Statistics and Computing (2022) 32:24
https://doi.org/10.1007/s11222-022-10086-2

A numerically stable algorithm for integrating Bayesian models using Markov melding

Andrew A. Manderson[1,2] · **Robert J. B. Goudie**[1]

Received: 9 December 2020 / Accepted: 29 January 2022 / Published online: 18 February 2022

Abstract
When statistical analyses consider multiple data sources, Markov melding provides a method for combining the source-specific Bayesian models. Markov melding joins together submodels that have a common quantity. One challenge is that the prior for this quantity can be implicit, and its prior density must be estimated. We show that error in this density estimate makes the two-stage Markov chain Monte Carlo sampler employed by Markov melding unstable and unreliable. We propose a robust two-stage algorithm that estimates the required prior marginal self-density ratios using weighted samples, dramatically improving accuracy in the tails of the distribution. The stabilised version of the algorithm is pragmatic and provides reliable inference. We demonstrate our approach using an evidence synthesis for inferring HIV prevalence, and an evidence synthesis of A/H1N1 influenza.

Keywords Biased sampling · Data integration · Evidence synthesis · Kernel density estimation · Multi-source inference · Self-density ratio · Weighted sampling

1 Introduction

Many modern applied statistical analyses consider several data sources, which differ in size and complexity. The wide variety of problems and information sources has produced numerous methods for multi-source inference (Lanckriet et al. 2004; Coley et al. 2017; Besbeas and Morgan 2019), as well as general methodologies including evidence synthesis methods (Sutton and Abrams 2001; Ades and Sutton 2006; Spiegelhalter et al. 2004), and the data fusion model (Kedem et al. 2017). These methods require an appropriate joint model for all data, which can be challenging to specify.

An alternative approach is to model smaller, simpler aspects of the data, such that designing these *submodels* is easier, then combine the submodels. The premise is that the combination of many smaller submodels will serve as a good approximation to a larger joint model, which may be methodologically or computationally infeasible. *Markov melding* (Goudie et al. 2019) is a methodology for coherently combining these submodels. Specifically, Markov melding joins together submodels that share a common quantity ϕ into a single joint model. Consider M submodels indexed by $m = 1, \ldots, M$ that share ϕ, have submodel specific parameters ψ_m and submodel specific data Y_m, denoting the m^{th} submodel $\mathrm{p}_m(\phi, \psi_m, Y_m)$. Markov melding forms a single joint *melded model* $\mathrm{p}_{\mathrm{meld}}(\phi, \psi_1, \ldots, \psi_M, Y_1, \ldots, Y_M)$, which enables information to flow through ϕ from one model to another. The melded model posterior thus incorporates uncertainty from all sources of data.

Multi-stage sampling methods are useful, pragmatic tools for estimating complex joint models – such as $\mathrm{p}_{\mathrm{meld}}$ – in a computationally feasible manner, and have been applied in settings including statistical genetics and phylogeny (Tom et al. 2010), meta-analysis (Lunn et al. 2013; Blomstedt et al. 2019), spatial statistics (Hooten et al. 2019), and joint models in survival analysis (Mauff et al. 2020). Whilst it is preferable to sample the joint posterior directly, this is often infeasible due to the complexity of the model, the size of the data, the limitations of probabilistic programming languages such as `JAGS` and `Stan`, or the complications of re-expressing complicated submodels in a common programming language (Johnson et al. 2020). Improving the stability of multi-stage estimation techniques is thus of interest to applied statisticians.

✉ Andrew A. Manderson
andrew.manderson@mrc-bsu.cam.ac.uk

[1] MRC Biostatistics Unit, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK

[2] The Alan Turing Institute, British Library, London, UK

Evidence synthesis models consider multiple sources of data (evidence), including randomised controlled trials or observational studies, to understand complex phenomena. Each source of data has an associated submodel and set of parameters it informs; combining all the submodels requires assuming deterministic or probabilistic relationships between the submodel-specific parameters. For example, De Angelis et al. (2014) collected many surveys of partially overlapping subpopulations, albeit at different frequencies, and combined these in an evidence synthesis model to estimate human immunodeficiency virus (HIV) prevalence in the United Kingdom. An introduction to evidence synthesis can be found in Chapter 8 of Spiegelhalter et al. (2004); other applications include estimating the prevalence of campylobacteriosis (Albert et al. 2011) and influenza (Presanis et al. 2014).

We can form evidence synthesis models by applying Markov melding to the various sources of data and their submodels. However, the common quantity ϕ may be a complex, non-invertible function of the parameters in one of the submodels. This is a challenge for Markov melding, as the method requires the prior marginal density of ϕ under each submodel $\mathrm{p}_m(\phi)$, which may not be analytically tractable. Instead, prior samples of ϕ are drawn, and the prior marginal density is estimated using a kernel density estimate (KDE) $\widehat{\mathrm{p}}_m(\phi)$ (Wand and Jones 1995). However, the use of a KDE in lieu of the analytic density function has poor implications for the numerical stability of the Markov Chain Monte Carlo (MCMC) method used to estimate the melded posterior, even in our low dimensional examples. Specifically, we illustrate that the multi-stage MCMC sampler of Goudie et al. (2019) is sensitive to error in $\widehat{\mathrm{p}}_m(\phi)$, particularly in low probability regions.

To address this sensitivity, we first note that Markov melding strictly only requires an estimate of the *self-density ratio* (Hiraoka et al. 2014), $\mathrm{r}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}) = \mathrm{p}_m(\phi_{\mathrm{nu}}) \,/\, \mathrm{p}_m(\phi_{\mathrm{de}})$, as we will show in Sect. 2. In Sect. 3 we develop methodology that reduces the error in the self-density ratio estimate $\widehat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ by using weighted-sample KDEs (Vardi 1985; Jones 1991), which are more accurate in low probability regions. Multiple weighted-sample estimates of $\widehat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ are combined via a weighted average to further improve performance. We call this methodology *weighted-sample self-density ratio estimation* (WSRE), and demonstrate the effectiveness of our methodology in two examples. The first is a toy example from Ades and Cliffe (2002). We show that output from the multi-stage estimation process that uses WSRE is closer to reference samples than the *naive* approach, which uses a single KDE for $\widehat{\mathrm{p}}_m(\phi)$. The second example is an involved evidence synthesis, previously considered in Goudie et al. (2019). Here we show that the multi-stage estimation process that employs WSRE produces plausible samples, whilst the naive approach produces nonsensical results. In these examples ϕ is a 1 or 2 dimensional quantity. We discuss the applicability of our method for higher dimensional ϕ in Sect. 6.

2 Markov melding

The Markov melding framework is able to join together any number of submodels which share a common component ϕ. As the examples in this paper only consider two submodels, we limit our exposition to the $M = 2$ model case; for the more general case see Goudie et al. (2019). Markov melding constructs a joint model using the conditional distributions for submodel-specific parameters ψ_m and data Y_m, denoted $\mathrm{p}_m(\psi_m, Y_m \mid \phi)$. These conditional distributions are then combined with a global prior for ϕ called the *pooled prior* $\mathrm{p}_{\mathrm{pool}}(\phi)$, which we discuss in Sect. 2.1. Mathematically, assuming that the supports of the relevant conditional, joint, and marginal distributions containing ϕ are appropriate, we define the *melded joint distribution* as

$$\begin{aligned}
&\mathrm{p}_{\mathrm{meld}}(\phi, \psi_1, \psi_2, Y_1, Y_2) \\
&\quad = \mathrm{p}_{\mathrm{pool}}(\phi)\mathrm{p}_1(\psi_1, Y_1 \mid \phi)\mathrm{p}_2(\psi_2, Y_2 \mid \phi) \qquad (1) \\
&\quad = \mathrm{p}_{\mathrm{pool}}(\phi)\frac{\mathrm{p}_1(\phi, \psi_1, Y_1)}{\mathrm{p}_1(\phi)}\frac{\mathrm{p}_2(\phi, \psi_2, Y_2)}{\mathrm{p}_2(\phi)}. \qquad (2)
\end{aligned}$$

The submodel-specific conditional densities $\mathrm{p}_m(\psi_m, Y_m) \mid \phi)$ may be analytically intractable. Hence, it is easier to work with the submodel-joint densities $\mathrm{p}_m(\phi, \psi_m, Y_m)$ and their prior marginal distributions $\mathrm{p}_m(\phi)$ as specified in Eq. (2), because the former can be factorised into the data generating process specified during submodel construction.

2.1 Forming the pooled prior

The pooled prior should represent previous knowledge of ϕ in the absence of other information. A general approach to constructing $\mathrm{p}_{\mathrm{pool}}(\phi)$ is to consider a weighted combination of the prior marginal distributions $\mathrm{p}_m(\phi)$, with submodel weights λ_m. Selection of the pooling method and specific values of the weights is a topic covered in much detail elsewhere (Clemen and Winkler 1999; O'Hagan et al. 2006); a full summary of this field is beyond the scope of this article. For the examples considered in this paper we form $\mathrm{p}_{\mathrm{pool}}(\phi)$ via logarithmic pooling: $\mathrm{p}_{\mathrm{pool}}(\phi) \propto \mathrm{p}_1(\phi)^{\lambda_1}\mathrm{p}_2(\phi)^{\lambda_2}$, with $\lambda_1 = \lambda_2 = \frac{1}{2}$. Logarithmic pooling also allows us to use the methodology we develop in Sect. 3 in the pooled prior.

2.2 Two-stage Markov chain Monte Carlo sampler

Directly estimating the melded model's posterior distribution

$$\mathrm{p}_{\mathrm{meld}}(\phi, \psi_1, \psi_2 \mid Y_1, Y_2) \propto \mathrm{p}_{\mathrm{pool}}(\phi) \frac{\mathrm{p}_1(\phi, \psi_1, Y_1)}{\mathrm{p}_1(\phi)} \frac{\mathrm{p}_2(\phi, \psi_2, Y_2)}{\mathrm{p}_2(\phi)}. \tag{3}$$

necessitates simultaneously evaluating both submodels. This can be impractical if the submodels are implemented in different probabilistic programming languages or have bespoke implementations. We use the two-stage Markov chain Monte Carlo (MCMC) sampler of Goudie et al. (2019) to sample from $\mathrm{p}_{\mathrm{meld}}(\phi, \psi_1, \psi_2 \mid Y_1, Y_2)$ without the need to evaluate Eq. (3) all at once. This involves a two-stage MCMC procedure, first sampling from a partial product of the terms in Eq. (3), then using these samples as a proposal distribution in the second stage. The result is a convenient cancellation of the common terms in the stage two acceptance probability, whilst still ensuring that the final samples come from the melded posterior distribution of Eq. (3).

In stage one of the sampler we may, for example, opt to target the first submodel p_1, but with an (improper) flat prior for ϕ

$$\mathrm{p}_{\mathrm{meld},1}(\phi, \psi_1 \mid Y_1) \propto \frac{\mathrm{p}_1(\phi, \psi_1, Y_1)}{\mathrm{p}_1(\phi)},$$

so we construct a standard Markov chain in which a proposed move from $(\phi, \psi_1) \rightarrow (\phi^*, \psi_1^*)$, with proposal density $\mathrm{q}(\phi^*, \psi_1^* \mid \phi, \psi_1)$, is accepted with probability

$$\alpha((\phi^*, \psi_1^*), (\phi, \psi_1)) = \frac{\mathrm{p}_1(\phi^*, \psi_1^*, Y_1)\mathrm{p}_1(\phi)\mathrm{q}(\phi, \psi_1 \mid \phi^*, \psi_1^*)}{\mathrm{p}_1(\phi, \psi_1, Y_1)\mathrm{p}_1(\phi^*)\mathrm{q}(\phi^*, \psi_1^* \mid \phi, \psi_1)}. \tag{4}$$

This Markov chain asymptotically emits samples from $\mathrm{p}_{\mathrm{meld},1}$.

In stage two we update ϕ and ψ_2 using Metropolis-within-Gibbs updates, targeting the full melded posterior distribution of Eq. (3). Updating ϕ uses the stage one samples as a proposal distribution. For a sample of size N from $\mathrm{p}_{\mathrm{meld},1}$ denoted $\{\phi_n^{(\mathrm{meld},1)}\}_{n=1}^N$ we sample an index n^* uniformly at random between 1 and N, and use the corresponding value as the proposal $\phi^* = \phi_{n^*}^{(\mathrm{meld},1)}$. This results in a stage two acceptance probability for a move from $\phi \rightarrow \phi^*$ of

$$\alpha(\phi^*, \phi) = \frac{\mathrm{p}_{\mathrm{pool}}(\phi^*)\mathrm{p}_1(\phi^*, \psi_1, Y_1)\mathrm{p}_2(\phi^*, \psi_2, Y_2)\mathrm{p}_1(\phi)\mathrm{p}_2(\phi)}{\mathrm{p}_{\mathrm{pool}}(\phi)\mathrm{p}_1(\phi, \psi_1, Y_1)\mathrm{p}_2(\phi, \psi_2, Y_2)\mathrm{p}_1(\phi^*)\mathrm{p}_2(\phi^*)} \frac{\mathrm{p}_1(\phi, \psi_1, Y_1)\mathrm{p}_1(\phi^*)}{\mathrm{p}_1(\phi^*, \psi_1, Y_1)\mathrm{p}_1(\phi)}$$

$$= \frac{\mathrm{p}_{\mathrm{pool}}(\phi^*)\mathrm{p}_2(\phi^*, \psi_2, Y_2)\mathrm{p}_2(\phi)}{\mathrm{p}_{\mathrm{pool}}(\phi)\mathrm{p}_2(\phi, \psi_2, Y_2)\mathrm{p}_2(\phi^*)}, \tag{5}$$

since all stage one terms cancel, providing a form of "modularisation" in the algorithm. The update for ψ_2 has an acceptance probability for a move from $\psi_2 \rightarrow \psi_2^*$, drawn from a proposal distribution $\mathrm{q}(\psi_2^* \mid \psi_2)$, of

$$\alpha(\psi_2^*, \psi_2) = \frac{\mathrm{p}_2(\phi, \psi_2^*, Y_2)}{\mathrm{p}_2(\phi, \psi_2, Y_2)} \frac{\mathrm{q}(\psi_2 \mid \psi_2^*)}{\mathrm{q}(\psi_2^* \mid \psi_2)},$$

as all terms that do not contain ψ_2 cancel. Samples from the melded posterior distribution for ψ_1, $\mathrm{p}_{\mathrm{meld}}(\psi_1 \mid Y_1, Y_2)$, can be obtained by storing the indices n used to draw values of ϕ from $\{\phi_n^{(\mathrm{meld},1)}\}_{n=1}^N$ in stage two. The stored indices are then used to resample the stage one samples $\{\psi_{1,n}^{(\mathrm{meld},1)}\}_{n=1}^N$ yielding samples from $\mathrm{p}_{\mathrm{meld}}(\psi_1 \mid Y_1, Y_2)$.

An interesting property of Equations (4) and (5) is that our interaction with the unknown prior marginal distribution is limited to the *self-density ratio* $\mathrm{r}(\phi, \phi^*) = \mathrm{p}_m(\phi) / \mathrm{p}_m(\phi^*)$. In Sect. 3 we develop methodology that uses self-density ratios to improve the numerical stability of the acceptance probability calculations.

We do not have to target $\mathrm{p}_{\mathrm{meld},1}(\phi, \psi_1 \mid Y_1)$ with an improper prior in stage one; we are free to choose any of the components of Eq. (3). The choice of stage one components will affect MCMC mixing, yet is often constrained by the practicalities of sampling the subposterior distributions. In the example of Sect. 5 the common quantity ϕ is a non-invertible function of parameters in p_1, and it is possible to sample from the subposterior $\mathrm{p}_1(\phi, \psi_1 \mid Y_1)$ using `JAGS`. Hence, we draw stage one samples from $\mathrm{p}_1(\phi, \psi_1 \mid Y_1)$, with stage two, implemented partially in `Stan`, accounting for the remaining terms: $1 / \mathrm{p}_1(\phi)$, $\mathrm{p}_2(\phi, \psi_2 \mid Y_2) / \mathrm{p}_2(\phi)$, and $\mathrm{p}_{\mathrm{pool}}(\phi)$. This process highlights another interesting advantage of Markov melding; we can use samples produced from one statistical software package in combination with a model implemented in another, mixing and matching as is most convenient.

2.3 Naive prior marginal estimation

The expressions in Eqs. (4) and (5) explicitly include both models' prior marginal distributions $\mathrm{p}_m(\phi)$ for $m = 1, 2$, and implicitly includes them in $\mathrm{p}_{\mathrm{pool}}(\phi)$. In our examples we do not have analytic expressions for these marginals. More generally, if ϕ is not a root node in the directed acyclic graph representation of either submodel (see e.g. π_{12} in Fig. 1), or is the aggregate output of a non-invertible deterministic link function, then the analytic form of $\mathrm{p}_m(\phi)$ will likely be intractable.

The approach proposed by Goudie et al. (2019), which we call the *naive* approach, estimates the prior marginal dis-

tributions by sampling $\mathrm{p}_m(\phi, \psi_m, Y_m)$ for each model using simple Monte Carlo, as the samples of ϕ will be distributed according to the correct marginal, and employs a standard KDE $\widehat{\mathrm{p}}_m(\phi)$ (Wand and Jones 1995). The two-stage sampler then targets the corresponding estimate of the melded posterior

$$\hat{\mathrm{p}}_{\mathrm{meld}}(\phi, \psi_1, \psi_2 \mid Y_1, Y_2) \propto \hat{\mathrm{p}}_{\mathrm{pool}}(\phi) \frac{\mathrm{p}_1(\phi, \psi_1, Y_1)}{\hat{\mathrm{p}}_1(\phi)} \frac{\mathrm{p}_2(\phi, \psi_2, Y_2)}{\hat{\mathrm{p}}_2(\phi)}, \tag{6}$$

where $\hat{\mathrm{p}}_{\mathrm{pool}}(\phi)$ is the approximation to $\mathrm{p}_{\mathrm{pool}}(\phi)$ obtained by plugging in $\hat{\mathrm{p}}_m(\phi)$ for $m = 1, 2$.

2.4 Numerical issues in the naive approach

Sampling the melded posterior using Eq. (6) can be numerically unstable. Say we propose a move from $\phi \to \phi^*$, where ϕ^* is particularly improbable under p_m. The KDE estimate at this value, $\widehat{\mathrm{p}}_m(\phi^*)$, is poor in terms of relative error

$$\left\| \frac{\widehat{\mathrm{p}}_m(\phi^*) - \mathrm{p}_m(\phi^*)}{\mathrm{p}_m(\phi^*)} \right\|_1,$$

particularly in the tails of the distribution (Koekemoer and Swanepoel 2008). In our experience, the KDE is typically an underestimate in the tails, which can lead to an explosion in the self-density ratio estimate $\hat{\mathrm{r}}(\phi, \phi^*) = \hat{\mathrm{p}}_m(\phi) / \hat{\mathrm{p}}_m(\phi^*)$. Hence, improbable values for ϕ^* are accepted far too often. Once at this improbable value, i.e. when ϕ is improbable under $\mathrm{p}_m(\phi)$, the error in the KDE then leads to a dramatically reduced value for the acceptance probability. This results in Markov chains that get stuck at improbable values. For example, see the top left panel of Fig. 5.

In which stage this instability arises depends on which prior marginal densities are intractable, and how the terms in Eq. (3) are apportioned across the stages. In the example of Sect. 4, $\mathrm{p}_1(\phi)$ is unknown and is part of both stage one (in Eq. (4)) and stage two (via $\mathrm{p}_{\mathrm{pool}}(\phi)$ in Eq. (5)). Thus both stages are numerically brittle. Our second example, contained in Sect. 5, represents a more typical scenario, where the first submodel posterior is used as the proposal for the melded posterior. In this case, all unknown prior marginal terms are factorised into the stage two target, and the instability is confined to the second stage.

3 Self-density ratio estimation

As described in Sect. 2.4, the self-density ratios associated with both $\mathrm{p}_1(\phi)$ and $\mathrm{p}_2(\phi)$ may be required by the two-stage MCMC algorithm. To simplify notation, we consider in this section a generic joint density $\mathrm{p}(\phi, \gamma)$ that we can evaluate pointwise, but whose marginal $\mathrm{p}(\phi) = \int \mathrm{p}(\phi, \gamma)\mathrm{d}\gamma$ we cannot obtain analytically. Our interest is in the *self-density ratio* evaluated at ϕ_{nu} and ϕ_{de} (the subscripts are abbreviations of numerator and denominator respectively) which we denote as

$$\mathrm{r}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}) = \frac{\mathrm{p}(\phi_{\mathrm{nu}})}{\mathrm{p}(\phi_{\mathrm{de}})} = \frac{\int \mathrm{p}(\phi_{\mathrm{nu}}, \gamma)\mathrm{d}\gamma}{\int \mathrm{p}(\phi_{\mathrm{de}}, \gamma)\mathrm{d}\gamma}.$$

In our examples we set $\phi_{\mathrm{nu}} = \phi$ and $\phi_{\mathrm{de}} = \phi^*$ for use in Eqs. (4) and (5); and define $\gamma = (\psi_m, Y_m)$ and $\mathrm{p} = \mathrm{p}_m$ where $m = 1$ or 2 as appropriate (see Sects. 4 and 5 for details).

To avoid the numerical issues associated with the naive approach, we need to improve the ratio estimate $\widehat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ for improbable values of ϕ_{nu} and ϕ_{de}, e.g. values more than two standard deviations away from the mean. The fundamental flaw in the naive approach in this context is that it minimises the absolute error in the high density region (HDR) of $\mathrm{p}(\phi)$, i.e. the region $R_\varepsilon(\mathrm{p}(\phi)) = \{\phi : \mathrm{p}(\phi) > \varepsilon\}$. But this is not necessarily the sole region of interest, and we are concerned with minimising the relative error. To address this we reweight $\mathrm{p}(\phi)$ towards a particular region, and thus obtain a more accurate estimate in that region. We then exploit the fact that we only interact with the prior marginal distribution via its self-density ratio to combine estimates from multiple reweighted distributions.

3.1 Single weighting function

We can shift $\mathrm{p}(\phi)$ by multiplying the joint distribution $\mathrm{p}(\phi, \gamma)$ by a known weighting function $\mathrm{w}(\phi; \xi)$, controlled by parameter ξ, then account for this shift in our KDE. This will improve the accuracy of the KDE in the region to which we shift the marginal. We first generate N samples denoted $\{(\phi_n, \gamma_n)\}_{n=1}^N$, from a weighted version of the joint distribution

$$\{(\phi_n, \gamma_n)\}_{n=1}^N \sim \frac{1}{Z_1} \mathrm{p}(\phi, \gamma)\mathrm{w}(\phi; \xi), \tag{7}$$

where $Z_1 = \iint \mathrm{p}(\phi, \gamma)\mathrm{w}(\phi; \xi)\mathrm{d}\phi\mathrm{d}\gamma$ is the normalising constant. The samples $\{\phi_n^{(\mathrm{s})}\}_{n=1}^N$, obtained by ignoring the samples of γ_n, are distributed according to a weighted version $\mathrm{s}(\phi; \xi)$ of the marginal distribution $\mathrm{p}(\phi)$

$$\{\phi_n^{(\mathrm{s})}\}_{n=1}^N \sim \frac{1}{Z_2} \mathrm{p}(\phi)\mathrm{w}(\phi; \xi) = \mathrm{s}(\phi; \xi),$$

where $Z_2 = \int \mathrm{p}(\phi)\mathrm{w}(\phi; \xi)\mathrm{d}\phi$. Typically (7) cannot be sampled by simple Monte Carlo; instead we employ MCMC.

Using the samples $\{\phi_n^{(\mathrm{s})}\}_{n=1}^N$ from $\mathrm{s}(\phi; \xi)$ we compute a weighted kernel density estimate (Jones 1991), with band-

width h, kernel K_h, and normalising constant Z_3

$$\hat{\mathrm{p}}(\phi) = \frac{1}{Z_3 N h} \sum_{n=1}^{N} (\mathrm{w}(\phi_n; \xi))^{-1} \mathrm{K}_h(\phi - \phi_n^{(\mathrm{s})}), \tag{8}$$

and form our weighted-sample self-density ratio estimate

$$\hat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}) = \frac{\hat{\mathrm{p}}(\phi_{\mathrm{nu}})}{\hat{\mathrm{p}}(\phi_{\mathrm{de}})} = \frac{\sum_{n=1}^{N} (\mathrm{w}(\phi_n; \xi))^{-1} \mathrm{K}_h(\phi_{\mathrm{nu}} - \phi_n^{(\mathrm{s})})}{\sum_{n=1}^{N} (\mathrm{w}(\phi_n; \xi))^{-1} \mathrm{K}_h(\phi_{\mathrm{de}} - \phi_n^{(\mathrm{s})})}.$$

The cancellation of the normalisation constant Z_3 is crucial, as accurately estimating constants like Z_3 is known to be challenging.

3.2 Choice of weighting function

The choice of $\mathrm{w}(\phi; \xi)$ affects both the validity and efficacy of our methodology. The weighted marginal $\mathrm{s}(\phi; \xi)$ must satisfy the requirements for a density for our method to be valid. Hence, the specific form of $\mathrm{w}(\phi; \xi)$ is subject to some restrictions. Our first requirement is that $\mathrm{w}(\phi; \xi) > 0$ for all ϕ in the support of $\mathrm{p}(\phi, \gamma)$. We also require that the weighted joint distribution, defined in (7), has finite integral, to ensure that it can be normalised to a probability distribution, and that the marginal $\mathrm{s}(\phi; \xi)$ is positive over the support of interest, also with finite integral.

3.3 Multiple weighting functions

The methodology of Sect. 3.1 produces a single estimate $\hat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ using $\hat{\mathrm{p}}(\phi)$ from Eq. (8). It is accurate for values in the HDR of $\mathrm{s}(\phi; \xi)$, i.e. $R_\varepsilon(\mathrm{s}(\phi; \xi))$, and we can control the location of $R_\varepsilon(\mathrm{s}(\phi; \xi))$ through ξ. This is similar to importance sampling, with $\mathrm{s}(\phi; \xi)$ acting as the proposal density. Nakayama (2011) notes importance sampling can be used to improve the mean square error (MSE) of a KDE in a specific local region, at the cost of an increase in global MSE. To ameliorate the decrease in global performance, we specify multiple regions in which we want accurate estimates for $\hat{\mathrm{p}}(\phi)$, and then combine the corresponding estimates of $\hat{\mathrm{r}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ to provide a single estimate that is accurate across all regions.

We elect to use W different weighting functions, indexed by $w = 1, \ldots, W$, with function-specific parameters ξ_w denoted $\mathrm{w}(\phi; \xi_w)$. Samples are then drawn from each of the W weighted distributions $\mathrm{s}_w(\phi; \xi_w) \propto \mathrm{p}(\phi)\mathrm{w}(\phi; \xi_w)$. Denote the samples from the w^{th} weighted distribution by $\{\phi_n^{(\mathrm{s}_w)}\}_{n=1}^N$. Each set of samples produces a separate ratio estimate $\hat{\mathrm{r}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ in the manner described in Sect. 3.1.

Each individual $\hat{\mathrm{r}}_w$ is accurate (in terms of relative accuracy) only in the HDR of $\mathrm{s}_w(\phi; \xi_w)$. Thus, when combining multiple ratio estimates, simply taking the mean of all $w = 1, \ldots, W$ estimates (for a specific value of ϕ_{nu} and ϕ_{de}) would not make use of our knowledge about the region in which $\hat{\mathrm{r}}_w$ is accurate. We therefore propose a weighted average of all the individual ratio estimates, where the weights approximately come from $\mathrm{s}_w(\phi_{\mathrm{nu}}; \xi_w)\mathrm{s}_w(\phi_{\mathrm{de}}; \xi_w)$ – this quantity is largest when $\hat{\mathrm{r}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ is most accurate. This ensures the more accurate terms are given more weight in our final estimate. Specifically, we use $\{\phi_n^{(\mathrm{s}_w)}\}_{n=1}^N$ to compute a standard KDE of $\mathrm{s}_w(\phi; \xi_w)$

$$\hat{\mathrm{s}}_w(\phi; \xi_w) = \frac{1}{Nh} \sum_{n=1}^{N} \mathrm{K}_h(\phi - \phi_n^{(\mathrm{s}_w)}).$$

Finally, we form the weighted-sample self-density ratio estimate $\hat{\mathrm{r}}_{\mathrm{WSRE}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$, which is a weighted mean of the individual ratio estimates

$$\hat{\mathrm{r}}_{\mathrm{WSRE}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}) = \frac{1}{Z_4} \sum_{w=1}^{W} \hat{\mathrm{s}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}; \xi_w) \hat{\mathrm{r}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}),$$

where $\hat{\mathrm{s}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}; \xi_w) = \hat{\mathrm{s}}_w(\phi_{\mathrm{nu}}; \xi_w)\, \hat{\mathrm{s}}_w(\phi_{\mathrm{de}}; \xi_w)$ and $Z_4 = \sum_{w=1}^{W} \hat{\mathrm{s}}_w(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}}; \xi_w)$.

3.4 Choosing values for ξ_w

Consider a D-dimensional $\phi = (\phi^{[1]}, \ldots, \phi^{[D]})$ where $\phi^{[d]}$ is the d^{th} component of ϕ, for $d = 1, \ldots, D$. Assume we have a compact region of interest for the d^{th} component denoted $A_d = [a_d, b_d] \subseteq \mathrm{supp}(\phi^{[d]})$, such that the overall region of interest A can be defined as the Cartesian product of component-wise regions of interest $A = \times_{d=1}^{D} A_d$. We are interested in accurately evaluating the self-density ratio for two points in this region. We will obtain W choices for ξ_w by specifying V weighting functions for each of the D components, such that $W = V^D$.

Assume that the weighting function $\mathrm{w}(\phi; \xi)$ is composed of D independent component weighting functions

$$\mathrm{w}(\phi; \xi) = \prod_{d=1}^{D} \mathrm{m}(\phi^{[d]}; \xi^{[d]}),$$

where $\xi^{[d]}$ is the d^{th} component of ξ. We can then define the marginal of the weighted target

$$\mathrm{t}(\phi^{[d]}; \xi^{[d]}) = \int \mathrm{s}(\phi; \xi) \mathrm{d}\phi^{[-d]},$$

where $\phi^{[-d]}$ represents the $D-1$ components of ϕ that are not $\phi^{[d]}$. For typical choices of ξ and $\mathrm{w}(\phi; \xi)$, the corresponding HDR of $\mathrm{t}(\phi^{[d]}; \xi^{[d]})$ does not span the region of interest. That is, $|R_\varepsilon(\mathrm{t}(\phi^{[d]}; \xi^{[d]}))| \ll |A_d|$.

Our aim is to choose, for each of the d components, values $v = 1, \ldots, V$ of $\xi^{[d]}$ denoted $\{\xi_{v,d}\}_{v=1}^{V}$, yielding weighting functions $\mathrm{m}(\phi^{[d]}, \xi_{v,d})$ and corresponding $\mathrm{t}(\phi^{[d]}, \xi_{v,d})$, such that $\bigcup_{v=1}^{V} R_\varepsilon(\mathrm{t}(\phi^{[d]}; \xi_{v,d})) \approx A_d = [a_d, b_d]$. We employ the following heuristic argument, first choosing a "minimum" $\xi_{1,d}$ and a "maximum" $\xi_{V,d}$ such that

$$\xi_{1,d}\colon\ a_d \in R_\varepsilon(\mathrm{t}(\phi^{[d]}; \xi_{1,d})),$$
$$\xi_{V,d}\colon\ b_d \in R_\varepsilon(\mathrm{t}(\phi^{[d]}; \xi_{V,d})).$$

In words, we choose a minimum value $\xi_{1,d}$ so that the corresponding HDR of the weighted target includes the lower limit of the region of interest. An analogous argument is used to choose the maximum $\xi_{V,d}$. We then interpolate $V-2$ values between $\xi_{1,d}$ and $\xi_{V,d}$ ensuring that there is sufficient, but not complete, overlap between the corresponding HDRs.

Denote an element from the set of all W possible values for the parameter of the weighting function with $\xi_w \in \times_{d=1}^{D}\{\xi_{v,d}\}_{v=1}^{V}$, noting that ξ_w is a D-vector.

The practitioner typically has some knowledge of $p(\phi)$ and A from prior predictive checks and previous attempts at running the two-stage sampler. Thus only a small number of trial-and-error attempts should be needed to determine $\xi_{1,d}$ and $\xi_{V,d}$ for all dimensions. These attempts are also used to check for overlap between the HDRs, and increase V if the overlap is insufficient. Section 6 contains further discussion of this selection process and its relationship to umbrella sampling (Torrie and Valleau 1977)

3.5 Practicalities and software

In our examples we use Gaussian density functions for $\mathrm{m}(\phi^{[d]}; \xi_{v,d})$,

$$\mathrm{m}(\phi^{[d]}; \xi_{v,d}) = \frac{1}{\sqrt{2\pi\sigma_{v,d}^2}} \exp\left\{-\frac{1}{2\sigma_{v,d}^2}(\phi - \mu_{v,d})^2\right\},$$

with $\xi_{v,d} = (\mu_{v,d}, \sigma_{v,d}^2)$, though we fix $\sigma_{v,d}^2 = \sigma_d^2$ for all v. Our definition of sufficient overlap is that 0.95 empirical quantile of $\mathrm{t}(\phi^{[d]}; \xi_{v,d})$ is equal or slightly greater than the 0.05 empirical quantile of $\mathrm{t}(\phi^{[d]}; \xi_{v+1,d})$, for $v = 1, \ldots, V-1$.

Our implementation of our WSRE methodology is available in an `R` (R Core Team 2021) package at https://github.com/hhau/wsre. It is built on top of `Stan` (Carpenter et al. 2017) and `Rcpp` (Eddelbuettel and François 2011). Package users supply a joint density $\mathrm{p}(\phi, \gamma)$ in the form of a `Stan` model; choose the parameters ξ_w of each of the W weighting functions; and the number of samples N drawn from each $\mathrm{s}_w(\phi; \xi_w)$. The combined estimate $\hat{\mathrm{r}}_{\mathrm{WSRE}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ is returned. A vignette on using `wsre` is included in the package, and documents the specific form of `Stan` model required.

4 An evidence synthesis for estimating the efficacy of HIV screening

To illustrate our approach we artificially split an existing joint model into two submodels, then compare the melded posterior estimates obtained by the two-stage algorithm using the naive and WSRE approaches. Artificially splitting this joint model serves several purposes: it demonstrates that the numerical instability can occur in a simple, low dimensional setting; we can obtain a good parametric approximation to the prior marginal to use as a reference; and the simplicity of the model allows us to focus on our methodology, not the complexity of the model.

4.1 Model

The model is a simple evidence synthesis model for inferring the efficacy of HIV screening in prenatal clinics (Ades and Cliffe 2002), and has 8 *basic parameters* $\rho_1, \rho_2, \ldots, \rho_8$, which are group membership probabilities for particular risk groups and subgroups thereof. The first risk group partitions the prenatal clinic attendees into those born in sub-Saharan Africa (SSA), injecting drug users (IDU), and the remaining women. These groups have corresponding membership probabilities ρ_1, ρ_2, and $1 - \rho_1 - \rho_2$. The groups are subdivided based on whether they are infected with HIV, with group specific HIV positivity ρ_3, ρ_4 and ρ_5 respectively; and if they had already been diagnosed before visiting the clinic, with pre-clinical diagnosis probabilities ρ_6, ρ_7 and ρ_8. An additional probability is also included in the model, denoted ρ_9, which considers the prevalence of HIV serotype B. This parameter enables the inclusion of study 12, which further informs the other basic parameters. Table 1 summarises the full joint model, including the $s = 1, \ldots, 12$ studies with observations y_s and sample size n_s; the basic parameters $\rho_1, \ldots, \rho_9$; and the link functions that relate the study proportions $\pi_1, \ldots, \pi_{12}$ to the basic parameters.

We make one small modification to original model of Ades and Cliffe (2002), to better highlight the impact of WSRE on the melded posterior estimate. The original model adopts a flat, Beta(1, 1) prior for ρ_9. This induces a prior on π_{12} that is not flat, but not overly informative, making it difficult to demonstrate the issues caused by an inaccurate density estimate of the tail of the prior marginal distribution. Instead, we adopt a Beta(3, 1) prior for ρ_9. This prior would have been reasonable for the time and place in which the original evidence synthesis was constructed, since the distribution of HIV serotypes differs considerably between North America and sub-Saharan Africa (Hemelaar 2012).

Table 1 HIV example: Study probabilities, their relationships to the basic parameters, and data used to inform the probabilities. For full details on the sources of the data see Ades and Cliffe (2002).

Parameter	Data		
	y	n	y/n
$\pi_1 = \rho_1$	11,044	104,577	0.106
$\pi_2 = \rho_2$	12	882	0.014
$\pi_3 = \rho_3$	252	15,428	0.016
$\pi_4 = \rho_4$	10	473	0.021
$\pi_5 = \frac{\rho_4\rho_2+\rho_5(1-\rho_1-\rho_2)}{1-\rho_1}$	74	136,139	0.001
$\pi_6 = \rho_3\rho_1 + \rho_4\rho_2 + \rho_5(1-\rho_1-\rho_2)$	254	102,287	0.002
$\pi_7 = \frac{\rho_6\rho_3\rho_1}{\rho_6\rho_3\rho_1+\rho_7\rho_4\rho_2+\rho_8\rho_5(1-\rho_1-\rho_2)}$	43	60	0.717
$\pi_8 = \frac{\rho_7\rho_4\rho_2}{\rho_7\rho_4\rho_2+\rho_8\rho_5(1-\rho_1-\rho_2)}$	4	17	0.235
$\pi_9 = \frac{\rho_6\rho_3\rho_1+\rho_7\rho_4\rho_2+\rho_8\rho_5(1-\rho_1-\rho_2)}{\rho_3\rho_1+\rho_4\rho_2+\rho_5(1-\rho_1-\rho_2)}$	87	254	0.343
$\pi_{10} = \rho_7$	12	15	0.800
$\pi_{11} = \rho_9$	14	118	0.119
$\pi_{12} = \frac{\rho_4\rho_2+\rho_9\rho_5(1-\rho_1-\rho_2)}{\rho_4\rho_2+\rho_5(1-\rho_1-\rho_2)}$	5	31	0.161

The code to reproduce this example is available at https://github.com/hhau/presanis-conflict-hiv-example.

4.2 Submodels formed by splitting

In the full joint model study 12 informs the probability π_{12}, and provides indirect evidence for the basic parameters through the deterministic link function

$$\pi_{12} = \frac{\rho_2\rho_4 + \rho_9\rho_5(1-\rho_1-\rho_2)}{\rho_2\rho_4 + \rho_5(1-\rho_1-\rho_2)}.$$

Figure 1 is a DAG of the basic parameters in the full model that relate to π_{12}. We consider splitting the model at the node corresponding to the expected proportion π_{12} in study 12, i.e. we set the common quantity $\phi = \pi_{12}$.

The first submodel ($m = 1$) considers data from studies 1 to 11 $Y_1 = (y_1, \ldots, y_{11})$, corresponding study proportions $(\pi_1, \ldots, \pi_{11})$, and all basic parameters $\psi_1 = (\rho_1, \ldots, \rho_9)$. Note that the study proportions are implicitly defined because they are deterministic functions of the basic parameters. The joint distribution of this submodel is

$$\mathrm{p}_1(\rho_1, \ldots, \rho_9, y_1, \ldots, y_{12}) = \mathrm{p}(\rho_1)\ldots\mathrm{p}(\rho_9)\prod_{s=1}^{11}\mathrm{p}(y_s \mid \pi_s).$$

The prior $\mathrm{p}_1(\pi_{12})$ on the common quantity $\phi = \pi_{12}$ is implicitly defined, so its analytic form is unknown, hence it needs to be estimated.

The second submodel ($m = 2$) pertains specifically to study 12, with data $Y_2 = y_{12}$, the study 12 specific probability $\phi = \pi_{12}$, and $\psi_1 = \varnothing$. The joint distribution is $\mathrm{p}_2(\pi_{12}, y_{12}) = \mathrm{p}_2(\pi_{12})\mathrm{p}(y_{12} \mid \pi_{12})$. In more complex examples $\mathrm{p}_2(\phi)$ may be implicitly defined, and contribute substantially to the melded posterior. However, in this simple example we are free to choose $\mathrm{p}_2(\pi_{12}) = \mathrm{p}_2(\phi)$, and opt for a Beta(1, 1) prior.

4.3 Self-density ratio estimation

We now compute the self-density ratio estimate $\hat{\hat{\mathrm{r}}}_{\mathrm{WSRE}}(\phi_{\mathrm{nu}}, \phi_{\mathrm{de}})$ of $\mathrm{p}_1(\phi_{\mathrm{nu}}) / \mathrm{p}_1(\phi_{\mathrm{de}})$. In the notation defined in Sect. 3.4, this example has $D = 1$, and we use $V = W = 7$ Gaussian weighting functions. We fix the variance parameter of the weighting function $\sigma_w^2 = 0.08^2$ for all w, and use the heuristic described in Sect. 3.4 to choose values for the mean parameter of the weighting function. Specifically, we set the minimum to be $\xi_{1,1} = \mu_1 = 0.05$, with maximum $\xi_{7,1} = \mu_7 = 0.08$ and 5 additional, equally spaced values between these extrema. We draw 3000 MCMC samples in total, apportioned equally across the 7 weighting functions. We thus draw 428 post warmup MCMC samples from each weighted target.

4.4 Results

We compare the melded posterior obtained by the naive approach and using WSRE. For a fair comparison, we estimate the prior marginal distribution of interest $\widehat{\mathrm{p}}_1(\phi)$ using 3000 Monte Carlo samples, This set-up is slightly advantageous for the naive approach, which uses Monte Carlo samples, rather than the MCMC samples of the self-density ratio estimate; the naive approach makes use of a sample comprised of 3000 effective samples, whilst the self-density ratio estimate uses fewerthan 3000 effective samples. A reference estimate of the melded posterior is obtained using a parametric density estimate $\hat{\mathrm{p}}_{\mathrm{ref},1}(\phi)$ for the unknown prior marginal, based on 5×10^6 prior samples. The reference sample also

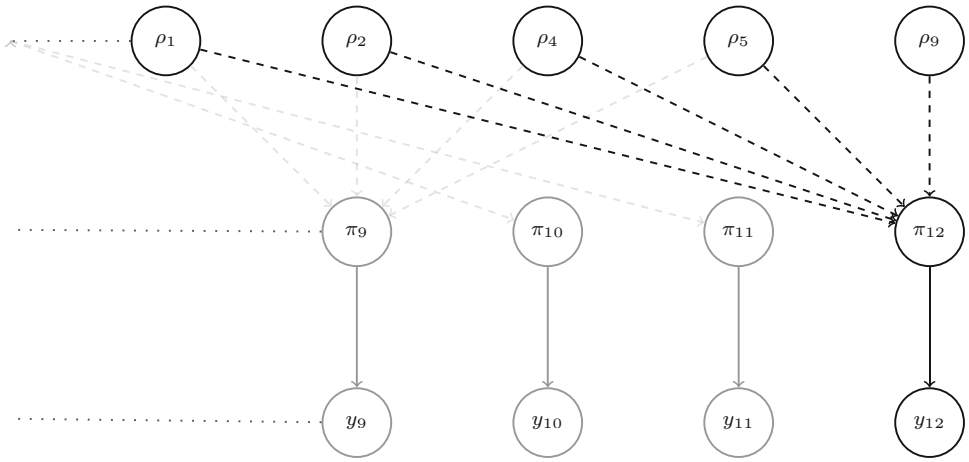

Fig. 1 Partial directed acyclic graph (DAG) for the HIV model. The top row only depicts nodes that relate to π_{12}. Dashed lines indicate deterministic relationships between nodes, some of which are non-invertible. Solid lines indicate stochastic relationships

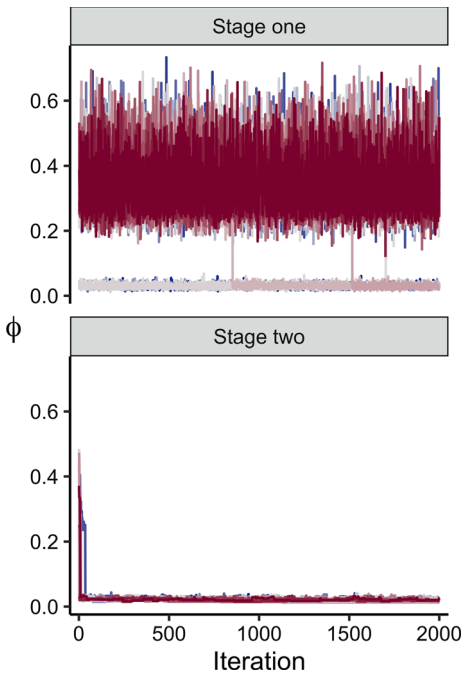

Fig. 2 *Top*: Stage one trace plot for ϕ using the naive method. At any moment in time chains can jump to the spurious mode, which is an artefact of $\hat{p}_1(\phi)$. *Bottom*: Corresponding stage two trace plot. The stage two target has the same numerical instability, and because the stage one samples are the proposal distribution, all chains encounter the instability

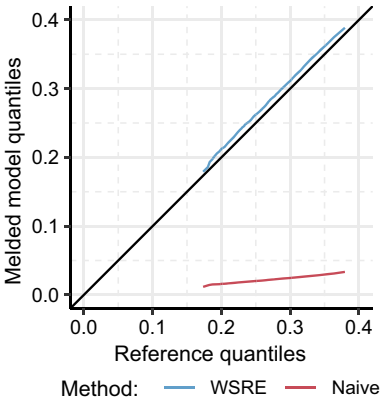

Fig. 3 Quantile-quantile plot of the melded posterior quantiles using the WSRE approach (blue) and the naive approach (red). Both methods are comparable to the quantiles from the reference sample (x-axis)

contains some error, as $\hat{p}_{\text{ref},1}(\phi)$ is not perfect. However, in the absence of an analytic form for $p_1(\phi)$ it serves as a very close approximation. We estimate the melded posterior using the two-stage sampler of Sect. 2.2, targeting in stage one

$$p_{\text{meld},1}(\phi, \psi_1 \mid Y_1) \propto \frac{p_1(\phi, \psi_1, Y_1)}{p_1(\phi)}. \tag{9}$$

and the full melded posterior in stage two. To demonstrate the numerical instability of interest, we run 24 chains that target $p_{\text{meld},1}$ in (9) using the naive approach.

The top panel of Fig. 2 displays the trace plot of the post-warmup samples. Many chains have already converged to a spurious model around $\phi \approx 0.02$, and other chains jump to this mode after a variable number of additional iterations. As discussed in Sect. 2.4, this mode is an artefact of the naive KDE employed for $\hat{p}_1(\phi)$, and is also visible in the corresponding stage two trace plot (bottom panel of Fig. 2). Because the stage one samples act as the proposal for stage two, all stage two chains quickly jump to the spurious mode.

The samples surrounding the spurious mode introduce substantial bias in estimate of the melded posterior under the naive approach. This is visible in the quantile-quantile plot in Fig. 3, where the naive approach produces an implausible estimate compared to the reference quantiles. In contrast, the WSRE approach rectifies the numerical instability, and uses the two-stage sampler to produce a sensible estimate of the melded posterior.

5 An evidence synthesis to estimate the severity of the H1N1 pandemic

We now consider a more involved example, where the prior for the common quantity does not have an analytical form under either submodel, and the two priors contain a substantially different quantity of information. Presanis et al. (2014) undertook a large evidence synthesis in order to estimate the severity of the H1N1 pandemic amongst the population of England. This model combines independent data on the number of suspected influenza cases in hospital's intensive care unit (ICU) into a large severity model. Here, we reanalyse the model introduced in Goudie et al. (2019) that uses Markov melding to join the independent ICU model ($m = 1$) with a simplified version of the larger, complex severity model ($m = 2$). In this example the melded model has no obvious implied joint model, so there are no simple "gold standard" joint model estimates to use as a baseline reference. However, we demonstrate that the naive approach is highly unstable, whereas the WSRE approach produces stable results. The code to reproduce all figures and outputs for this example is available at https://github.com/hhau/full-melding-example.

5.1 ICU submodel

The data for the ICU submodel ($m = 1$) are aggregate weekly counts of patients in the ICU of all the hospitals in England, for 78 days between December 2010 and February 2011. Observations were recorded of the number of children $a = 1$ and adults $a = 2$ in the ICU on days $U = \{8, 15, \ldots, 78\}$, and we denote a specific weekly observation as $y_{a,t}$ for $t \in U$.

To appropriately model the temporal nature of the weekly ICU data we use a time inhomogeneous, thinned Poisson process with rate parameter $\lambda_{a,t}$ for $t \in T$ where $T = \{1, 2, \ldots, 78\}$. This is the expected number of new ICU admissions; the corresponding age group specific ICU exit rate is μ_a. There is also a discrepancy between the observation times U and the daily support of our Poisson process T. We address this in the observation model

$$y_{a,t} \sim \text{Pois}(\eta_{a,t}), \ t \in U, \tag{10}$$

$$\eta_{a,t} = \sum_{u=1}^{t} \lambda_{a,u} \exp\{-\mu_a(t-u)\}, \ t \in T, \tag{11}$$

through different supports for t in Eqs. (10) and (11). An identifiability assumption of $\eta_{a,1} = 0$ is required, which enforces the reasonable assumption that no H1N1 influenza patients were in the ICU at time $t = 0$.

Weekly virological positivity data are available at weeks $V = \{1, \ldots, 11\}$, and inform the proportion of influenza cases which are attributable to the H1N1 virus $\pi^{\text{pos}}_{a,t}$. The virology data consists of the number of H1N1-positive swabs $z^{\text{pos}}_{a,v}$ and the total number of swabs $n^{\text{pos}}_{a,z}$ tested for influenza that week. This proportion relates the counts to $\pi^{\text{pos}}_{a,t}$ via a truncated uniform prior on $\pi^{\text{pos}}_{a,t}$,

$$\pi^{\text{pos}}_{a,t} \sim \text{Unif}(\omega_{a,v}, 1), \ t \in T$$
$$z^{\text{pos}}_{a,v} \sim \text{Bin}(n^{\text{pos}}_{a,z}, \omega_{a,v}), \ v \in V,$$

with $v = 1$ for $t = 1, 2, \ldots, 14$, and $v = \lfloor (t-1)/7 \rfloor$ for $t = 15, 16, \ldots, 78$ to align the temporal indices. The positivity proportion $\pi^{\text{pos}}_{a,t}$ is combined with $\lambda_{a,t}$ to compute the lower bound on the total number of H1N1 cases $\phi_a = \sum_{t \in T} \pi^{\text{pos}}_{a,t} \lambda_{a,t}$ where $\phi = (\phi_1, \phi_2)$ is the quantity common to both submodels. This summation is a non-invertible function, which necessitates either considering this model in stage one of our two-stage sampler, or appropriately augmenting the definition of ϕ_a such that it is invertible. We elect to consider this submodel in stage one, and further discuss model ordering in Sect. 6.

Lastly, we specify priors for the remaining parameters. A lognormal random-walk is used for the expected number of new admissions

$$\log(\lambda_{a,1}) \sim \text{Unif}(0, 250), \ \log(\lambda_{a,t}) \sim \text{N}(\log(\lambda_{a,t-1}), \nu_a^{-2}),$$
$$\nu_a \sim \text{Unif}(0.1, 2.7),$$

for $t = 2, 3, \ldots, 78$ and $a = 1, 2$. Age group specific exit rates have informative priors (Presanis et al. 2014)

$$\mu_1 = \exp\{-\alpha\}, \ \mu_2 = \exp\{-(\alpha + \beta)\},$$
$$\alpha \sim \text{N}(2.7058, 0.0788^2), \ \beta \sim \text{N}(-0.4969, 0.2048^2),$$

and the lower bound on the positivity proportion has a flat prior, $\omega_{a,v} \sim \text{Unif}(0, 1)$, for $v \in V$.

5.2 Severity submodel

A simplified version of the large severity model ($m = 2$) of Presanis et al. (2014) is considered here, in which parts of the severity model are collapsed into informative priors. The cumulative number of ICU admissions ϕ_a is assumed to be an underestimate of the true number of ICU admissions due to H1N1, χ_a. This motivates

$$\phi_a \sim \text{Bin}(\chi_a, \pi^{\text{det}}), \ \pi^{\text{det}} \sim \text{Beta}(6, 4),$$
$$\chi_1 \sim \text{LN}(4.93, 0.17^2), \ \chi_2 \sim \text{LN}(7.71, 0.23^2),$$

where π^{det} is the age group constant probability of detection, and the priors on χ_a appropriately summarise the remainder of the large severity model.

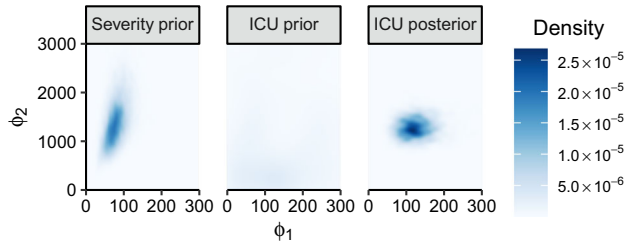

Fig. 4 Heatmap of the severity submodel prior $\mathrm{p}_2(\phi)$, ICU submodel prior $\mathrm{p}_1(\phi)$, and the stage one (ICU submodel) posterior $\mathrm{p}_1(\phi \mid Y_1)$

5.3 Prior distributions, stage one target

Figure 4 displays $\mathrm{p}_m(\phi)$ for both submodels, as well as the subposterior for the ICU ($m = 1$) submodel $\mathrm{p}_1(\phi \mid Y_1)$. The melded posterior will be largely influenced by the product of $\mathrm{p}_1(\phi \mid Y_1)$ and $\mathrm{p}_2(\phi)$, since $\mathrm{p}_1(\phi)$ is effectively uniform (see the centre panel of Fig. 4), and there are no data observed in the severity submodel, i.e. $Y_2 = \varnothing$. In stage one we target the ICU submodel posterior $\mathrm{p}_1(\phi, \psi_1 \mid Y_1)$, enabling the use of the original `JAGS` (Plummer 2018) implementation. These samples for ϕ are displayed in the right panel of Fig. 4, and we see that whilst there is substantial overlap with $\mathrm{p}_2(\phi)$ (left panel), $\mathrm{p}_1(\phi \mid Y_1)$ is more disperse, particularly for ϕ_1. Our region of interest is thus the HDR of $\mathrm{p}_1(\phi \mid Y_1)$, as the two-stage sampler involves evaluating the samples from $\mathrm{p}_1(\phi \mid Y_1)$ under $\mathrm{p}_2(\phi)$.

5.4 Self-density ratio estimation

The stage two acceptance probability for a move from $\phi \to \phi^*$ where $\phi^* \sim \mathrm{p}_1(\phi, \psi_1 \mid Y_1)$ is

$$\alpha(\phi^*, \phi) = \frac{\mathrm{p}_{\text{pool}}(\phi^*)\mathrm{p}_2(\phi^*, \psi_2, Y_2)\mathrm{p}_1(\phi)\mathrm{p}_2(\phi)}{\mathrm{p}_{\text{pool}}(\phi)\mathrm{p}_2(\phi, \psi_2, Y_2)\mathrm{p}_1(\phi^*)\mathrm{p}_2(\phi^*)}. \tag{12}$$

In both the severity and ICU submodels, the prior marginal distribution $\mathrm{p}_m(\phi)$ is unknown. This necessitates estimating the self-density ratio for both $\mathrm{p}_1(\phi)$ and $\mathrm{p}_2(\phi)$. However, the uniformity of $\mathrm{p}_1(\phi)$ corresponds to a self-density ratio that is effectively 1 everywhere. In contrast, the severity submodel prior marginal $\mathrm{p}_2(\phi)$ is clearly not uniform over our region of interest; appropriately estimating the melded posterior thus requires an accurate estimate of $\mathrm{p}_2(\phi) / \mathrm{p}_2(\phi^*)$.

To obtain the WSRE estimate of $\mathrm{p}_2(\phi) / \mathrm{p}_2(\phi^*)$ we first note that, in the notation of Section 3.4, $\phi = (\phi_1, \phi_2) = (\phi^{[1]}, \phi^{[2]})$, thus $D = 2$. We set $V = 10$, resulting in $W = 10^2$ weighting functions, and we draw 10^3 samples from each weighted target for a total of 10^5 MCMC samples. The Gaussian weighting functions for the first component $\mathrm{m}(\phi^{[1]}; \xi_{v,1})$ have the variance parameter fixed to 25^2; for the mean parameter we set $\xi_{1,1} = 30$, $\xi_{V,1} = 275$, with 8 other equally spaced values between these extrema. For $\mathrm{m}(\phi^{[2]}; \xi_{v,2})$ we set the variance parameter to 250^2, with $\xi_{1,2} = 500$, $\xi_{V,2} = 3000$, and interpolate 8 equally spaced values between the extrema. The values for the variance parameters are based on the empirical variance of the samples in Fig. 4.

5.5 Results

We compare the melded posterior estimate obtained using WSRE against the naive approach. For the latter we draw 10^5 samples from $\mathrm{p}_2(\phi)$ so that both approaches have the same number of samples, although the naive approach has a larger effective sample size. Figure 5 displays trace plots of 15 stage two MCMC chains, where $\alpha(\phi^*, \phi)$ is computed using the naive approach (left column), and using WSRE (right column). The erroneous behaviour displayed in the left column is due to underestimation of the tails of $\widehat{\mathrm{p}}_2(\phi)$ using a standard KDE. This underestimation results in an overestimation of the acceptance probability for proposals in the tails of $\widehat{\mathrm{p}}_2(\phi)$, since the proposal term $\widehat{\mathrm{p}}_2(\phi^*)$ is in the denominator of Eq. (12). Hence, moves to improbable values of ϕ^* have acceptance probabilities that are dominated by Monte Carlo error. Once at this improbable value the error then has the opposite effect; the underestimate yields chains unable to move back to probable values. This produces the monotonic step-like behaviour seen in the top left panel of Fig. 5. Although this behaviour is not visible in all 15 chains, it will eventually occur if the chains are run for more iterations, as a sufficiently improbable value for ϕ^* will be proposed. The results from this sampler are thus unstable.

Whilst there is no baseline "truth" to compare to in this example, the sampler that employs $\hat{\mathrm{r}}_{\text{WSRE}}(\phi, \phi^*)$ as an estimate of $\mathrm{p}_2(\phi) / \mathrm{p}_2(\phi^*)$ produces plausible results, in contrast to the naive approach. No step-like behaviour is visible when employing the WSRE approach (right column of 5). Whilst the between-chain mixing is not optimal, this can be ameliorated by running the chains for longer, which cannot be said for the naive method. This improved behaviour is obtained using the same number of samples from the prior marginal distribution, or weighted versions thereof. Users of this algorithm can be much more confident that the results are not artefactual.

6 Discussion

The complexity of many phenomena necessitates intricate, large models. Markov melding allows the practitioner to channel modelling efforts into smaller, simpler submodels, each of which may have data associated with it, then coherently combine these smaller models and disparate data. Multi-stage, sequential sampling methods, such as the

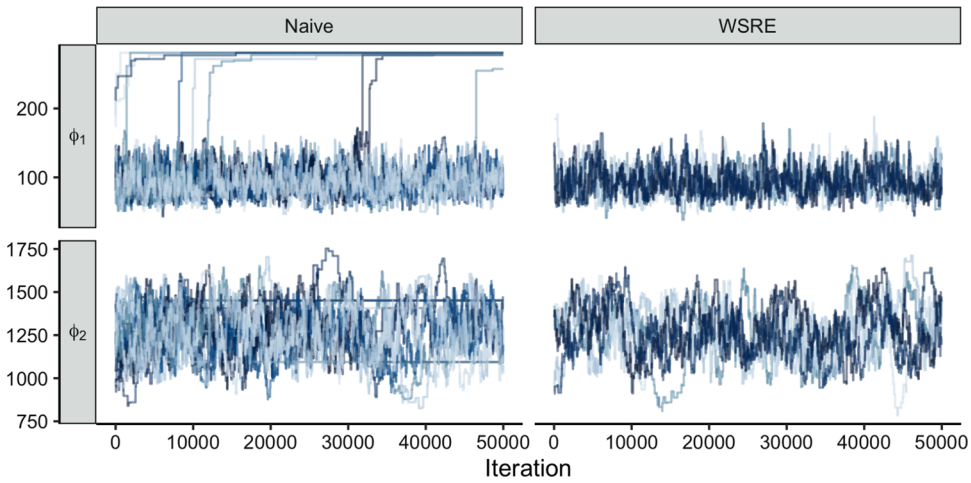

Fig. 5 Trace plots of 15 replicate stage two chains for ϕ_1 and ϕ_2, using the naive approach (left column) and the WSRE approach (right column)

sampler used for Markov melding, are important tools for estimating these models in a pragmatic, computationally feasible manner.

In particular, when an analytic form of the prior marginal distribution is not available, we have demonstrated that the two-stage sampling process is particularly sensitive to the corresponding KDE in regions of low probability. Tail probability estimation is an important and recurrent challenge in statistics (Hill 1975; Béranger et al. 2019). We addressed this issue in the Markov melding context by noting that we can limit our focus to the self-density ratio estimate, and sample weighted distributions to improve performance in low probability areas, for lower computational cost than simple Monte Carlo. Our examples show that for equivalent sample sizes, we improve the estimation of the melded posterior compared to the naive approach.

The issue addressed in this paper arises to due differences in the intermediary distributions of the two-stage sampling process, particularly where the proposal distribution is wider than the target distribution. The presence or absence of this issue is dependent upon the order in which the components of the melded model are considered in the sampling process, which is often constrained by the link function used to define ϕ in each model. In both our examples the link function is non-invertible. Goudie et al. (2019) show extensions of the link function that render it invertible are valid; that is, the model is theoretically invariant to the choice of extension. However, the practical performance of the two-stage sampler is heavily dependent on the appropriateness of such extensions, and designing such extensions is extremely challenging. Hence, the ordering of the submodels in the two-stage sampler is often predetermined; we are practically constrained by the non-invertible link function. In our examples this corresponds to sampling the less informative model for ϕ first. If we are free to choose the ordering of the two-stage sampler, we may still prefer to sample the wider model first, as the melded posterior is more likely to be captured in a reweighted sample from a wider distribution than such a sample from a narrow distribution. However, if the melded posterior distribution is substantially narrower than the stage-one target distribution then we are susceptible to the sample degeneracy and impoverishment problem (Li et al. 2014). Addressing this issue in the melding context, whilst retaining the computational advantages of the two-stage sampler, is an avenue for future work.

The examples we consider contain 1 or 2 dimensional ϕ. For higher dimensional ϕ we anticipate encountering issues associated with the curse of dimensionality. Specifically, the decrease in accuracy of any KDE and increase in the required number of weighting functions will scale exponentially with dimension. Applying the argument in Sect. 3.4 to locate these additional weighting functions will be challenging. As such we recommend WSRE, like other KDE methods, for settings where $\dim(\phi) \leq 5$ (Wand and Jones 1995). This requirement may be relaxed when there is structure in ϕ that allows it to be split into lower-dimensional components, such as when ϕ contains a collection of subject-specific parameters that are independent a priori. More generally, in high dimensions almost everywhere is a 'region of low probability' and the performance of KDEs is known to be poor, making choosing both an appropriate number of weighting functions and their parameters difficult. Machine learning methods have proven to be effective for estimating densities of moderate to high dimension (see Wang and Scott (2019) for a review), however the performance of these methods in low probability regions has not, to our knowledge, been thoroughly investigated.

There are potential alternatives to our weighted-sample self-density ratio estimation technique. Umbrella sampling (Torrie and Valleau 1977; Matthews et al. 2018) aims to accurately estimate the tails of a density $\mathrm{p}(\phi)$ by constructing an estimate $\widehat{\mathrm{p}}(\phi)$ from W sets of weighted samples $\{\phi_{n,w}\}_{n=1}^{N} \sim \mathrm{s}_w(\phi; \xi_w)$, However, umbrella sampling requires estimates of the normalising constants $Z_{2,w} = \int \mathrm{s}_w(\phi; \xi_w)\mathrm{d}\phi$ to combine the density estimates computed from each weighted

sample. Our approach is able to avoid computing normalising constants by focusing on the self-density ratio. Umbrella sampling also requires choosing the location of the weighting functions, i.e. choosing ξ_w appropriately. A heuristic strategy, similar to that of Sect. 3.4, is seen as necessary by Torrie and Valleau (1977). Adaptive procedures that automatically choose values of ξ_w based on other criteria exist, but these assume that $s_w(\phi)$ is a Gaussian distribution (Mitsuta et al. 2018) or operate on a predefined grid of possible values (Wojtas-Niziurski et al. 2013). We cannot use the generic tempering methodology advocated by Matthews et al. (2018), as sampling from $p(\phi, \gamma)^{1/\tau}$, for $\tau > 1$, does not generally produce marginal samples from $p(\phi)^{1/\tau}$.

Another possibility would be to sample p_{meld} using a pseudo-marginal approach (Andrieu and Roberts 2009). A necessary condition of the pseudo-marginal approach is that we possess an unbiased estimate of the target distribution. Kernel density estimation produces biased estimates of $p(\phi)$ for finite N. A KDE can be debiased (Calonico et al. 2018; Cheng and Chen 2019), but doing so requires substantial computational effort. Moreover, we also require an unbiased estimate of $1 / p(\phi)$. Debiasing estimates of $1 / p(\phi)$ is possible with pseudo-marginal methods like Russian roulette (Lyne et al. 2015), but Park and Haran (2018) observe prohibitive computational costs when doing so. The presence of both $p_{pool}(\phi)$ and $1 / p(\phi)$ in the melded posterior further complicates the production of an unbiased estimate, particularly when $p_{pool}(\phi)$ is formed via logarithmic pooling.

Acknowledgements We acknowledge Anne Presanis for many helpful discussions about this issue, and thank the referees for useful comments on an earlier version of this paper. This work was supported by The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/N510129/1] and the UK Medical Research Council [programme code MC_UU_00002/2].

References

Ades, A., Cliffe, S.: Markov chain Monte Carlo estimation of a multiparameter decision model: Consistency of evidence and the accurate assessment of uncertainty. Med. Decis. Making **22**(4), 359–371 (2002). https://doi.org/10.1177/027298902400448920

Ades, A.E., Sutton, A.J.: Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. J. R. Stat. Soc. A. Stat. Soc. **169**(1), 5–35 (2006). https://doi.org/10.1111/j.1467-985X.2005.00377.x

Albert, I., Espié, E., de Valk, H., Denis, J.B.: A Bayesian evidence synthesis for estimating campylobacteriosis prevalence. Risk Anal. **31**(7), 1141–1155 (2011). https://doi.org/10.1111/j.1539-6924.2010.01572.x

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **37**(2), 697–725 (2009). https://doi.org/10.1214/07-AOS574

Besbeas, P., Morgan, B.: Exact inference for integrated population modelling. Biometrics **75**(2), 475–484 (2019). https://doi.org/10.1111/biom.13045

Blomstedt, P., Mesquita, D., Lintusaari, J., Sivula, T., Corander, J., Kaski, S.: Meta-analysis of Bayesian analyses. arXiv e-prints arXiv:1904.04484 (2019)

Béranger, B., Duong, T., Perkins-Kirkpatrick, S.E., Sisson, S.A.: Tail density estimation for exploratory data analysis using kernel methods. Journal of Nonparametric Statistics **31**(1), 144–174 (2019). https://doi.org/10.1080/10485252.2018.1537442

Calonico, S., Cattaneo, M.D., Farrell, M.H.: On the effect of bias estimation on coverage accuracy in nonparametric inference. J. Am. Stat. Assoc. **113**(522), 767–779 (2018). https://doi.org/10.1080/01621459.2017.1285776

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1), 1–32 (2017). https://doi.org/10.18637/jss.v076.i01

Cheng, G., Chen, Y.C.: Nonparametric inference via bootstrapping the debiased estimator. Electron. J. Stat. **13**(1), 2194–2256 (2019). https://doi.org/10.1214/19-EJS1575

Clemen, R.T., Winkler, R.L.: Combining probability distributions from experts in risk analysis. Risk Anal. **19**(2), 187–203 (1999). https://doi.org/10.1111/j.1539-6924.1999.tb00399.x

Coley, R.Y., Fisher, A.J., Mamawala, M., Carter, H.B., Pienta, K.J., Zeger, S.L.: A Bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. Biometrics **73**(2), 625–634 (2017). https://doi.org/10.1111/biom.12577

De Angelis, D., Presanis, A.M., Conti, S., Ades, A.E.: Estimation of HIV Burden through Bayesian Evidence Synthesis. Stat. Sci. (2014). https://doi.org/10.1214/13-STS428

Eddelbuettel, D., François, R.: Rcpp: Seamless R and C++ integration. Journal of Statistical Software **40**(8), 1–18 (2011). https://doi.org/10.18637/jss.v040.i08

Goudie, R.J.B., Presanis, A.M., Lunn, D., De Angelis, D., Wernisch, L.: Joining and splitting models with Markov melding. Bayesian Anal. **14**(1), 81–109 (2019). https://doi.org/10.1214/18-BA1104

Hemelaar, J.: The origin and diversity of the HIV-1 pandemic. Trends Mol. Med. **18**(3), 182–192 (2012). https://doi.org/10.1016/j.molmed.2011.12.001

Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**(5), 1163–1174 (1975)

Hiraoka, K., Hamada, T., Hori, G.: Estimators for unnormalized statistical models based on self density ratio. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4523–4527 (2014). https://doi.org/10.1109/ICASSP.2014.6854458

Hooten, M.B., Johnson, D.S., Brost, B.M.: Making recursive Bayesian inference accessible. Am. Stat. (2019). https://doi.org/10.1080/00031305.2019.1665584

Johnson, D.S., Brost, B.M., Hooten, M.B.: Greater than the sum of its parts: Computationally flexible Bayesian hierarchical modeling. arXiv:2010.12568 (2020)

Jones, M.C.: Kernel density estimation for length biased data. Biometrika **78**(3), 511–519 (1991)
Kedem, B., De Oliveira, V., Sverchkov, M.: Statistical data fusion. World Sci. (2017). https://doi.org/10.1142/10282
Koekemoer, G., Swanepoel, J.W.: Transformation kernel density estimation with applications. J. Comput. Graph. Stat. **17**(3), 750–769 (2008)
Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. Bioinformatics **20**(16), 2626–2635 (2004). https://doi.org/10.1093/bioinformatics/bth294
Li, T., Sun, S., Sattar, T.P., Corchado, J.M.: Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. Expert Syst. Appl. **41**(8), 3944–3954 (2014). https://doi.org/10.1016/j.eswa.2013.12.031
Lunn, D., Barrett, J., Sweeting, M., Thompson, S.: Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. J. R. Stat. Soc. Ser. C **62**(4), 551–572 (2013). https://doi.org/10.1111/rssc.12007
Lyne, A.M., Girolami, M., Atchadé, Y., Strathmann, H., Simpson, D.: On russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. Stat. Sci. **30**(4), 443–467 (2015). https://doi.org/10.1214/15-STS523
Matthews, C., Weare, J., Kravtsov, A., Jennings, E.: Umbrella sampling: A powerful method to sample tails of distributions. Mon. Not. R. Astron. Soc. **480**, 4069–4079 (2018). https://doi.org/10.1093/mnras/sty2140
Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., Rizopoulos, D.: Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. Stat. Comput. **30**(4), 999–1014 (2020). https://doi.org/10.1007/s11222-020-09927-9
Mitsuta, Y., Kawakami, T., Okumura, M., Yamanaka, S.: Automated exploration of free energy landscapes based on umbrella integration. Int. J. Mol. Sci. **19**(4), 937 (2018). https://doi.org/10.3390/ijms19040937
Nakayama, M.K.: Asymptotic properties of kernel density estimators when applying importance sampling. In: Proceedings of the 2011 Winter Simulation Conference (WSC), pp. 556–568 (2011). https://doi.org/10.1109/WSC.2011.6147785
O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T.: Uncertain judgements: Eliciting experts' probabilities. Stat. Pract. (2006). https://doi.org/10.1002/0470033312
Park, J., Haran, M.: Bayesian inference in the presence of intractable normalizing functions. J. Am. Stat. Assoc. **113**(523), 1372–1390 (2018). https://doi.org/10.1080/01621459.2018.1448824
Plummer, M.: Rjags: Bayesian graphical models using MCMC (2018). R package version 4-8
Presanis, A.M., Pebody, R.G., Birrell, P.J., Tom, B.D.M., Green, H.K., Durnall, H., Fleming, D., De Angelis, D.: Synthesising evidence to estimate pandemic (2009) A/H1N1 influenza severity in 2009–2011. Annal. Appl. Stat. **8**(4), 2378–2403 (2014). https://doi.org/10.1214/14-AOAS775
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021)
Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Statistics in Practice. Wiley, Chichester ; Hoboken, NJ (2004)
Sutton, A.J., Abrams, K.R.: Bayesian methods in meta-analysis and evidence synthesis. Stat. Methods Med. Res. **10**(4), 277–303 (2001). https://doi.org/10.1177/096228020101000404
Tom, J.A., Sinsheimer, J.S., Suchard, M.A.: Reuse, recycle, reweigh: Combating influenza through efficient sequential Bayesian computation for massive data. Annal. Appl. Stat. **4**(4), 1722–1748 (2010). https://doi.org/10.1214/10-AOAS349
Torrie, G., Valleau, J.: Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J. Comput. Phys. **23**(2), 187–199 (1977). https://doi.org/10.1016/0021-9991(77)90121-8
Vardi, Y.: Empirical distributions in selection bias models. Ann. Stat. **13**(1), 178–203 (1985). https://doi.org/10.1214/aos/1176346585
Wand, M., Jones, M.: Kernel Smoothing. Chapman and Hall/CRC (1995)
Wang, Z., Scott, D.W.: Nonparametric density estimation for high-dimensional data-Algorithms and applications. WIREs Comput. Stat. **11**(4), e1461 (2019). https://doi.org/10.1002/wics.1461
Wojtas-Niziurski, W., Meng, Y., Roux, B., Bernèche, S.: Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. J. Chem. Theory Comput. **9**(4), 1885–1895 (2013). https://doi.org/10.1021/ct300978b