i^C} p_i^{C r_i^C} (1 - p_i^C)^{N_1 - r_i^C} \frac{\omega}{\sqrt{2\pi}} \exp\left(-\frac{\omega}{2} \left(\log \frac{p_i^C}{1 - p_i^C} - (\mu_0 - \kappa/2) \right)^2 \right) \frac{1}{p_i^C (1 - p_i^C)} \\
& \cdot \prod_{i=1}^{n_1} \binom{N_1}{r_i^T} p_i^{T r_i^T} (1 - p_i^T)^{N_1 - r_i^T} \frac{\omega}{\sqrt{2\pi}} \exp\left(-\frac{\omega}{2} \left(\log \frac{p_i^T}{1 - p_i^T} - (\mu_0 + \kappa/2) \right)^2 \right) \frac{1}{p_i^T (1 - p_i^T)} \\
& \cdot \frac{\tau_0}{\sqrt{2\pi}} \exp\left(-\frac{\tau_0}{2} (\mu - \mu_0)^2 \right) \cdot \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) \\
\propto & \prod_{i=1}^{n_1} \{ p_i^{C r_i^C - 1} (1 - p_i^C)^{N_1 - r_i^C - 1} \} \, \omega^{n_1} \exp\left(-\frac{\omega}{2} \sum_{i=1}^{n_1} \left(\log \frac{p_i^C}{1 - p_i^C} - (\mu_0 - \kappa/2) \right)^2 \right) \\
& \cdot \prod_{i=1}^{n_1} \{ p_i^{T r_i^T - 1} (1 - p_i^T)^{N_1 - r_i^T - 1} \} \, \omega^{n_1} \exp\left(-\frac{\omega}{2} \sum_{i=1}^{n_1} \left(\log \frac{p_i^T}{1 - p_i^T} - (\mu_0 + \kappa/2) \right)^2 \right) \\
& \cdot \exp\left(-\frac{\upsilon_0}{2} (\kappa - \kappa_0)^2 \right) \upsilon^{a-1} \exp(-b\upsilon)
\end{aligned}
\tag{3.30}
$$

Often it is more computationally convenient to work in terms of logarithms as this allows products to be replaced with sums. In particular, by differencing the logarithm of two parameter settings and then exponentiating the result we obtain acceptance probabilities in a more numerical stable way. The log-posterior is then

$$
\begin{aligned}
& \log p(\kappa, \upsilon, p^C, p^T \mid r^C, r^T) \\
\propto & \sum_{i=1}^{n_1} \{ (r_i^C - 1) \log p_i^C + (N_1 - r_i^C - 1) \log(1 - p_i^C) \\
& + (r_i^T - 1) \log p_i^T + (N_1 - r_i^T - 1) \log(1 - p_i^T) \\
& - \frac{\omega}{2} \left(\log \frac{p_i^C}{1 - p_i^C} - (\mu_0 - \kappa/2) \right)^2 - \frac{\omega}{2} \left(\log \frac{p_i^T}{1 - p_i^T} - (\mu_0 + \kappa/2) \right)^2 \} \\
& + 2 n_1 \omega - \frac{\upsilon_0}{2} (\kappa - \kappa_0)^2 + (a - 1) \log \upsilon - b\upsilon := l_1.
\end{aligned}
\tag{3.31}
$$

To target the posterior (3.30) we use RWM-within-Gibbs. The initialisation values are chosen to be 0.5 for each of the probabilities and otherwise chosen according to the prior means $(\kappa^{(0)}, \upsilon^{(0)}) = (\kappa_0, a/b) = (0, 1/3)$. The proposal distribution for each component is Gaussian with individual componentwise standard deviation parameters tuned to achieve close to the optimal one-dimensional acceptance rate 0.44 (Roberts, Gelman, and Gilks (1997) note that there is “little value” in finely tuning algorithms to exact values and so the tuning is not done exhaustively - see Table 3.1).

Table 3.1: Submodel 1 scaling and resultant acceptance rates for each component

	κ	υ	p_1^C	p_2^C	p_3^C	p_4^C	p_5^C	p_1^T	p_2^T	p_3^T	p_4^T	p_5^T
σ	3.000	3.500	0.300	0.250	0.300	0.250	0.200	0.150	0.150	0.250	0.300	0.200
Acceptance	0.427	0.452	0.384	0.407	0.398	0.381	0.411	0.438	0.437	0.406	0.397	0.447

To determine an appropriate number of Markov chain samples for the problem at hand one method is to consider the ESS. To this end, Vats, Flegal, and Jones (2015) calculate a lower bound on the number of effective samples required to estimate a vector of length p with $100(1-\alpha)\%$ confidence and a relative tolerance of ϵ. Additionally, they define a multivariate ESS which takes into account cross-correlation across the Markov chain components. Both these methods are implemented in the `R` package `mcmcse` (J. M. Flegal et al. 2017).

In order to estimate the mean, of length $p = 12$, with 97.5% confidence and a tolerance of $\epsilon = 0.025$ the required number of effective samples is approximately 40000. Based on a short trial chain, the number of simulations required to achieve this ESS is 2800000 (although this may seem high, at every iteration only one component is proposed and therefore the chain is highly auto-correlated). That being said, the algorithm is not prohibitively computationally intensive, so we choose to run 5000000 iterations.

Table 3.2 shows that κ, υ and 6/10 of the probability parameters are estimated such that the truth lies less than one posterior standard deviation from the posterior mean. Traceplots and histograms for all of the parameters are plotted, for example κ and υ in Figure 3.3. Based on this output, it would appear that the Markov chains reach convergence.

Table 3.2: Submodel 1 posterior mean (PM), true value and posterior standard deviation (PSD) for each component

	κ	υ	p_1^C	p_2^C	p_3^C	p_4^C	p_5^C	p_1^T	p_2^T	p_3^T	p_4^T	p_5^T
PM	-0.314	2.991	0.359	0.265	0.502	0.219	0.172	0.124	0.124	0.262	0.546	0.215
Truth	-0.250	4.000	0.214	0.360	0.464	0.296	0.236	0.134	0.324	0.308	0.421	0.221
PSD	1.229	1.731	0.102	0.093	0.106	0.088	0.079	0.068	0.069	0.093	0.107	0.087

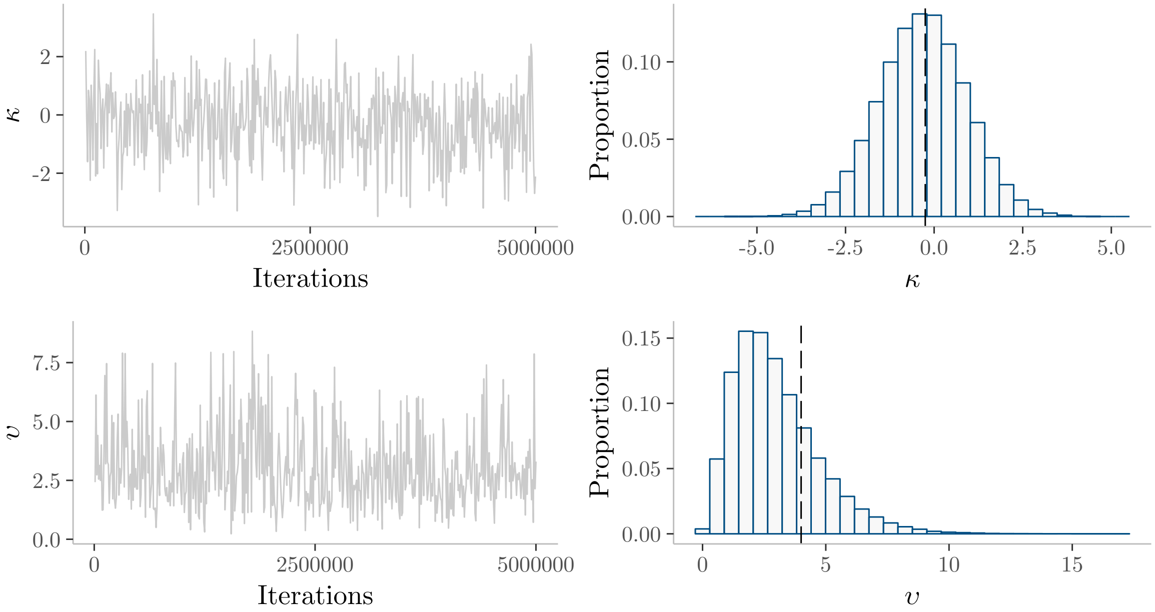


Figure 3.3: Submodel 1 posterior traceplots and histograms for the parameters κ and υ, with truth shown on the histogram as a dashed black line (as it will be for all histograms to follow). Similar plots for the ten other probability parameters are presented in Appendix A. Note that for this, and all subsequent, traceplots the chains have first been thinned for graphical display reasons

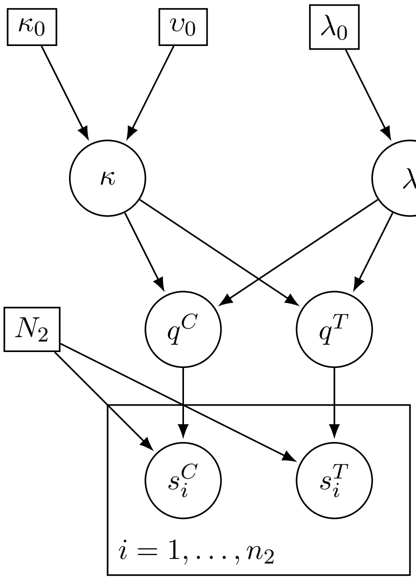


Figure 3.4: DAG representing Submodel 2

Submodel 2 Suppose that a second collection of $n_2 = 10$ trials of the same treatment have been conducted in a different region with treatment and control groups of size $N_2 = 15$. It is believed that there is less variability in the quality of hospitals and therefore that a fixed-effects meta-analysis is the most appropriate model for this data.

The full submodel $p_2(\kappa, \lambda, q^C, q^T, s^C, s^T)$ is given by

$$\kappa \sim \mathcal{N}(\kappa_0, \upsilon_0^{-1}), \tag{3.32}$$

$$\lambda \sim \mathcal{N}(\lambda_0, \gamma_0^{-1}), \tag{3.33}$$

$$q^C = \frac{e^{\lambda - \kappa/2}}{1 + e^{\lambda - \kappa/2}}, \ q^T = \frac{e^{\lambda + \kappa/2}}{1 + e^{\lambda + \kappa/2}}, \tag{3.34}$$

$$s_i^C \sim \mathrm{Bin}(N_2, q^C), \ s_i^T \sim \mathrm{Bin}(N_2, q^T), \quad i = 1, \ldots, n_2, \tag{3.35}$$

where s^T and s^C are vectors of length n_2 such that the observed data for this submodel is $Y_2 = (s^T, s^C)$. This submodel has a similar structure to Submodel 1, the difference informally being that the plate has been moved down only to include the trial outcomes.

Assume the authors of this study have more informative prior beliefs (following a subjective Bayesian approach) about κ, setting $\kappa_0 = -0.1$ and $\upsilon_0 = 0.5$. The prior on λ is similar to that of each of the μ_i in Submodel 1, such that $\lambda_0 = 0$ and $\gamma_0 = 0.25$.

As with the first submodel, we forward simulate observed data by setting $\kappa = -0.25$ and

$\lambda = -1$. The observed data $Y_2 = y_2$ is then $s^C = \{3,4,5,7,3,7,7,5,5,2\}$ and $s^C = \{3,4,5,7,3,7,7,5,5,2\}$.

To facilitate Markov melding, inference about the underlying parameters κ and λ is conducted rather than the probabilities q^C and q^T - which typically would seem to be the more natural option. Indeed for a fixed effects meta-analysis more generally it would likely be preferable to directly place priors on the probabilities (for example a Beta prior would be conjugate to the Binomial likelihood). Setting this fact aside, the posterior distribution given (s^C, s^T) is

$$\begin{aligned}
&p_2(\kappa, \lambda \,|\, s^C, s^T) \\
&\propto \prod_{i=1}^{n_2} \binom{N_2}{s_i^C} \left(\frac{e^{\lambda-\kappa/2}}{1+e^{\lambda-\kappa/2}}\right)^{s_i^C} \left(\frac{1}{1+e^{\lambda-\kappa/2}}\right)^{N_2 - s_i^C} \left(\frac{1}{1+e^{\lambda-\kappa/2}}\right)^2 \\
&\cdot \prod_{i=1}^{n_2} \binom{N_2}{s_i^T} \left(\frac{e^{\lambda+\kappa/2}}{1+e^{\lambda+\kappa/2}}\right)^{s_i^T} \left(\frac{1}{1+e^{\lambda+\kappa/2}}\right)^{N_2 - s_i^T} \left(\frac{1}{1+e^{\lambda+\kappa/2}}\right)^2 \\
&\cdot \frac{\upsilon_0}{\sqrt{2\pi}} \exp\left(-\frac{\upsilon_0}{2}(\kappa-\kappa_0)^2\right) \cdot \frac{\gamma_0}{\sqrt{2\pi}} \exp\left(-\frac{\gamma_0}{2}(\lambda-\lambda_0)^2\right) \\
&\propto \left(\frac{e^{\lambda-\kappa/2}}{1+e^{\lambda-\kappa/2}}\right)^{\sum_{i=1}^{n_2} s_i^C} \left(\frac{1}{1+e^{\lambda-\kappa/2}}\right)^{2+n_2N_2-\sum_{i=1}^{n_2} s_i^C} \\
&\cdot \left(\frac{e^{\lambda+\kappa/2}}{1+e^{\lambda+\kappa/2}}\right)^{\sum_{i=1}^{n_2} s_i^T} \left(\frac{1}{1+e^{\lambda+\kappa/2}}\right)^{2+n_2N_2-\sum_{i=1}^{n_2} s_i^T} \\
&\cdot \exp\left(-\frac{\upsilon_0}{2}(\kappa-\kappa_0)^2\right) \cdot \exp\left(-\frac{\gamma_0}{2}(\lambda-\lambda_0)^2\right),
\end{aligned} \tag{3.36}$$

where again transformation of variables is used (both of which with Jacobians $(1+e^{\lambda-\kappa/2})^{-2}$). The log posterior is

$$\begin{aligned}
&\log p(\kappa, \lambda \,|\, s^C, s^T) \\
&\propto \sum_{i=1}^{n_2} s_i^C \log\left(\frac{e^{\lambda-\kappa/2}}{1+e^{\lambda-\kappa/2}}\right) + (2 + n_2N_2 - \sum_{i=1}^{n_2} s_i^C) \log\left(\frac{1}{1+e^{\lambda-\kappa/2}}\right) \\
&+ \sum_{i=1}^{n_2} s_i^T \log\left(\frac{e^{\lambda+\kappa/2}}{1+e^{\lambda+\kappa/2}}\right) + (2 + n_2N_2 - \sum_{i=1}^{n_2} s_i^T) \log\left(\frac{1}{1+e^{\lambda+\kappa/2}}\right) \\
&- \frac{\upsilon_0}{2}(\kappa-\kappa_0)^2 - \frac{\gamma_0}{2}(\lambda-\lambda_0)^2 := l_2.
\end{aligned} \tag{3.37}$$

To target Equation (3.36) we again use the RWM-within-Gibbs algorithm with acceptance rates tuned to 0.44. The initialisation values are chosen based on the a-priori means $(\kappa^{(0)}, \lambda^{(0)}) =$

$(0.1, 0)$. The number of samples used (1000000) is based on similar considerations to that of Submodel 1 (as we are only estimating a $p = 2$ dimensional mean the required effective samples is much lower). The resulting traceplots, histograms and posterior summaries are shown in Figure 3.5 and Table 3.4 respectively. Once again the MCMC output looks free from any significant problems. Both κ and λ are accurately recovered: well within posterior mean plus or minus one posterior standard deviation.

Table 3.3: Submodel 2 scaling and acceptance rate for κ and λ

	κ	λ
σ	0.550	0.300
Acceptance	0.469	0.445

Table 3.4: Submodel 2 posterior mean (PM), true value and posterior standard deviation (PSD) for κ and λ

	κ	λ
PM	-0.221	-0.887
Truth	-0.250	-1.000
PSD	0.250	0.126

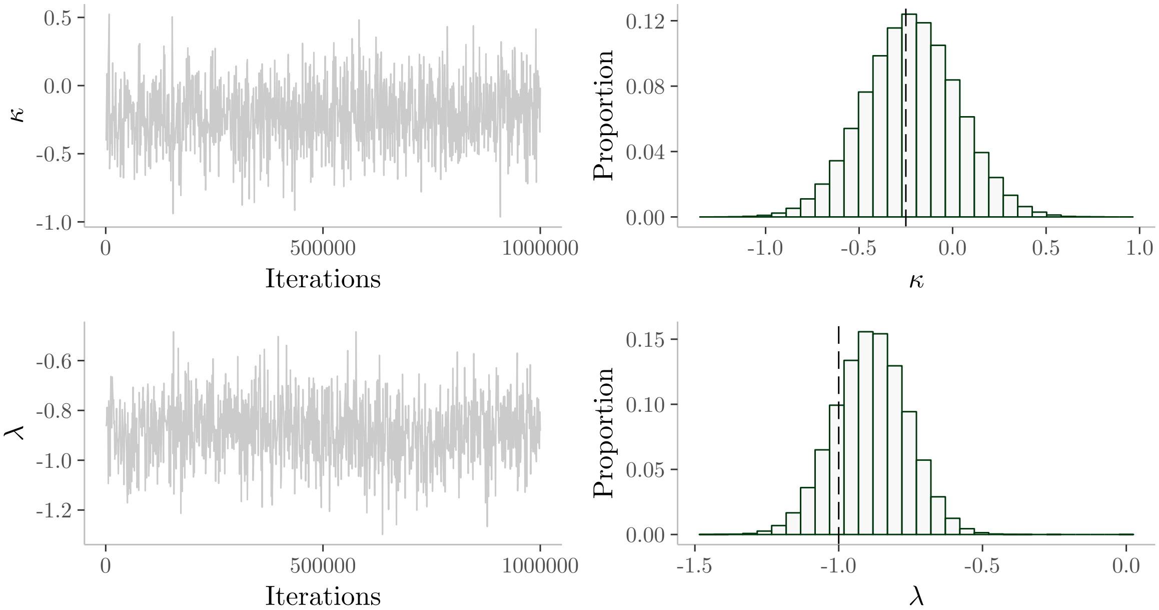


Figure 3.5: Submodel 2 posterior traceplots and histograms for the parameters κ and λ

Melded model By Equation (2.35), the Markov melded posterior is

$$\begin{aligned} & p_{\text{meld}}(\kappa, \upsilon, p^C, p^T, \lambda \,|\, r^C, r^T, s^C, s^T) \\ & \propto p_{\text{pool}}(\kappa) \frac{p_1(\kappa, \upsilon, p^C, p^T \,|\, r^C, r^T)}{p_1(\kappa)} \frac{p_2(\kappa, \upsilon \,|\, s^C, s^T)}{p_2(\kappa)}. \end{aligned} \tag{3.38}$$

For computational simplicity, PoE pooling $p_{\text{pool}}(\kappa) \propto p_1(\kappa) p_2(\kappa)$ is chosen such that Equation (3.38) simplifies to

$$p_{\text{meld}}(\kappa, \upsilon, p^C, p^T, \lambda \,|\, r^C, r^T, s^C, s^T) \propto p_1(\kappa, \upsilon, p^C, p^T \,|\, r^C, r^T) p_2(\kappa, \upsilon \,|\, s^C, s^T). \tag{3.39}$$

Taking the logarithm, we have that the log posterior is simply a sum of the submodel log posteriors, l_1 and l_2

$$\begin{aligned} & \log p_{\text{meld}}(\kappa, \upsilon, p^C, p^T, \lambda \,|\, r^C, r^T, s^C, s^T) \\ & \propto \log p_1(\kappa, \upsilon, p^C, p^T \,|\, r^C, r^T) + \log p_2(\kappa, \upsilon \,|\, s^C, s^T) \\ & \propto l_1 + l_2. \end{aligned} \tag{3.40}$$

To target Equation (3.39) we use 5000000 iterations of RWM-within-Gibbs with each component tuned to 0.44 acceptance rate. Each parameter is initialised as before, either based on a-priori means or 0.5 for probabilities.

Table 3.5: Melded model scaling and acceptance rate for each component

	κ	υ	p_1^C	p_2^C	p_3^C	p_4^C	p_5^C	p_1^T	p_2^T	p_3^T	p_4^T	p_5^T	λ
σ	0.600	3.500	0.300	0.250	0.300	0.250	0.200	0.150	0.15	0.250	0.300	0.200	0.400
Acceptance	0.433	0.452	0.383	0.407	0.397	0.381	0.411	0.438	0.44	0.406	0.396	0.449	0.358

Table 3.6: Melded model posterior mean (PM), true value and posterior standard deviation (PSD) for each component

	κ	υ	p_1^C	p_2^C	p_3^C	p_4^C	p_5^C	p_1^T	p_2^T	p_3^T	p_4^T	p_5^T	λ
PM	-0.224	2.998	0.359	0.264	0.500	0.218	0.172	0.124	0.124	0.262	0.546	0.215	-0.888
Truth	-0.250	4.000	0.214	0.360	0.464	0.296	0.236	0.134	0.324	0.308	0.421	0.221	-1.000
PSD	0.244	1.738	0.101	0.093	0.107	0.087	0.079	0.068	0.068	0.093	0.106	0.087	0.126

The posterior means for κ are -0.314 (Submodel 1), -0.221 (Submodel 2) and -0.224 (melded model) respectively. This result is not due to Monte carlo error: as well as the ESS considerations, 100 replications of each simulation gives maximum and minimum posterior means for κ to be [-0.323, -0.3] (Submodel 1), [-0.223, -0.218] (Submodel 2) and [-0.227, -0.222] (melded model). Notably, the melded model estimate is far closer to that of Submodel 2. This is relatively unsurprising since Submodel 2 is substantially more informative than Submodel 1. Firstly, the sample size is larger: $2N_1n_1 = 100$ patients in Y_1 compared with $2N_2n_2 = 300$ patients in Y_2. Additionally, due to the random-effects (remembering that the data is simulated) there is more variation in the Submodel 1 data and thus the signal is weaker than for the fixed-effects Submodel 2.

The fitted submodel specific parameters in the melded model (a little disappointingly) appear to be essentially the same as when fitting each of the submodels alone. Just as with Submodel 1, the same 6/10 intervals generated by posterior mean plus or minus one posterior standard deviation contain the truth. □

Having introduced and demonstrated a successful method for Markov melding inference, in the final part of this chapter we discuss one further algorithm extending the MWG approach.

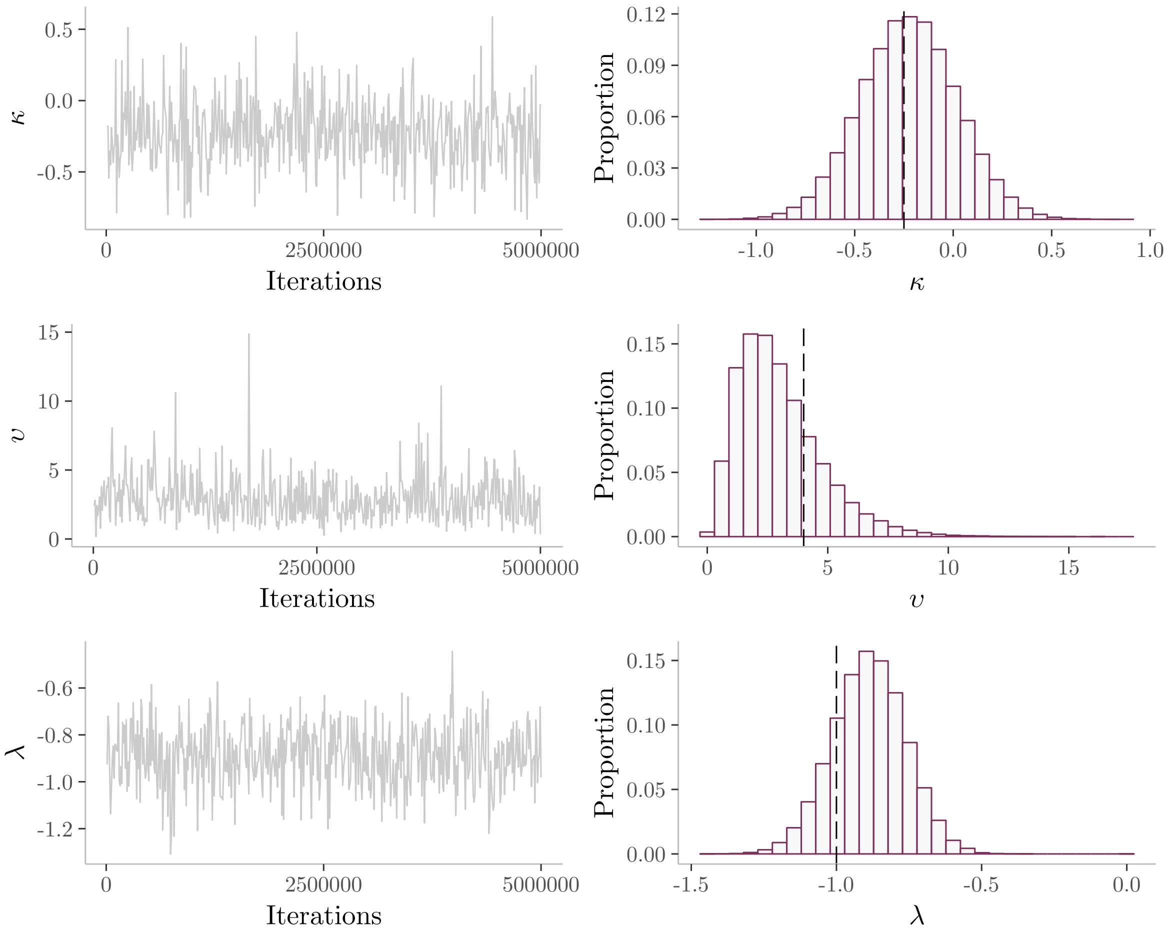


Figure 3.6: Melded model posterior traceplots and histograms for the parameters κ, υ and λ. As with Submodel 1, plots for the other parameters are presented in Appendix A

3.3 Sequential Monte Carlo

In some situations, where the submodels are complex, it may be challenging (or computationally inefficient) to directly sample from the full Markov melded posterior using MWG. Goudie et al. (2019) therefore propose a sequential approach, in which information from the submodels is added step-by-step. At a high-level, it is an example of Sequential Monte Carlo (SMC) which we outline the connection to here.

Consider a probability density function $\pi(x)$ as before defined on the state space $\mathcal{X}$. Following Lindsten et al. (2017), if $\mathcal{X}$ is a product space it may be decomposed into a Cartesian product of subspaces according to

$$\mathcal{X} := \mathcal{X}_T = \widetilde{\mathcal{X}}_1 \times \widetilde{\mathcal{X}}_2 \times \cdots \times \widetilde{\mathcal{X}}_T. \tag{3.41}$$

Elements $x \in \mathcal{X}$ in the product space may be written as $x := x_T = (\widetilde{x}_1, \ldots, \widetilde{x}_T)$. Both the elements of the subspaces and the subspaces themselves are denoted with tildes, so that $\widetilde{x}_t \in \widetilde{\mathcal{X}}_t$, for $1 \leq t \leq T$. A sequence of auxiliary probability distributions $\pi_1, \ldots \pi_T$ can be defined on the spaces $\mathcal{X}_t$ of lower dimension $1 \leq t \leq T$ than $\mathcal{X}_T$. $\pi_t(x_t) := \pi_t(\widetilde{x}_1, \cdots, \widetilde{x}_t)$ for $t = 1, \ldots, T$. The final auxiliary distribution is chosen to coincide with the target distribution, such that

$$\pi = \pi_T. \tag{3.42}$$

Collections of auxiliary distributions with the two properties, (3.41) and (3.42), are called a sequential decomposition of the target distribution π.

Sequential Monte Carlo (SMC) methods are a general class of Monte Carlo algorithms designed for situations where sequential decompositions are applicable. The distributions $\pi_1, \pi_2, \cdots, \pi_T = \pi$ are approximated sequentially, allowing information from previous iterations to inform the sampling strategy as it proceeds.

3.3.1 Multi-stage Metropolis-within-Gibbs

Define $(\phi, \psi_1, \ldots, \psi_M) = \theta_M = \theta \in \Theta$, then the full parameter space Θ may be decomposed such that $\Theta = \Phi \times \Psi_1 \times \cdots \times \Psi_M$. Define the sequence of subspaces $\Theta_\ell = \Phi \times \Psi_1 \times \cdots \times \Psi_m$

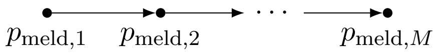


Figure 3.7: Computational flow of the multi-stage MWG sampler for Markov melding. Stages labeled with target auxiliary distribution. Adapted to Markov melding context from Lindsten et al. (2017)

with respective elements $\theta_\ell = (\phi, \psi_1, \dots, \psi_\ell)$ for $\ell = 1, \dots, M$. The subspaces are of increasing dimension such that $\Theta_1 \subseteq \cdots \subseteq \Theta_\ell \subseteq \cdots \subseteq \Theta_M$ with the final subspace $\Theta_M = \Theta$.

It remains to construct a sequence of auxiliary probability distributions $p_{\text{meld},l}$ defined on the subspaces Θ_ℓ such that the M^{th} distribution coincides with the posterior

$$p_{\text{meld},M}(\theta_M) = p_{\text{meld},M}(\phi, \psi_1, \dots, \psi_M) = p_{\text{meld}}(\phi, \psi_1, \dots, \psi_M \mid y_1, \dots, y_M) \tag{3.43}$$

The posterior (2.35) includes a product of the submodel posteriors $p_m(\phi, \psi_m \mid y_m)$ and conditional on ϕ, the m^{th} submodel is the only source of information about ψ_m. It therefore seems reasonable that when $\Psi_m \subseteq \Theta_\ell$ this information $p_m(\phi, \psi_m \mid y_m)$ about ψ_m is included in the auxiliary probability distribution.

On the other hand, informally speaking, information about the link parameter is ϕ is dispersed across the M submodels. For this reason there is flexibility about how the auxiliary distributions should be defined to take this into account. This flexibility can be encapsulated by the possible factorisations of the pooled prior

$$p_{\text{pool}}(\phi) = \prod_{m=1}^{M} p_{\text{pool},m}(\phi). \tag{3.44}$$

Two example factorisations are $p_{\text{pool},m}(\phi) = p_{\text{pool}}(\phi)^{1/M}$ and $p_{\text{pool},m}(\phi) = p_m(\phi)$.

Therefore, the ℓ^{th} stage posterior over the parameters $\theta_\ell \in \Theta_\ell$ is defined to be

$$p_{\text{meld},l}(\theta_\ell \mid y_1, \dots, y_\ell) \propto \prod_{m=1}^{\ell} \left(\frac{p_m(\phi, \psi_m, y_m)}{p_m(\phi)} p_{\text{pool},m}(\phi) \right) \tag{3.45}$$

where indeed Equation (3.43) holds.

To sample from this sequence of auxiliary distributions, Goudie et al. (2019) generalise a previous two-stage computational approach of Lunn et al. (2013). With the introduction of each additional parameter, the relevant Markov chains may be initialised as before at the corresponding elements of $(\phi^{(0)}, \psi_1^{(0)}, \dots, \psi_M^{(0)})$.

Stage 1. For the first stage, the auxiliary target distribution is the 1st stage posterior $p_{\text{meld},1}(\phi, \psi_1 \mid y_1)$. Samples $(\phi^{(h,1)}, \psi_1^{(h,1)})$ for $h = 1, \dots, H_1$ from this distribution can typically be drawn using standard MCMC methods such as MWG. If $p_{\text{pool},1}(\phi) = p_1(\phi)$ then the first stage posterior corresponds to the standard posterior $p_m(\phi, \psi_m \mid y_m)$. This may present computational advantages as often a given submodel may have been fit and samples will be already available. This factor may motivate setting the first submodel to be the most computational intensive submodel for which samples are available.

Stage ℓ. Following stage $\ell - 1$, samples $(\theta^{(h,\ell-1)})$ for $h = 1, \dots, H_{\ell-1}$ from the $(\ell-1)^{\text{th}}$ stage posterior $p_{\text{meld},\ell-1}(\theta_{\ell-1} \mid y_1, \dots, y_{\ell-1})$ are available. In stage ℓ a MWG sampler targeting $p_{\text{meld},l}(\theta_{\ell-1}, \psi_\ell \mid y_1, \dots, y_\ell)$ on the space $\Theta_\ell = \Theta_{\ell-1} \times \Psi_\ell$ is constructed.

Keeping $\theta_{\ell-1}$ fixed, the parameter ψ_m is updated as usual using a Metropolis-Hastings step. This chain can be initialised at $\psi_m^{(0)}$. The target-to-proposal density ratio is equivalent, with respect to the calculation of r as in Equation (3.14), to

$$R(\theta_{\ell-1}, \psi_\ell^\star, \theta_{\ell-1}, \psi_\ell) = p_\ell(\phi, \psi_\ell^\star, y_\ell) \cdot \frac{1}{q(\psi_\ell^\star \mid \psi_\ell)} \tag{3.46}$$

Now, to update the parameters $\theta_{\ell-1}$ the empirical distribution generated by the draws $(\theta^{(h,\ell-1)})$ for $h = 1, \dots, H_{\ell-1}$ is used as proposal. This is equivalent to drawing an index $d \sim \mathcal{U}(\{1, \dots, H_{\ell-1}\})$ and setting $\theta_{\ell-1}^\star = \theta_{\ell-1}^{(d,\ell-1)}$, a process commonly known as resampling. The resulting target-to-proposal density ratio simplifies as a result

$$\begin{aligned}
&R(\theta_{\ell-1}^\star, \psi_\ell, \theta_{\ell-1}, \psi_\ell) \\
&= \frac{p_\ell(\phi^\star, \psi_\ell, y_\ell)}{p_\ell(\phi^\star)} p_{\text{pool},\ell}(\phi^\star) \prod_{m=1}^{\ell-1} \left(\frac{p_m(\phi^\star, \psi_m^\star, y_m)}{p_m(\phi^\star)} p_{\text{pool},m}(\phi^\star) \right) \cdot \frac{1}{q(\theta_{\ell-1}^\star \mid \theta_{\ell-1})} \\
&= \frac{p_\ell(\phi^\star, \psi_\ell, y_\ell)}{p_\ell(\phi^\star)} p_{\text{pool},\ell}(\phi^\star) \frac{p_{\text{meld},\ell-1}(\phi^\star, \theta_{\ell-1}^\star \mid y_1, \dots, y_{\ell-1})}{p_{\text{meld},\ell-1}(\phi^\star, \theta_{\ell-1}^\star \mid y_1, \dots, y_{\ell-1})} \\
&= \frac{p_\ell(\phi^\star, \psi_\ell, y_\ell)}{p_\ell(\phi^\star)} p_{\text{pool},\ell}(\phi^\star), \qquad (3.47)
\end{aligned}$$

facilitating fast computation.

In the Markov melding framework, it is assumed that the set of studies to be synthesised and the statistical models for each study $m = 1, \dots, M$ are prespecified. In reality this is often not the case: practitioners may first fit a single submodel and then, based on the results, decide whether (or not) to include further evidence. It could be argued that the above multi-stage MWG algorithm would computationally facilitate this sequential workflow. However, for good reason, proponents of meta-analysis have suggested a systematic approach to study selection to avoid bias.

Chapter 4

Conclusion and further work

Markov melding and related approaches are an ongoing area of methodological research with a broad range of possible applications across many domains. It is possible that integration into popular (Gibbs-sampling-based) statistical software such as `BUGS` (D. J. Lunn et al. 2000) or `JAGS` (Plummer and others 2003) could enhance the rate of adoption by practitioners.

In isolation of any computational considerations, Chapter 2 and Example 2.2 demonstrate the validity of the approach. However, in almost all applied settings the computational techniques discussed in Chapter 3 will be required. Example 3.2 is a more realistic demonstration than Example 2.2, involving the fitting of two relatively complex submodels. That being said, the examples considered in this dissertation are still idealised in a number of ways.

The data we have used was simulated from the model, such that for some parameter setting the true data generating mechanism can exactly be recovered. This situation is known as the $\mathcal{M}$-closed world (Bernardo and Smith 2009). Increasingly, statisticians are acknowledging that real world inference typically takes place in the $\mathcal{M}$-open world where their models are (at least to some extent) misspecified. Evidence synthesis procedures such as Markov melding have particular reason to be worried by model misspecification. As M submodels are involved and information flows between the submodels the risks of overall model misspecification are amplified (Jacob et al. 2017).

Preliminary conflict analysis (Presanis et al. 2013) between the submodels may provide some security against misspecification, with Goudie et al. (2019) advising that Markov melding should not be used when the exists “strong conflict between evidence components”. Indeed, Goudie et al. (2019) also suggest one application of Markov melding may be in systematic conflict assessment. One aspect of Markov melding which we have not yet discussed is that of model splitting, the inverse operation to model joining (Goudie et al. 2019). Supposing parts of a suitable larger model are found to be in conflict, then model splitting could be used to remove dependence on the offending submodels. Alternatively, model splitting may be computationally preferable depending on the cost of fitting the large joint model.

Prior specification has been a historic focus of model robustness research in Bayesian statistics (Watson and Holmes 2016). For Markov melding, taking the submodel priors as given, sensitivity (of computational efficiency as well as the outcomes) to the pooling operation could be investigated further. Although PoE pooling is attractive in order to simplify some operations, the strength of its aggregated beliefs may be cause for concern.

As a final point, the examples we considered featured only $M = 2$ submodels. There is significant research and practical interest in techniques which "scale-up" Bayesian statistics by facilitating the handling of larger quantities of data. Markov melding may be prohibitively computationally expensive for joining large numbers of submodels. In the multi-stage Metropolis-within-Gibbs algorithm each stage ℓ can only be completed after the previous stage $\ell - 1$ has been. An alternative, proposed by Lindsten et al. (2017) in a broader SMC context, is to instead utilise divide-and-conquer by operating on a tree $\mathcal{T}$ rather than the chain $\{1, \dots, M\}$. Particularly for joining many (at least $M \geq 4$) submodels, this may result in significant computational gains as the resulting algorithm would be more easily parallelisable. Figure 4.1 shows a proposed computational flow for this problem, the task then being to to define the auxiliary distributions π and their domain in order to arrive at $\pi_{1:4} = p_{\text{meld}}$ at the root of the tree. Information about the link parameter from each submodel must be incorporated, so ϕ will be be present in each independent branch. It therefore seems natural that π_1 be defined on $\Phi \times \Psi_1$, π_2 on $\Phi \times \Psi_2$ and $\pi_{1,2}$ on $\Phi \times \Psi_1 \times \Psi_2$. Further research could investigate whether computationally efficient MWG algorithms could be found in this setting.

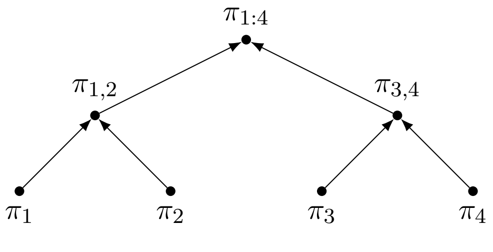


Figure 4.1: Computational flow for divide-and-conquer Markov melding with $M = 4$ submodels.

That said, it is unclear as to how often $M \geq 4$ submodels will be available. However, it is foreseeable that as data infrastructure improves more generally, it could become increasingly commonplace.

Appendix A

Appendix

A.1 Submodel 1

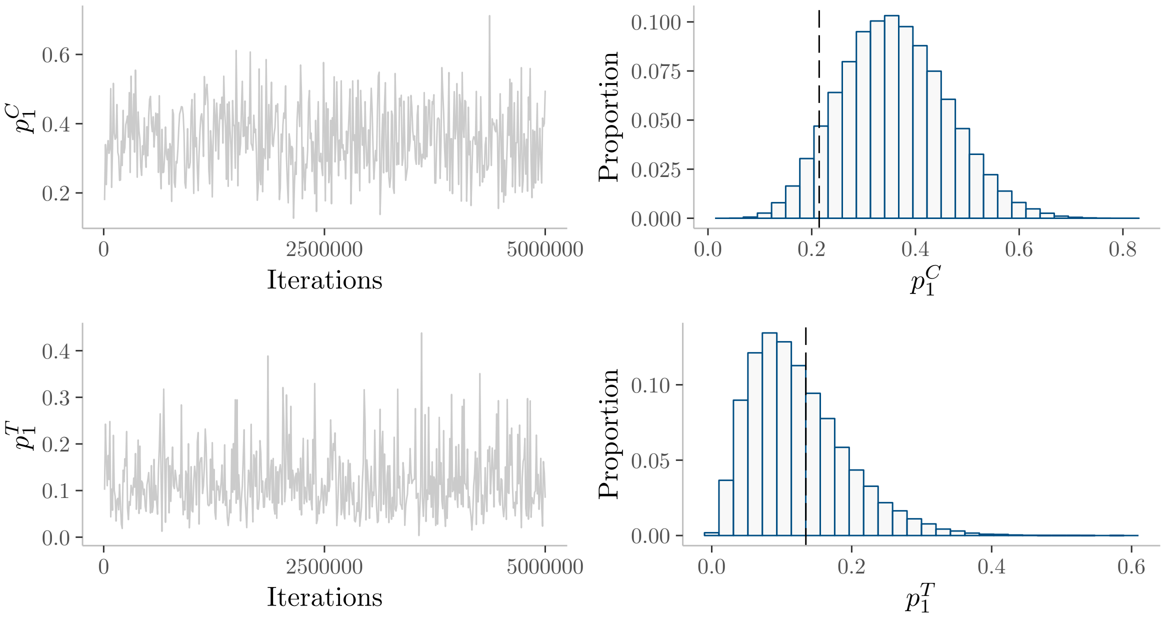


Figure A.1: Submodel 1 posterior traceplots and histograms for p_1^C and p_1^T

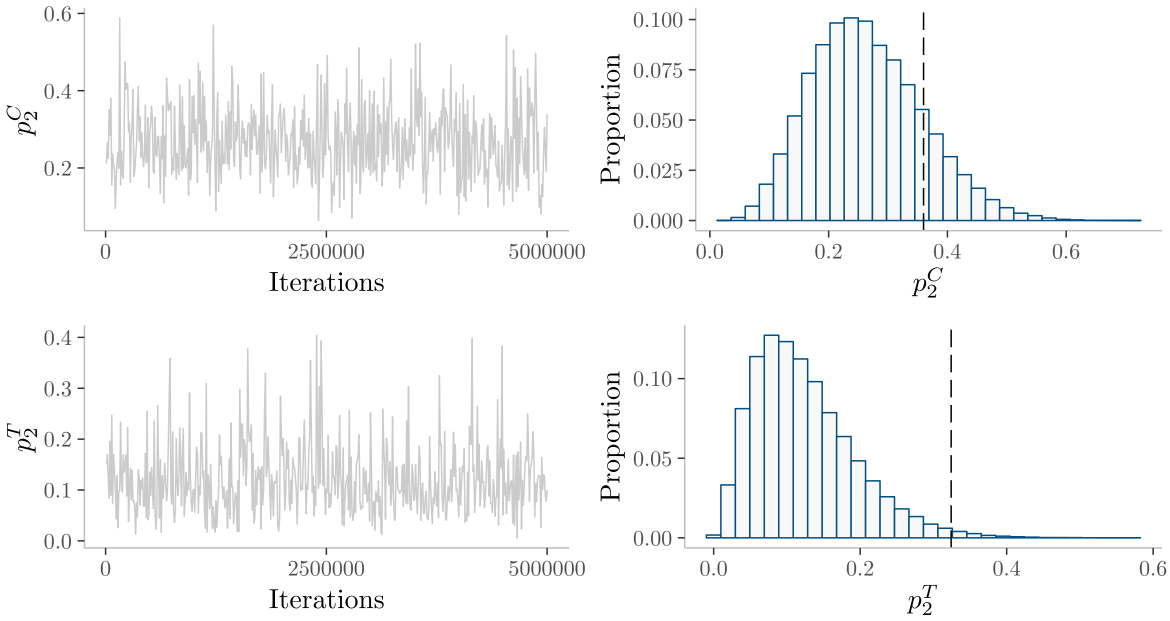


Figure A.2: Submodel 1 posterior traceplots and histograms for p_2^C and p_2^T

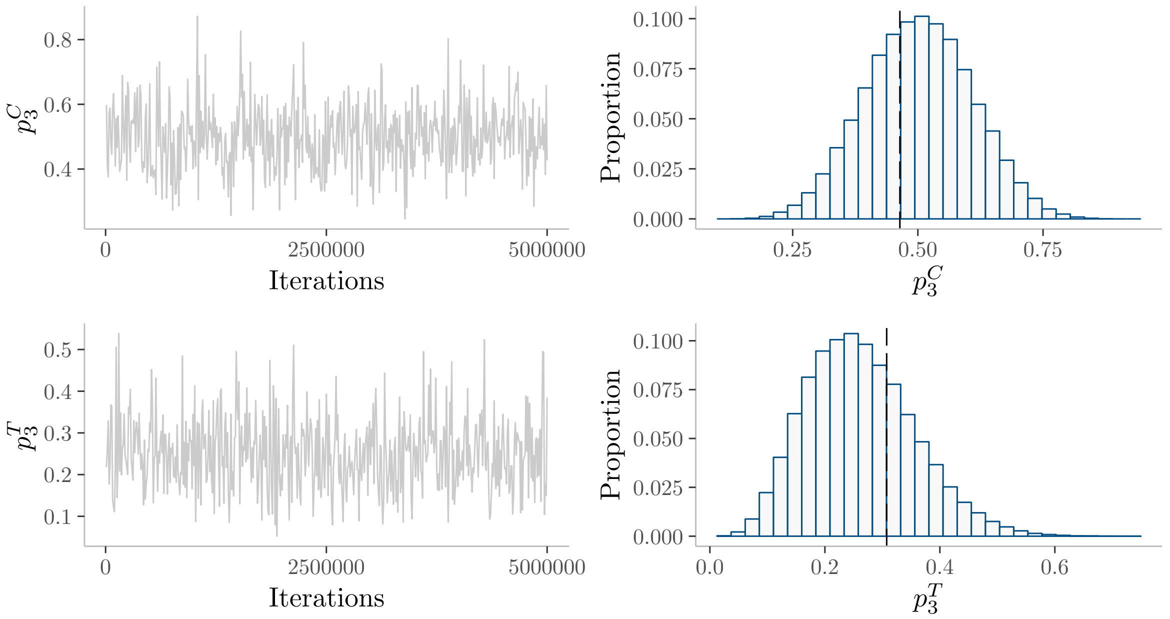


Figure A.3: Submodel 1 posterior traceplots and histograms for p_3^C and p_3^T

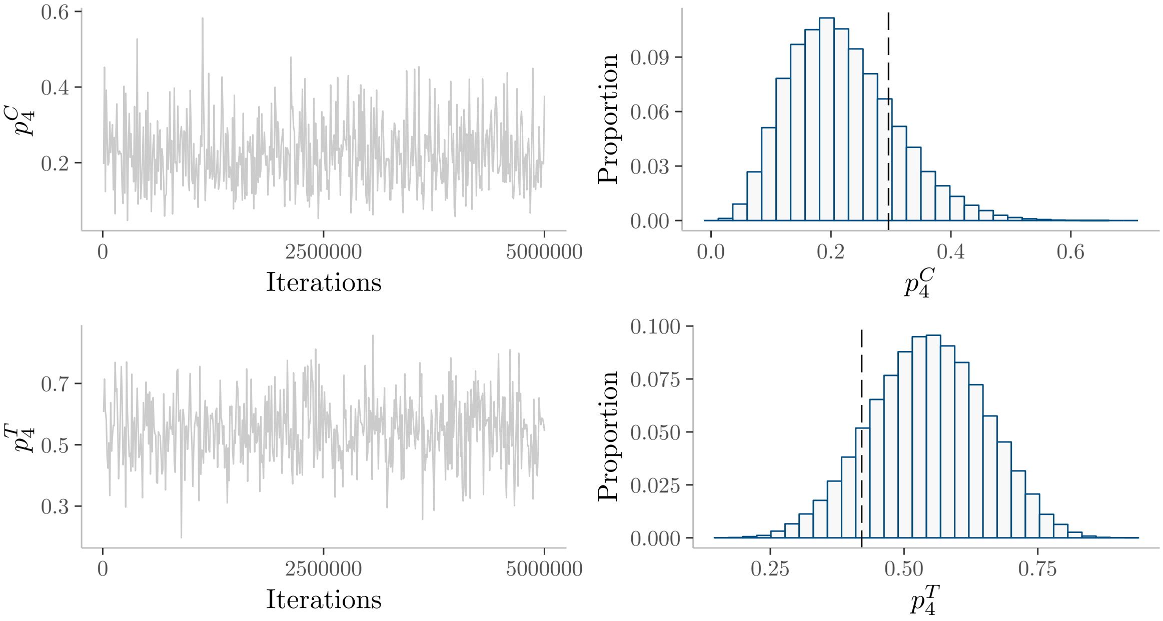


Figure A.4: Submodel 1 posterior traceplots and histograms for p_4^C and p_4^T

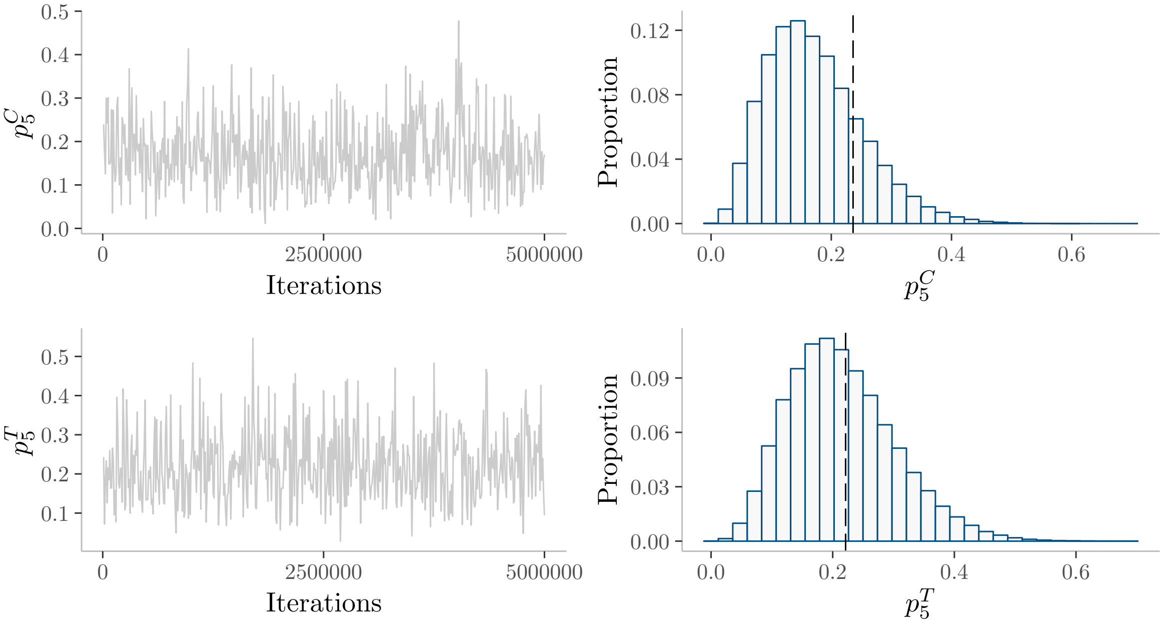


Figure A.5: Submodel 1 posterior traceplots and histograms for p_5^C and p_5^T

A.2 Melded model

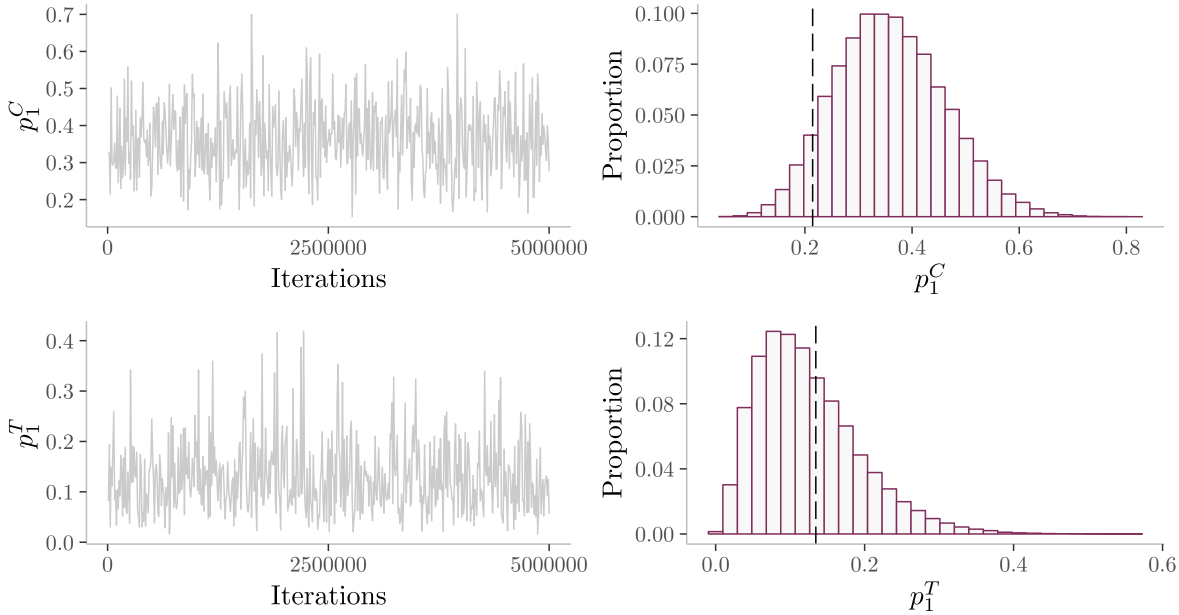


Figure A.6: Melded model posterior traceplots and histograms for p_1^C and p_1^T

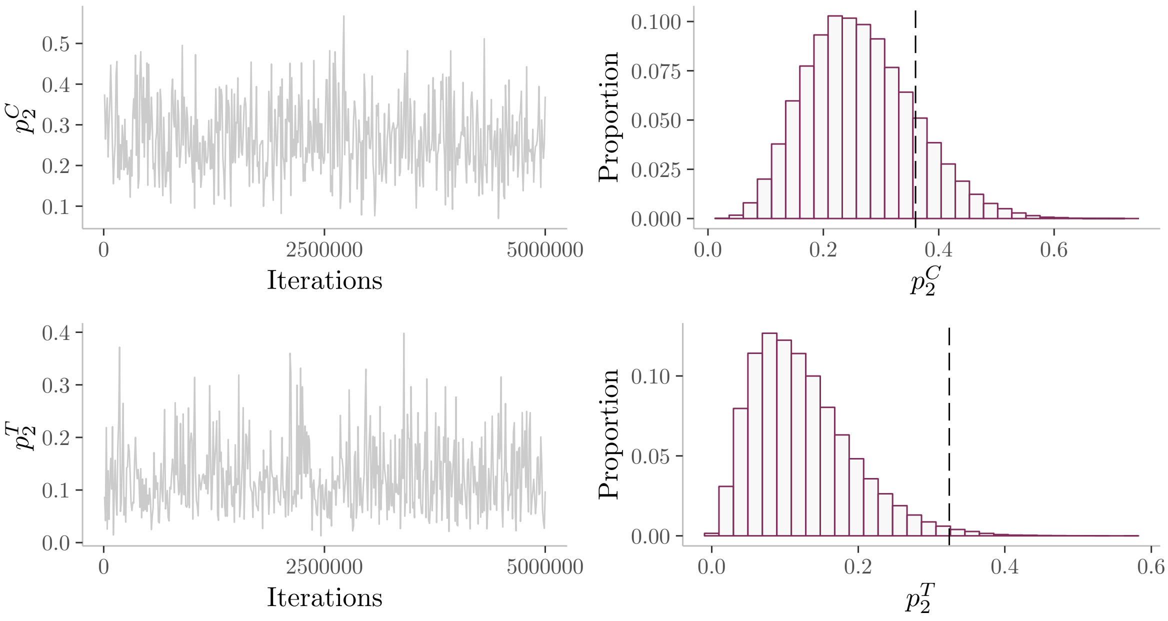


Figure A.7: Melded model posterior traceplots and histograms for p_2^C and p_2^T

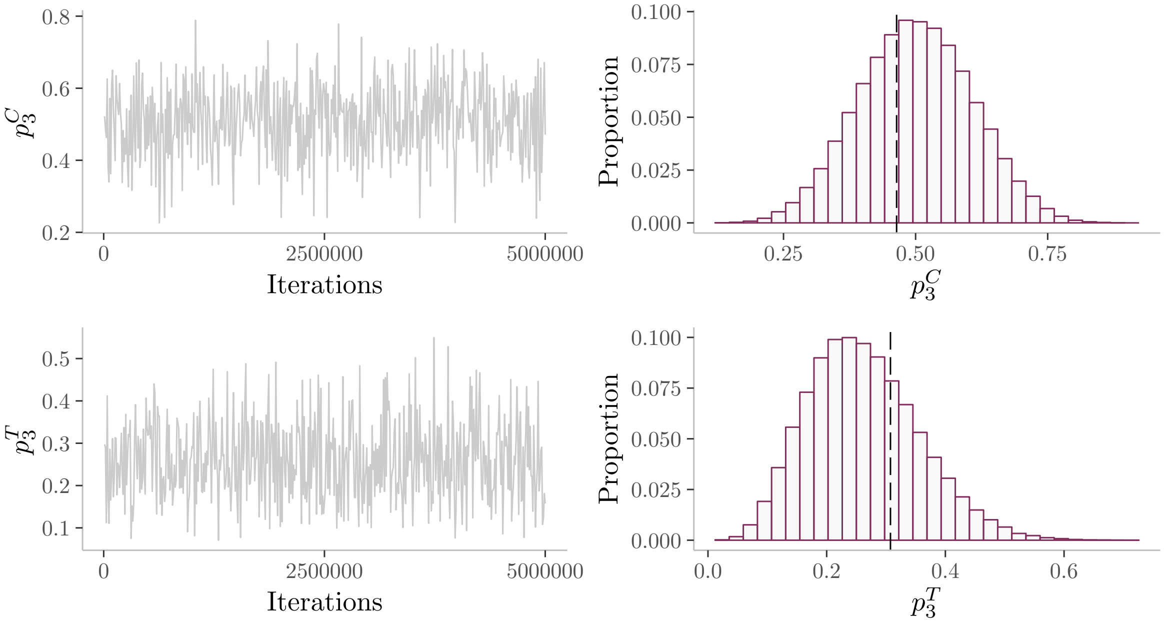


Figure A.8: Melded model posterior traceplots and histograms for p_3^C and p_3^T

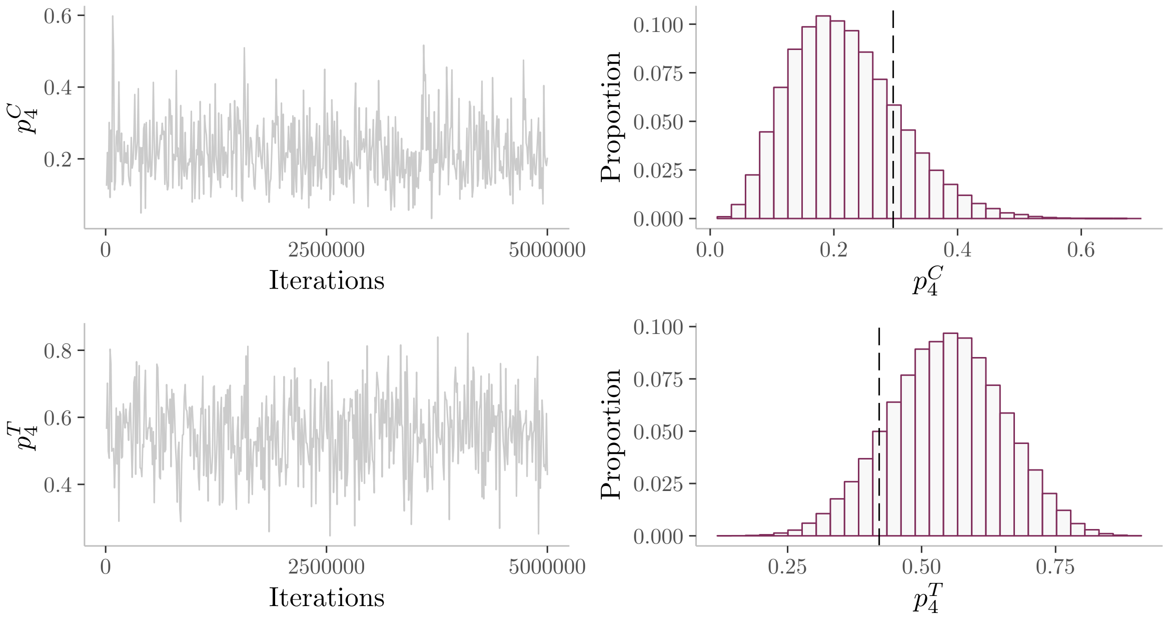


Figure A.9: Melded model posterior traceplots and histograms for p_4^C and p_4^T

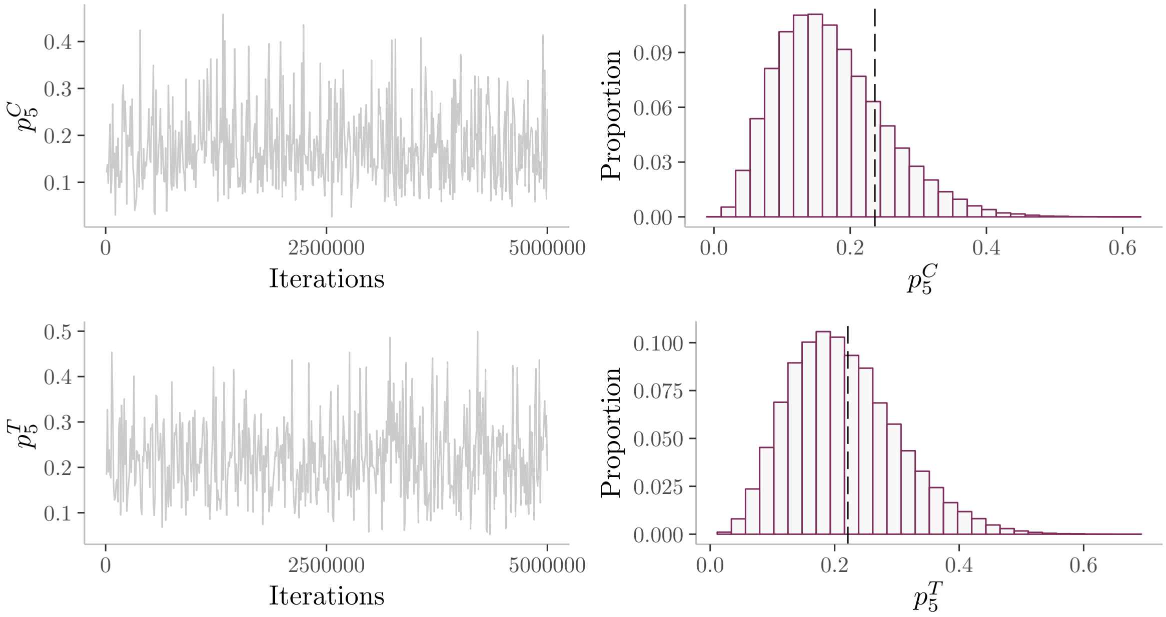


Figure A.10: Melded model posterior traceplots and histograms for p_5^C and p_5^T

Bibliography

Ades, AE, and S Cliffe. 2002. "Markov Chain Monte Carlo Estimation of a Multiparameter Decision Model: Consistency of Evidence and the Accurate Assessment of Uncertainty." *Medical Decision Making* 22 (4). Sage Publications Sage CA: Thousand Oaks, CA: 359–71.

Ashby, Deborah, and Adrian FM Smith. 2000. "Evidence-Based Medicine as Bayesian Decision-Making." *Statistics in Medicine* 19 (23). Wiley Online Library: 3291–3305.

Bernardo, José M, and Adrian FM Smith. 2009. *Bayesian Theory*. Vol. 405. John Wiley & Sons.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Clemen, Robert T, and Robert L Winkler. 1999. "Combining Probability Distributions from Experts in Risk Analysis." *Risk Analysis* 19 (2). Springer: 187–203.

Dawid, A Philip, and Steffen L Lauritzen. 1993. "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models." *The Annals of Statistics* 21 (3). Institute of Mathematical Statistics: 1272–1317.

Flegal, James M., John Hughes, Dootika Vats, and Ning Dai. 2017. *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, Denver, CO, Coventry, UK; Minneapolis, MN.

Gelman, Andrew, and Donald B Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4). Institute of Mathematical Statistics: 457–72.

Geman, Stuart, and Donald Geman. 1987. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." In *Readings in Computer Vision*, 564–84. Elsevier.

Gilks, Walter R, Sylvia Richardson, and David Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. Chapman; Hall/CRC.

Goudie, Robert JB, Anne M Presanis, David Lunn, Daniela De Angelis, and Lorenz Wernisch.

2019. “Joining and Splitting Models with Markov Melding.” *Bayesian Analysis* 14 (1). Europe PMC Funders: 81.

Hastings, W Keith. 1970. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications.” Oxford University Press.

Hickman, Matthew, Daniela De Angelis, Hayley Jones, Ross Harris, Nicky Welton, and AE Ades. 2013. “Multiple Parameter Evidence Synthesis-a Potential Solution for When Information on Drug Use and Harm Is in Conflict.” *Addiction* 108 (9). Wiley Online Library: 1529–31.

Hinton, Geoffrey E. 2002. “Training Products of Experts by Minimizing Contrastive Divergence.” *Neural Computation* 14 (8). MIT Press: 1771–1800.

Jacob, Pierre E, Lawrence M Murray, Chris C Holmes, and Christian P Robert. 2017. “Better Together? Statistical Learning in Models Made of Modules.” *arXiv Preprint arXiv:1708.08719*.

Lindsten, Fredrik, Adam M Johansen, Christian A Naesseth, Bonnie Kirkpatrick, Thomas B Schön, JAD Aston, and Alexandre Bouchard-Côté. 2017. “Divide-and-Conquer with Sequential Monte Carlo.” *Journal of Computational and Graphical Statistics* 26 (2). Taylor & Francis: 445–58.

Lunn, David J, Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. “WinBUGS-a Bayesian Modelling Framework: Concepts, Structure, and Extensibility.” *Statistics and Computing* 10 (4). Springer: 325–37.

Lunn, David, Jessica Barrett, Michael Sweeting, and Simon Thompson. 2013. “Fully Bayesian Hierarchical Modelling in Two Stages, with Application to Meta-Analysis.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62 (4). Wiley Online Library: 551–72.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. “Equation of State Calculations by Fast Computing Machines.” *The Journal of Chemical Physics* 21 (6). AIP: 1087–92.

Muller, Peter. 1991. “A Generic Approach to Posterior Integration and Gibbs Sampling.” *Technical Report*, 91–09.

Neal, Peter, and Gareth Roberts. 2006. “Optimal Scaling for Partially Updating Mcmc Algorithms.” *The Annals of Applied Probability* 16 (2). Institute of Mathematical Statistics: 475–515.

O’Hagan, Anthony, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities.* John Wiley & Sons.

Plummer, Martyn, and others. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 124:10. 125. Vienna, Austria.

Poole, David, and Adrian E Raftery. 2000. "Inference for Deterministic Simulation Models: The Bayesian Melding Approach." *Journal of the American Statistical Association* 95 (452). Taylor & Francis Group: 1244–55.

Presanis, Anne M, David Ohlssen, David J Spiegelhalter, Daniela De Angelis, and others. 2013. "Conflict Diagnostics in Directed Acyclic Graphs, with Applications in Bayesian Evidence Synthesis." *Statistical Science* 28 (3). Institute of Mathematical Statistics: 376–97.

Presanis, Anne M, Richard G Pebody, Paul J Birrell, Brian DM Tom, Helen K Green, Hayley Durnall, Douglas Fleming, Daniela De Angelis, and others. 2014. "Synthesising Evidence to Estimate Pandemic (2009) A/H1N1 Influenza Severity in 2009-2011." *The Annals of Applied Statistics* 8 (4). Institute of Mathematical Statistics: 2378–2403.

Public Health England. 2008. "HIV: Overall Prevalence." https://www.gov.uk/guidance/hiv-overall-prevalence.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raftery, Adrian E, Geof H Givens, and Judith E Zeh. 1995. "Inference from a Deterministic Population Dynamics Model for Bowhead Whales." *Journal of the American Statistical Association* 90 (430). Taylor & Francis Group: 402–16.

Robert, Christian, and George Casella. 2013. *Monte Carlo Statistical Methods.* Springer Science & Business Media.

Roberts, Gareth, Andrew Gelman, and Walter R Gilks. 1997. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *The Annals of Applied Probability* 7 (1). Institute of Mathematical Statistics: 110–20.

Royal Society, Academy of Medical Sciences. 2018. "Evidence Synthesis for Policy: A Statement of Principles." https://royalsociety.org/-/media/policy/projects/evidence-synthesis/evidence-synthesis-statement-principles.pdf.

Schweder, Tore, and Nils Lid Hjort. 1996. "Bayesian Synthesis or Likelihood Synthesis -

What Does Borel's Paradox Say?" *Preprint Series. Statistical Research Report.* Matematisk Institutt, Universitetet i Oslo.

Shah, Rajen D, and Jonas Peters. 2018. "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure." *arXiv Preprint arXiv:1804.07203.*

Smith, Teresa, David J Spiegelhalter, and Andrew Thomas. 1995. "Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study." *Statistics in Medicine* 14 (24). Wiley Online Library: 2685–99.

Vats, Dootika, James M Flegal, and Galin L Jones. 2015. "Multivariate Output Analysis for Markov Chain Monte Carlo." *arXiv Preprint arXiv:1512.07713.*

Watson, James, and Chris Holmes. 2016. "Approximate Models and Robust Decisions." *Statistical Science* 31 (4). Institute of Mathematical Statistics: 465–89.

Welton, Nicky, Alexander J Sutton, Nicola Cooper, Keith R Abrams, and AE Ades. 2012. *Evidence Synthesis for Decision Making in Healthcare.* Vol. 132. John Wiley & Sons.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wolpert, Robert L. 1995. "Inference from a Deterministic Population Dynamics Model for Bowhead Whales: Comment." *Journal of the American Statistical Association* 90 (430). JSTOR: 426–27.

Xie, Yihui. 2016. *bookdown: Authoring Books and Technical Documents with R Markdown.* Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.