Statistics and Computing manuscript No.
(will be inserted by the editor)

Unbiased approximation of posteriors via coupled particle Markov chain Monte Carlo

Willem van den Boom · **Ajay Jasra** · **Maria De Iorio** · **Alexandros Beskos** · **Johan G. Eriksson**

Received: date / Accepted: date

Abstract Markov chain Monte Carlo (MCMC) is a powerful methodology for the approximation of posterior distributions. However, the iterative nature of MCMC does not naturally facilitate its use with modern highly parallel computation on HPC and cloud environments. Another concern is the identification of the bias and Monte Carlo error of produced averages. The above have prompted the recent development of fully ('embarrassingly') parallel unbiased Monte Carlo methodology based on coupling of MCMC algorithms. A caveat is that formulation of effective coupling is typically not trivial and requires model-specific technical effort. We propose coupling of MCMC chains deriving from sequential Monte Carlo (SMC) by considering adaptive SMC methods in combination with recent advances in unbiased estimation for state-space models. Coupling is then achieved at the SMC level and is, in principle, not problem-specific. The resulting methodology enjoys desirable theoretical properties. A central motivation is to extend unbiased MCMC to more challenging targets compared to the ones typically considered in the relevant literature. We illustrate the effectiveness of the algorithm via application to two complex statistical models: (i) horseshoe regression; (ii) Gaussian graphical models.

Willem van den Boom · Maria De Iorio · Johan G. Eriksson
National University of Singapore, Yong Loo Lin School of Medicine
E-mail: vandenboom@nus.edu.sg

Willem van den Boom · Maria De Iorio · Johan G. Eriksson
Agency for Science, Technology and Research, Singapore Institute for Clinical Sciences

Ajay Jasra
King Abdullah University of Science and Technology, Computer, Electrical and Mathematical Sciences and Engineering division, Thuwal, Saudi Arabia

Maria De Iorio · Alexandros Beskos
University College London, Department of Statistical Science, UK

Keywords Adaptive sequential Monte Carlo · Coupling · Embarrassingly parallel computing · Gaussian graphical model · Particle filter · Unbiased MCMC

Declarations

Funding

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier 2 (grant number MOE2019-T2-2-100) and the Singapore National Research Foundation under its Translational and Clinical Research Flagship Programme and administered by the Singapore Ministry of Health's National Medical Research Council (grant number NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014). Additional funding is provided by the Singapore Institute for Clinical Sciences, Agency for Science, Technology and Research.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare that relate to the content of this article.

Availability of data and material

The data are confidential human subject data, thus are not available.

Code availability

The scripts that produced the empirical results are available on `https://github.com/willemvandenboom/cpmcmc`.

1 Introduction

MCMC is a powerful methodology for the approximation of complex distributions. MCMC is intrinsically iterative and, while asymptotically unbiased, the size of the bias and the Monte Carlo error of generated estimates given a finite number of iterations are often difficult to quantify. Moreover, MCMC will typically not allow full exploitation of the computational potential of modern distributed-computing techniques. Recently, Jacob et al. (2020b) propose a method for unbiased MCMC estimation based on coupling of Markov chains, building on ideas by Glynn and Rhee (2014). The algorithm is embarrassingly parallel, and the unbiasedness provides immediate quantification of the Monte Carlo error. However, devising effective coupling for MCMC algorithms targeting a given posterior can be highly challenging. See e.g. the construction of coupled MCMC for a horseshoe regression model in Biswas et al. (2021) and the development for posteriors based on Hamiltonian Monte Carlo (HMC) in Heng and Jacob (2019).

Considering the scope for unbiased MCMC and its current limitations, we propose a coupled MCMC algorithm where the coupling mechanism is not in principle specific to the posterior at hand, resulting in a general recipe for unbiased MCMC. We do so by working on an appropriate augmented space. Specifically, we build on recent advances for unbiased estimation in state-space models (Middleton et al. 2019; Jacob et al. 2020a) which devise coupling strategies through particle filters independently of the shape of the target distribution. Our work extends these ideas by embedding a general posterior distribution in a state-space model using adaptive SMC (Del Moral et al. 2006). This results in a methodology that broadens the class of posteriors that can be treated via unbiased MCMC, by exploiting the coupling methods feasible within the SMC framework. As our methodology couples Markov chains that consist of particle MCMC methods, we refer for convenience to our method as 'coupled particle MCMC'. To illustrate the potential of our methodology, we apply coupled particle MCMC to (i) Gaussian graphical models, which are substantially more complex than models considered previously in the literature on unbiased MCMC, mainly due to discreteness and dimension of graph space; (ii) horseshoe regression.

The MCMC step resulting from our state space embedding combines particle MCMC methods by Middleton et al. (2019) and Jacob et al. (2020a), namely coupled particle independent Metropolis-Hastings (PIMH) and conditional SMC, respectively. The focus of these works is unbiased approximation for the smoothing distribution of a state-space model; the high-dimensionality therein relates to the number of time steps, whereas the state space is implicitly assumed to be low-dimensional. We consider more complex state spaces of a much higher dimension than Middleton et al. (2019) and Jacob et al. (2020a), provide effective adaptation of the SMC, and consider both PIMH and conditional SMC to improve mixing (as compared to using only PIMH) of the MCMC. We empirically investigate the trade-off between mixing and coupling via a number of numerical experiments.

The structure of the paper is as follows. Section 2 reviews unbiased MCMC estimation based on coupled Markov chains. Section 3 introduces coupled particle MCMC for unbiased estimation, for the case of a general posterior. Section 4 contains theoretical results for the meeting time of the coupled Markov chains and the resulting unbiased estimator deduced from the literature on coupled conditional SMC for state-space models (Lee et al. 2020). Section 5 applies the proposed general methodology to simulated data, including a case with horseshoe regression, where a comparison with the MCMC coupling approach in Biswas et al. (2021) is carried out. Section 6 considers Gaussian graphical models as an example where effective coupling of MCMC is highly non-trivial. We finish with a discussion in Section 7.

2 Unbiased MCMC with couplings

Before we introduce coupled particle MCMC, we review coupled MCMC and introduce notation. Denote the parameter space by $\mathscr{X} \subseteq \mathbb{R}^{d_x}$, $d_x \geq 1$. Let $x \in \mathscr{X}$ denote the parameter, $y \in \mathscr{Y} \subseteq \mathbb{R}^{d_y}$, $d_y \geq 1$, the data, and $\pi(x)$ the density of the posterior of interest w.r.t. some dominating measure. The unbiased construction requires a pair of coupled ergodic Markov chains $\{x(t-1), \bar{x}(t)\}$, $t \geq 1$, on $\mathscr{X} \times \mathscr{X}$ with both chains having $\pi(x)$ as equilibrium distribution. The coupling is such that $x(t-1)$ and $\bar{x}(t)$ have the same distribution for any $t \geq 1$, and the two chains meet at some random time $\tau < \infty$ a.s. so that $x(t) = \bar{x}(t)$ for $t \geq \tau$. Under standard conditions, the posterior expectation $\pi(h) = \int_{\mathscr{X}} h(x)\,\pi(x)\,dx$ of a statistic $h : \mathscr{X} \mapsto \mathbb{R}$ can be obtained as $\pi(h) = \lim_{t\to\infty} E\left[h\{x(t)\}\right]$, with all expectations assumed finite. Writing the limit as a telescoping sum and using that fact that $x(t-1), \bar{x}(t)$ admit the same law, gives for any fixed $k \geq 0$ (Glynn and Rhee 2014)

$$
\begin{aligned}
\pi(h) &= E[h\{x(k)\}] + \sum_{t=k+1}^{\infty} \left(E[h\{x(t)\}] - E[h\{x(t-1)\}]\right) \\
&= E[h\{x(k)\}] + \sum_{t=k+1}^{\infty} \left(E[h\{x(t)\}] - E[h\{\bar{x}(t)\}]\right) \\
&= E\left(h\{x(k)\} + \sum_{t=k+1}^{\tau-1} \left[h\{x(t)\} - h\{\bar{x}(t)\}\right]\right).
\end{aligned}
$$

We assume that the technical conditions that permit the exchange of summation and expectation in the last step hold in

our setting. Thus, the quantity

$$\hat{h}_k = h\{x(k)\} + \sum_{t=k+1}^{\tau-1} [h\{x(t)\} - h\{\bar{x}(t)\}] \quad (1)$$

is an unbiased estimator of $\pi(h)$. Process $\{\bar{x}(t)\}$ is typically initialised at $\bar{x}(1) = x(0) \sim p_0$, for a law p_0. Both chains will evolve marginally according to the same MCMC chain with the posterior $\pi(x)$ as invariant distribution, with a coupling applied for the joint transition $[x(t), \bar{x}(t) \mid x(t-1), \bar{x}(t-1)]$, $t \geq 2$. We adopt this setting for the rest of the paper.

Coupled particle MCMC requires some further notation. Denote the prior density on the parameter by $p(x)$. Let $p(y \mid x)$ be the density of the data y given x. The posterior density follows from Bayes' rule as $\pi(x) \propto p(x)\, p(y \mid x)$. For any $\alpha \in [0,1]$, we denote by $\pi_\alpha(x)$ the tempered posterior proportional to $p(x)\, p(y \mid x)^\alpha$. Thus, $\pi_0(x) = p(x)$ and $\pi_1(x) = \pi(x)$. All densities are assumed to be determined w.r.t. appropriate reference measures. As our method involves adaptive SMC, we assume we can sample x from its prior $p(x)$ and evaluate the likelihood $p(y \mid x)$. The method requires, for any $\alpha \in (0,1]$, the construction of an MCMC step with $\pi_\alpha(x)$ as its invariant distribution. We refer to this step as the 'inner' MCMC step as it is a part of a more involved 'outer' particle MCMC step. For integers $i \leq j$, denote the range $\{i, \ldots, j\}$ by $i{:}j$. We use the colon notation for collections of variables, i.e., $x_{i:j} = \{x_i, \ldots, x_j\}$ and $x^{i:j} = \{x^i, \ldots, x^j\}$.

3 Coupled particle MCMC

3.1 Feynman-Kac model & SMC sampler

The proposed unbiased estimation procedure builds on SMC. As a first step, we 'adapt' the SMC algorithm to the prior $p(x)$ and likelihood $p(y \mid x)$ under consideration. This is a preliminary step that determines the tempering constants and the inner MCMC steps, thus also the target Feynman-Kac model. By exploiting adaptive SMC, we are able to obtain a flexible and general construction (i.e. applicable to a large class of posterior distributions) for unbiased posterior approximation.

The adaptation produces a sequence of $S \geq 0$ temperatures, $0 < \alpha_1 < \cdots < \alpha_S < 1$, corresponding to bridging distributions $\pi_{\alpha_s}(x)$, $s = 1, \ldots, S$. Here, S and temperatures α_s, $s = 1, \ldots, S$, are chosen from this initial application of SMC with $N_0 \geq 1$ particles, so that importance weights meet an effective sample size (ESS) threshold as, e.g., in Jasra et al. (2010). The adaptation also determines the number m_s, $s = 1, \ldots, S$, of iterations of an 'inner' MCMC kernel $p_{\alpha_s}(dx' \mid x)$ that preserves $\pi_{\alpha_s}(x)$. More in details, for each s, we choose m_s via a criterion that requires sufficiently reduced sample correlation, for particles pre- and post-application of MCMC steps, over given scalar statistics of interest, $f_j : \mathscr{X} \to \mathbb{R}$, $j = 1, \ldots$. See Sections 5 and 6.4 for examples of such statistics for specific models, which in both cases include the log-likelihood $f_1(x) = \log\{p(y \mid x)\}$. Concatenating m_s inner MCMC steps increases diversity among the particles $x_s^{1:N_0}$ to avoid weight degeneracy. Kantas et al. (2014) consider similar adaptation of m_s. Algorithm 1 summarizes the adaptive procedure where $\widehat{\mathrm{corr}}(\cdot,\cdot)$ in Step 2c denotes the sample correlation.

Algorithm 1 Adaptation of the Feynman-Kac model.

Input: Number of particles N_0, ESS and correlation thresholds $\gamma_0, \zeta_0 \in [0,1]$.

1. Sample particles $x_0^{1:N_0}$ independently from $p(x)$. Set $s = 1$, $\alpha_0 = 0$.
2. Repeat while $\alpha_{s-1} < 1$:
   (a) Compute weights $w_s^i(\alpha) = p(y \mid x_{s-1}^i)^{\alpha - \alpha_{s-1}}$, $i = 1, \ldots, N_0$, and find
   $$\alpha_s = \min\left\{ \alpha \in (\alpha_{s-1}, 1] : 1/\sum_{i=1}^{N_0} \{w_s^i(\alpha)\}^2 \leq \gamma_0 N_0 \right\}.$$
   (b) Obtain $x_s^{1:N_0}$ by sampling with replacement from $x_{s-1}^{1:N_0}$ with probabilities proportional to $w_s^{1:N_0}(\alpha_s)$.
   (c) Let $x_{s,k}^{1:N_0}$ denote the position of particles $x_s^{1:N_0}$ after applying $k \geq 1$ MCMC transitions $p_{\alpha_s}(dx' \mid x)$, on each particle. Find
   $$m_s = \min_{k \geq 1}\Big\{ \max_j \big[\widehat{\mathrm{corr}}\{f_j(x_s^{1:N_0}), f_j(x_{s,k}^{1:N_0})\}\big] \leq \zeta_0 \Big\}.$$
   With an abuse of notation, let $x_s^{1:N_0}$ now be the particles after the application of m_s MCMC steps.
   (d) Set $s = s + 1$.

Output: Temperatures $0 < \alpha_1 < \cdots < \alpha_S < 1$, number of MCMC steps $m_{1:S}$.

Algorithm 2 SMC sampler.

Input: Number of particles N, temperatures $0 = \alpha_0 < \cdots < \alpha_S < 1$, number of MCMC steps $m_{1:S}$.

1. Sample N particles $x_0^{1:N}$ independently from $\pi_{\alpha_0}(x) = p(x)$.
2. For $s = 1, \ldots, S$:
   (a) Compute weights $w_s^i = p(y \mid x_{s-1}^i)^{\alpha_s - \alpha_{s-1}}$, $i = 1, \ldots, N$.
   (b) Determine $x_s^{1:N}$ by sampling with replacement from $x_{s-1}^{1:N}$ with probabilities proportional to $w_s^{1:N}$.
   (c) For each particle in $x_s^{1:N}$, carry out m_s MCMC transitions $p_{\alpha_s}(dx' \mid x)$. With an abuse of notation, let $x_s^{1:N}$ be the particles after application of the MCMC steps.
3. (a) Compute weights $w^i = p(y \mid x_S^i)^{1-\alpha_S}$, $i = 1, \ldots, N$.
   (b) Determine $x^{1:N}$ by sampling with replacement from $x_S^{1:N}$ with probabilities proportional to $w^{1:N}$.

Output: Set of particles $x^{1:N}$ that approximate the posterior $\pi(x)$.

Algorithm 1 determines the Feynman-Kac model, i.e. a distribution $\Pi(x_{0:S})$ with marginal $x_S \sim \pi(x)$. Then, Algorithm 2 describes a standard SMC sampler applied to such model corresponding to a bootstrap particle filter. Steps 2b and 3b of Algorithm 2 describe multinomial resampling. For

our empirical results, we replace multinomial with systematic resampling, as the latter reduces variability (Chopin and Papaspiliopoulos 2020, Section 9.7) and yields better mixing for the outer MCMC steps. Algorithm 7, in Appendix A, describes systematic resampling. Use of adaptive resampling (Chopin and Papaspiliopoulos 2020, Section 10.2) further reduces the variability of Monte Carlo estimates in our empirical results. That is, we only resample (Step 2b) if the ESS of the current weighted particle approximation falls below $N\gamma$ for a $\gamma \in [0,1]$.

Step 3b of Algorithm 2 is not required to approximate the posterior $\pi(x)$ as the pair $(w^{1:N}, x_S^{1:N})$ provides a weighted approximation. We include the step since conditional SMC will involve it. Section 3.3 discusses a Rao-Blackwellization approach enabling the use of the weighted approximation within coupled particle MCMC.

Algorithms 1 and 2 use different numbers of particles, namely N_0 and N, respectively. We choose N_0 to be larger than N as the the number N_0 is used once, 'off-line', and results provided by this single run of Algorithms 1 will fix all aspects of the Feynman-Kac model used by coupled particle MCMC. Choice of a large enough N_0 aims to facilitate stability for the obtained collection of temperatures and number of inner MCMC steps.

3.2 Coupling of particle MCMC

Having specified an SMC sampler, we derive a coupled particle MCMC algorithm. The outer MCMC step is constructed via the SMC sampler. Specifically, we borrow ideas from the particle filtering literature and define a coupling strategy of the outer MCMC step, the latter defined as a mixture of the coupled PIMH in Middleton et al. (2019, Algorithm 3) and the coupled conditional particle filters in Jacob et al. (2020a, Algorithm 2). PIMH and conditional SMC (Andrieu et al. 2010, Section 2.4) provide MCMC steps on the extended state space $x_{0:S} \in \mathscr{X}^{S+1}$ based on Algorithm 2. The invariant law on $\mathscr{X}^{S+1}$ of such MCMC has $\pi(x)$ as marginal distribution on $x_S \in \mathscr{X}$ (Andrieu et al. 2010, Theorems 2, 5). Therefore, the resulting MCMC algorithm can be run for two coupled chains to provide unbiased Monte Carlo estimation per (1). The added machinery of SMC provides more ways to couple the MCMC compared to using a less elaborate MCMC algorithm for sampling from the posterior $\pi(x)$.

Coupled PIMH results in smaller meeting times τ and worse mixing of the MCMC chain than conditional SMC in our experiments. The meeting times and the MCMC mixing both affect the variance of the resulting unbiased estimators. The empirical results show that neither PIMH nor conditional SMC yields universally better performance (across different posteriors). We thus also consider a mixture of them as the outer MCMC step. We mention here that in our experiments we do not encounter scenarios where such mixture yields notably more computationally efficient unbiased estimation than both PIMH and conditional SMC. Still, incorporation of the mixture allows for gaining further insight into the presented methods. Next, we describe coupled PIMH and coupled conditional SMC separately.

Algorithm 3 Coupled PIMH step.

Input: Current states $x_{0:S}$ and $\bar{x}_{0:S}$, along with the corresponding SMC estimates Z and $\bar{Z}$ of the marginal likelihood $p(y)$.

1. Sample $\tilde{x}_{0:S}$ and $\tilde{Z}$ using Algorithm 2 as the proposal for both chains:
   (a) Set $\tilde{x}_S$ equal to x^1 from the output of Algorithm 2.
   (b) For $s = S-1,\dots,0$, set $\tilde{x}_s$ equal to the element in $x_s^{1:N}$ which generated $\tilde{x}_{s+1}$ per Step 2b of Algorithm 2. In other words, $\tilde{x}_s$ is the ancestor of $\tilde{x}_{s+1}$.
   (c) Compute the corresponding marginal likelihood estimate $\tilde{Z}$ from the weights in Algorithm 2 as detailed in Del Moral et al. (2006, Section 3.2.1).
2. Perform two coupled Metropolis-Hastings accept-reject steps:
   (a) Sample $U \sim \mathscr{U}(0,1)$.
   (b) If $U < \tilde{Z}/Z$, then set $x_{0:S} = \tilde{x}_{0:S}$ and $Z = \tilde{Z}$.
   (c) If $U < \tilde{Z}/\bar{Z}$, then set $\bar{x}_{0:S} = \tilde{x}_{0:S}$ and $\bar{Z} = \tilde{Z}$.

Output: Updated states $x_{0:S}$ and $\bar{x}_{0:S}$, along with Z and $\bar{Z}$, obtained by application of a coupled PIMH kernel, with transitions that marginally preserve $\Pi(x_{0:S})$.

Coupled PIMH

Algorithm 3 details the coupled PIMH update which builds on the SMC sampler in Algorithm 2. SMC provides an unbiased estimate Z of the marginal likelihood $p(y)$ which enables the use of an independent Metropolis-Hastings step for two chains. The two steps are coupled by using the same proposal and the same uniform random variable U in the accept-reject step across both chains. Then, both chains meet as soon as they both accept the proposal in Step 2 at the same MCMC iteration. See Middleton et al. (2019) for a more elaborate introduction of coupled PIMH.

Coupled conditional SMC

Algorithm 4 details the coupled conditional SMC update which, like Algorithm 3, builds on the SMC sampler in Algorithm 2. The resampling of the particles in Steps 4b and 5b forms the main source of coupling across the two transition kernels applied on $x_{0:S}$ and $\bar{x}_{0:S}$. See Jacob et al. (2020a) for a more elaborate introduction of an algorithm closely related to Algorithm 4, namely the coupled conditional particle filter.

As in Jacob et al. (2020a) and Lee et al. (2020), we consider Algorithm 5 for the coupled resampling. Earlier applications of Algorithm 5 can be found in Chopin and Singh (2015) in the context of theoretical analysis of conditional SMC and in Jasra et al. (2017) within the setting

Algorithm 4 Coupled conditional SMC step.

Input: Current states $x_{0:S}$ and $\bar{x}_{0:S}$.

1. Set $x_s^1 = x_s$ and $\bar{x}_s^1 = \bar{x}_s$ for $s = 0, \ldots, S$.
2. Sample $N-1$ particles $x_0^{2:N}$ independently from $\pi_{\alpha_0}(x) = p(x)$.
3. Set $\bar{x}_0^i = x_0^i$ for $i = 2, \ldots, N$.
4. For $s = 1, \ldots, S$:
   (a) Compute the (unnormalised) weights $w_s^i = p(y \mid x_{s-1}^i)^{\alpha_s - \alpha_{s-1}}$, and $\bar{w}_s^i = p(y \mid \bar{x}_{s-1}^i)^{\alpha_s - \alpha_{s-1}}$, $i = 1, \ldots, N$.
   (b) Determine $x_s^{2:N}$ and $\bar{x}_s^{2:N}$ by coupled resampling (Algorithm 5) with replacement from $x_{s-1}^{1:N}$ and $\bar{x}_{s-1}^{1:N}$ with probabilities proportional to $w_s^{1:N}$ and $\bar{w}_s^{1:N}$, respectively.
   (c) Update x_s^i and $\bar{x}_s^i$ by applying m_s coupled inner MCMC steps that preserve (marginally) $\pi_{\alpha_s}(x)$ for $i = 2, \ldots, N$.
   (d) Form trajectory $x_{0:s}^i$ by joining the ancestors of x_s^i, $i = 1, \ldots, N$.
5. (a) Compute the weights $w^i = p(y \mid x_S^i)^{1-\alpha_S}$, $\bar{w}^i = p(y \mid \bar{x}_S^i)^{1-\alpha_S}$, $i = 1, \ldots, N$.
   (b) Determine $x_{0:S}$ and $\bar{x}_{0:S}$ by coupled resampling (Algorithm 5) with replacement from $\{x_{0:S}^i\}_{i=1}^N$ and $\{\bar{x}_{0:S}^i\}_{i=1}^N$ with probabilities proportional to $w^{1:N}$ and $\bar{w}^{1:N}$, respectively.
   (c) Compute the corresponding marginal likelihood estimates Z and $\bar{Z}$ from the weights $w_{1:S}^{1:N}$ and $\bar{w}_{1:S}^{1:N}$, respectively, as detailed in Del Moral et al. (2006, Section 3.2.1).

Output: Updated states $x_{0:S}$ and $\bar{x}_{0:S}$, along with Z and $\bar{Z}$, obtained by application of a coupled conditional SMC kernel, with transitions that marginally preserve $\Pi(x_{0:S})$.

Algorithm 5 Coupled resampling.

Input: Probability vectors $p_{1:N}$ and $\bar{p}_{1:N}$.

1. Compute $p_i^{\min} = \min(p_i, \bar{p}_i)$ for $i = 1, \ldots, N$.
2. (a) With probability $a = \sum_{i=1}^N p_i^{\min}$:
      i. Draw i according to the law on $1{:}N$ defined by the probability vector $p_{1:N}^{\min}/a$.
      ii. Let $\bar{i} = i$.
   (b) With probability $1-a$, draw i and $\bar{i}$ according to the laws defined by the probability vectors $(p_{1:N} - p_{1:N}^{\min})/(1-a)$ and $(\bar{p}_{1:N} - p_{1:N}^{\min})/(1-a)$, respectively.

Output: Samples i and $\bar{i}$ which are distributed according to $p_{1:N}$ and $\bar{p}_{1:N}$, respectively, and for which the probability of $i = \bar{i}$ is maximized.

of multilevel particle filtering. The algorithm samples from two discrete distributions such that the resulting two indices are equal with maximum probability (Jasra et al. 2017, Section 3.2).

We use systematic resampling (see Algorithm 7, in Appendix A) for our empirical results. Thus, we require a coupling for it. Chopin and Singh (2015) derive a modification of systematic resampling for use with conditional SMC. We propose a coupling of this conditional systematic resampling for Step 4b of Algorithm 4. Appendix A gives the details for these algorithms.

Step 4c of Algorithm 4 involves a coupling of the inner MCMC update for $\pi_{\alpha_s}(x)$ across two chains. The coupling should at least be 'faithful' (Rosenthal 1997), i.e., sustain any meeting of the chains. That is, if $x_s^i = \bar{x}_s^i$ initially, then that should still be true after the coupled inner MCMC updates of x_s^i and $\bar{x}_s^i$. A simple way to attain this minimal requirement is by using the same seed for the pseudorandom number generators in both chains. Better coupling of the inner MCMC updates could further improve the overall coupling of our method.

Algorithm 6 Coupled particle MCMC.

Input: Minimum number of outer MCMC steps l and probability ρ of using PIMH.

1. Initialize $x_{0:S}(0)$ by running the SMC algorithm as per Step 1 of Algorithm 3.
2. (a) With probability ρ, generate $x_{0:S}(1) \mid x_{0:S}(0)$ according to the PIMH algorithm, for instance by running Algorithm 3 with $x_{0:S} = \bar{x}_{0:S} = x_{0:S}(0)$, and set $\bar{x}_{0:S}(1)$ equal to the proposal $\tilde{x}_{0:S}$ from this PIMH.
   (b) With probability $1-\rho$, set $\bar{x}_{0:S}(1) = x_{0:S}(0)$ and generate $x_{0:S}(1) \mid x_{0:S}(0)$ according to the conditional SMC algorithm, for instance by running Algorithm 4 with $x_{0:S} = \bar{x}_{0:S} = x_{0:S}(0)$.
   (c) Set $t = 2$.
3. While $x_{0:S}(t-1) \neq \bar{x}_{0:S}(t-1)$ or $t \leq l$:
   (a) Draw $x_{0:S}(t) \mid x_{0:S}(t-1)$ and $\bar{x}_{0:S}(t) \mid \bar{x}_{0:S}(t-1)$ using the coupled PIMH step in Algorithm 3 with probability ρ, and the coupled conditional SMC step in Algorithm 4 with probability $1-\rho$.
   (b) Set $t = t+1$.

Output: Chains $\{x_{0:S}(t)\}_{t=0}^T$ and $\{\bar{x}_{0:S}(t)\}_{t=1}^T$, that meet at a random time $\tau \geq 1$, with $T = \max\{l, \tau\}$.

3.3 Unbiased Monte Carlo approximation

Algorithm 6 specifies the coupled MCMC algorithm resulting from a mixture of Algorithms 3 and 4 with parameter ρ. It provides chains $\{x(t)\}_{t=1}^T$ and $\{\bar{x}(t)\}_{t=1}^T$ which can be used to estimate expectations w.r.t. the posterior $\pi(x)$ via ergodic averages. By construction, $x(t-1)$ and $\bar{x}(t)$ have the same distribution for any $t \geq 1$. Also, they meet at some time τ which is almost surely finite under the conditions given in in Section 4 for $\rho = 0, 1$. This enables unbiased Monte Carlo estimation of $\pi(h)$ as described in (1).

Middleton et al. (2019, Section 2.2) propose the initialization in Step 2a of Algorithm 6. This choice allows for $\tau = 1$ as $x_{0:S}(1) = \bar{x}_{0:S}(1)$ if the PIMH step accepts. Moreover, $\Pr(\tau = 1) \geq 1/2$ if only PIMH is used, i.e. $\rho = 1$ (Middleton et al. 2019, Proposition 8). Note that a high $\Pr(\tau = 1)$ does not necessarily result in low variance Monte Carlo estimation. For instance, consider $\bar{x}(1) = x(0)$ and let $x(1) \mid x(0)$ follow a Metropolis-Hastings update. Then, we have that $\Pr(\tau = 1) = \Pr\{\bar{x}(1) = x(1)\}$, i.e. $\Pr(\tau = 1)$ corresponds to the Metropolis-Hastings rejection probability which will typically not be too low. However, the performance of the unbiased methodology based on coupling of such an MCMC kernel can be very poor.

The unbiasedness of the estimator $\hat{h}_k$ in (1) enables embarrassingly parallel computation for independent estimates. Consider R independent runs of Algorithm 6 resulting in

R independent copies of the estimator denoted by $\hat{h}_k^r$, $r = 1,\dots,R$. Then, $R^{-1}\sum_r^R \hat{h}_k^r$ is an unbiased estimator of $\pi(h)$. Its variance decreases linearly in R and can be estimated by its empirical variance. Moreover, the unbiasedness and independence of the R estimators enables the construction of confidence intervals for $\pi(h)$ which are exact as $R \to \infty$ per the central limit theorem.

In the context of particle filtering, Jacob et al. (2020a, Section 4) provide improvements to the unbiased estimator $\hat{h}_k$ in (1) that reduce its variance for each run of Algorithm 6. We apply these improvements to our setting for general posterior computation. Firstly, one can average over different k since $\hat{h}_k$ is unbiased for any $k \geq 1$. For any positive integers k and l with $k \leq l$, this results in the unbiased estimator

$$\begin{aligned}\bar{h}_k^l &= \frac{1}{l-k+1}\sum_{q=k}^{l}\hat{h}_q \\ &= \frac{1}{l-k+1}\sum_{q=k}^{l}\left(h\{x(q)\} + \sum_{t=q+1}^{\tau-1}[h\{x(t)\} - h\{\bar{x}(t)\}]\right) \\ &= \frac{1}{l-k+1}\sum_{t=k}^{l} h\{x(t)\} \\ &\qquad + \sum_{t=k+1}^{\tau-1}\frac{\min(l-k+1,t-k)}{l-k+1}[h\{x(t)\} - h\{\bar{x}(t)\}],\end{aligned} \tag{2}$$

where the penultimate quantity is an ergodic average and the last quantity is a bias correction term. The ergodic average term $(l-k+1)^{-1}\sum_{t=k}^{l} h\{x(t)\}$ discards the first $k-1$ steps in the chain $\{x(t)\}_{t=1}^{T}$ as burn-in iterations and uses $l-k+1$ recorded iterations.

As k increases, the variance of the bias correction decreases, and the bias correction equals zero for $k \geq \tau - 1$. Nonetheless, it is suboptimal to set k very large as that increases the variance of the ergodic average similar to when discarding too many iterations as burn-in in MCMC. One can pick k as a high percentile of the empirical distribution of the meeting time τ from multiple runs of Algorithm 6. Alternatively, the empirical variance of $\bar{h}_k^l$ can be minimized via grid search over k (Middleton et al. 2019, Appendix B.2), at the price of losing the unbiasedness of $\bar{h}_k^l$. Parameter l can be set to a large value within computational constraints as a larger l reduces the variance of the unbiased estimator in (2).

A second straightforward approach for variance reduction involves Rao-Blackwellization of the estimator $\hat{h}_k$ over the weighted particle approximation from SMC. So far, we have considered only the single particle selected by Algorithm 3 or Step 5b of Algorithm 4. It is more efficient to use all N particles via the weighted approximations defined by the pairs $(w^{1:N}, x_S^{1:N})$, $(\bar{w}^{1:N}, \bar{x}_S^{1:N})$. That is, $h\{x(t)\}$, $h\{\bar{x}(t)\}$ in (1), (2) can be replaced by $\sum_{i=1}^N w^i(t)\,h\{x_S^i(t)\}/\sum_{i=1}^N w^i(t)$ and $\sum_{i=1}^N \bar{w}^i(t)\,h\{\bar{x}_S^i(t)\}/\sum_{i=1}^N \bar{w}^i(t)$, respectively, where the weights and particles are given in Step 3b of Algorithm 2 or Step 5b of Algorithm 4.

4 Theoretical properties

Existing analysis of coupled conditional SMC applies to our context which aims to approximate the posterior $\pi(x)$. Building on Lee et al. (2020), we derive results for Algorithm 6 with $\rho = 0$, such that only conditional SMC is used. Appendix B contains the proofs which mostly consist of a mapping from the smoothing problem considered in Lee et al. (2020) to our context of posterior approximation. That mapping results in the following assumptions.

Assumption 1 The likelihood is bounded. That is, $\sup_{x\in\mathscr{X}} p(y \mid x) < \infty$ for the observed data y.

Assumption 2 The statistic $h\colon \mathscr{X} \to \mathbb{R}$, as in (1), is bounded. That is, $\sup_{x\in\mathscr{X}} |h(x)| < \infty$.

Many models satisfy Assumption 1, including the Gaussian graphical model in Section 6. The likelihood from linear regression in Section 5.2 violates it if the number of predictors is greater than the number of observations. The assumption relates to the boundedness of potential functions in the SMC literature, which is a common assumption (Del Moral 2004, Section 2.4.1). Andrieu et al. (2018, Theorem 1) show that Assumption 1 is essentially equivalent to uniform ergodicity of the Markov chains produced by conditional SMC in Algorithm 6 with $\rho = 0$.

In the context of particle filtering, Jacob et al. (2020a, Section 3) establish unbiasedness and finite variance of the estimator $\hat{h}_k$ in (1), like we do, but without the boundedness in Assumption 2 and instead use an assumption jointly on $h(x)$ and the Markov chain generated by conditional SMC. The simpler and more restrictive Assumption 2 provides a bound on the variance of $\hat{h}_k$ in terms of the number of particles N in Proposition 2.

Assumptions 1 and 2 imply Assumptions 6 and 4 of Middleton et al. (2019), respectively. Therefore, the estimator $\hat{h}_k$ from Algorithm 8 with $\rho = 1$, such that it uses only PIMH, is unbiased and has finite variance, and $\tau < \infty$ almost surely per Proposition 8 of Middleton et al. (2019) and its proof under Assumptions 1 and 2.

Proposition 1 *Suppose Assumption 1 holds. Consider Algorithm 6 with $\rho = 0$. Then, for any number of temperatures $S \geq 0$, there exists a $c < \infty$ such that for any number of particles $N \geq 2$ and any initial $(x_{0:S}, \bar{x}_{0:S})$, we have:*

(i) $\Pr(x'_{0:S} = \bar{x}'_{0:S}) \geq N/(N+c)$ *where $(x'_{0:S}, \bar{x}'_{0:S})$ are distributed per coupled conditional SMC in Algorithm 4.*
(ii) $\tau < \infty$ *almost surely.*
(iii) *The average meeting time $E(\tau) \leq 2 + c/N$.*

Proposition 1 contains no conditions on the quality of the coupling of the inner MCMC steps in Step 4c of Algorithm 4. This confirms that the SMC machinery by itself enables coupling and thus unbiased estimation for the posterior $\pi(x)$, though coupling of the inner MCMC steps can lead to substantially smaller meeting times as explored empirically in Appendix D, an aspect of the method that the above theoretical results fail to capture. Moreover, the meeting time τ can be made equal to 2 with arbitrarily high probability by increasing the number of particles N.

Whether a sufficient increase of N is practicable depends on the SMC sampler in Algorithm 2 which we adapt to $\pi(x)$. For instance, under additional assumptions, Theorem 9 from Lee et al. (2020) states that the meeting probability $\Pr(x'_{0:S} = \bar{x}'_{0:S})$ does not vanish if the number of particles scales as $N = \mathcal{O}(2^S S)$ where S is the number of temperatures.

Proposition 2 *Suppose Assumptions 1 and 2 hold. Consider the estimator $\hat{h}_k$ of $\pi(h)$ in (1) where the chains $x(t)$ and $\bar{x}(t)$ are generated by Algorithm 6 with $\rho = 0$. Denote the expectation and variance w.r.t. the resulting distribution on $\{x(t), \bar{x}(t)\}_{t=1}^{T}$ by $\hat{E}(\cdot)$ and $\hat{\mathrm{var}}(\cdot)$, respectively. Then, for any number of temperatures $S \geq 0$, we have:*

(i) *For any $k \geq 1$, $\hat{E}(\hat{h}_k) = \pi(h)$ and $\hat{\mathrm{var}}(\hat{h}_k) < \infty$.*
(ii) *There exists a $c < \infty$ such that for any $k \geq 1$ and for any number of particles $N \geq 2$ in Algorithm 4,*

$$|\hat{\mathrm{var}}(\hat{h}_k) - \mathrm{var}\{h(x)\}| \leq 16 \left(\frac{N+c}{N}\right)^2 \left(\frac{c}{N+c}\right)^{k/2} \sup_{x \in \mathscr{X}} |h(x) - \pi(h)|.$$

Proposition 2 confirms the unbiasedness of $\hat{h}_k$. Moreover, the variance penalty resulting from the bias correction sum in (1) vanishes as $N \to \infty$ or $k \to \infty$. The latter confirms the role of k as the number of burn-in iterations in the MCMC estimation. The constant c in Proposition 2 is the same as in Proposition 1. The rates of convergence implied by the upper bounds in Propositions 1 and 2 are probably conservative as noted by Lee et al. (2020).

5 Simulation studies

In all our applications, Algorithm 1 sets up a Feynman-Kac model via a preliminary run that uses $N_0 = 10^4$ particles, an ESS threshold of $\gamma_0 = 0.8$ to determine $\alpha_{0:S}$ and a correlation threshold of $\zeta_0 = 0.95$ for $m_{0:S}$. Subsequent executions of SMC algorithms use adaptive resampling with ESS threshold $\gamma = 0.5$, i.e. we resample, (Step 2b of Algorithm 2) only if ESS drops below γN. All empirical results use Rao-Blackwellization when computing $\bar{h}_k^l$.

To evaluate algorithmic performance for a test function h, we consider the product '$\hat{\mathrm{var}}(\bar{h}_k^l) \times$ time' of the empirical variance of $\bar{h}_k^l$ and the average computation time to obtain one $\bar{h}_k^l$ across the number, R, of independent runs. This product captures the trade-off between quantity and quality of unbiased estimates: the variance of the average of independently and identically distributed estimates $\bar{h}_k^l$ is proportional to $\hat{\mathrm{var}}(\bar{h}_k^l)$ and inversely proportional to the number of estimates. The number of estimates is in turn inversely proportional to the average computation time when working with a fixed computational budget. Thus, '$\hat{\mathrm{var}}(\bar{h}_k^l) \times$ time' is proportional to the variance of the final estimator for fixed computation time. We compute '$\hat{\mathrm{var}}(\bar{h}_k^l) \times$ time' for $l = 1, \ldots, l_{\max}$, for some $l_{\max}$, where $k \leq l$ is set for each l such that $\hat{\mathrm{var}}(\bar{h}_k^l)$ is minimized.

The mixing of the outer MCMC and the meeting times both influence $\mathrm{var}(\bar{h}_k^l)$ with larger meeting times and worse mixing corresponding to higher variance, though the manner in which such effects combine is non-trivial. Thus, we present mixing and meeting times in our empirical results for further insight. Mixing is monitored via calculation of the integrated autocorrelation time for each run based on the chain $\sum_{i=1}^N w^i(t)\, h\{x_S^i(t)\} / \sum_{i=1}^N w^i(t)$, $t = l/2, \ldots, l$, computed via the `R` package `LaplacesDemon` (Statisticat, LLC. 2020).

5.1 Mixture of Gaussians

The first set of results are motivated by the model and SMC considered in Section B.2 of Middleton et al. (2019). The likelihood is

$$p(y \mid x) = \prod_{i=1}^{d_y} \frac{1}{d_x} \sum_{j=1}^{d_x} \mathcal{N}(y_i \mid x_j, 1)$$

and the prior is uniform over the hypercube $[-10, 10]^{d_x}$. We consider $d_x = 2$ and $d_y = 100$, and simulate data y from $p(y \mid x)$ under true values $x^* = (-3, 0)^\top$. Thus, the posterior $\pi(x) \propto p(x)\, p(y \mid x)$ is multimodal. The inner MCMC step is a random walk Metropolis, with proposed transition following a Gaussian distribution with identity covariance, as in Middleton et al. (2019). We couple the inner MCMC across two chains by using the same seed for the pseudorandom number generator in each of the two MCMC updates, resulting in a common random number coupling.

The scalar statistics used to determine $m_{1:S}$ are the log-likelihood $f_1(x) = \log\{p(y \mid x)\}$ and L_2-norm $f_2(x) = \|x\|$. We aim to estimate $\pi(h)$ for $h(x) = x_1 + x_2 + x_1^2 + x_2^2$. We run Algorithm 6 for $R = 1024$ times for each $\rho = 0, 1/2, 1$, with a maximum value $l_{\max} = 10^4$ and $N = 25$ particles.

Figure 1 shows that meeting times are lowest for Algorithm 6 with PIMH while mixing of the outer MCMC step is best with conditional SMC. A mixed set-up with $\rho = 1/2$, aimed at a trade-off between these extremes, has coupling and mixing performance in between $\rho = 0$ and $\rho = 1$. Nonetheless, criterion '$\hat{\mathrm{var}}(\bar{h}_k^l) \times$ time', defined in

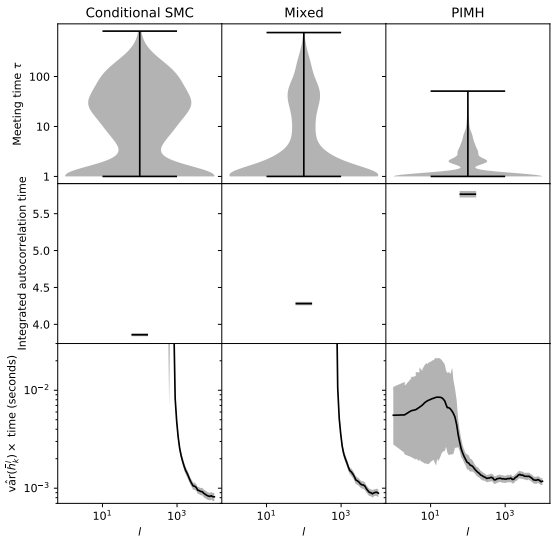


Fig. 1 Results from execution of Algorithm 6 with $\rho = 0$ (conditional SMC), $\rho = 1/2$ (mixed) and $\rho = 1$ (PIMH), for the case of the mixture of Gaussians. The top row contains violin plots of $\log(\tau)$. The bottom two rows show the integrated autocorrelation time and $\hat{\text{var}}(\bar{h}_k^l) \times$ time as a function of l, with their means and 95% bootstrapped confidence intervals in black and gray, respectively.

the second paragraph of Section 5, is slightly lower for $\rho = 1/2$ than for conditional SMC only for values of l where PIMH is superior, a pattern observed also for scenarios described in Appendix D. The criterion is lowest for conditional SMC owing to its better mixing but only if Algorithm 6 is run for sufficiently more iterations than the largest meeting times. Finally, we compare our methodology with the coupled HMC of Heng and Jacob (2019) in Appendix C showing the coupled particle MCMC is more efficient in the scenario considered, though it must be noted that HMC is not well suited for multimodal targets.

5.2 Horseshoe regression

Background

Biswas et al. (2021) recently proposed a coupled MCMC for horseshoe regression (Carvalho et al. 2009). This section compares the coupling and its resulting unbiased estimation with the proposed coupled particle MCMC in Algorithm 6.

We consider the standard likelihood for linear regression $p(y \mid x) = \mathcal{N}(y \mid W\beta, \sigma^2 I_{d_y})$ with y a d_y-dimensional vector of observations, W a $d_y \times p$ design matrix and σ^2 the error variance. The main object of inference is the p-dimensional coefficient vector β. The horseshoe prior on β is one of the most popular global-local shrinkage priors for state-of-the-art Bayesian variable selection (Bhadra et al. 2019). It is defined by $\beta_j \mid \sigma^2, \xi, \eta_j \sim \mathcal{N}\{0, \sigma^2/(\xi\eta_j)\}$ independently for $j = 1, \ldots, p$ where ξ is a global precision parameter with prior $\sqrt{\xi} \sim C^+(0,1)$ and η_j is a local precision parameter with prior $\sqrt{\eta_j} \sim C^+(0,1)$ independently for $j = 1, \ldots, p$. Here, $C^+(0,1)$ denotes the standard half-Cauchy distribution. That is, if $t \sim C^+(0,1)$, then the density of t is $p(t) = 2/\{\pi(1+t^2)\}$ for $t > 0$ and zero otherwise.

We follow the simulation set-up of Biswas et al. (2021, Section 3). The gamma prior $1/\sigma^2 \sim \text{Gamma}(1,1)$ completes the Bayesian model specification. The elements of the matrix W are sampled independently from the standard Gaussian distribution. We draw $y \sim \mathcal{N}(W\beta_*, \sigma_*^2 I_n)$ for some true β_* and true error variance σ_*^2. We set $\sigma_*^2 = 8$, $d_y = 100$, $p = 20$, $\beta_{*,j} = 2^{(9-j)/4}$ for $j = 1, \ldots, 10$ and $\beta_{*,j} = 0$ for $j = 11, \ldots, p$.

Coupled particle MCMC for horseshoe regression

We construct our inner MCMC steps by making use of Algorithm 1 from Biswas et al. (2021). Here, $x = (\beta, \eta, \sigma^2, \xi)$ since the MCMC provides a Markov chain on these parameters jointly. Our method requires an inner MCMC step that is invariant w.r.t. $\pi_\alpha(x) \propto p(x)\, p(y \mid x)^\alpha$, e.g., Step 2c of Algorithm 2. Here, $p(y \mid x)^\alpha$ equals $p(y \mid x)$ with n, y and W replaced by αn, $\sqrt{\alpha}y$ and $\sqrt{\alpha}W$, respectively. Thus, the required inner MCMC for $\pi_\alpha(x)$ follows by carrying out these replacements in the inner MCMC step as developed for the full posterior $\pi(x)$.

The scalar statistics used for determining the number of MCMC steps $m_{1:S}$ are the log-likelihood function, that is $f_1(x) = \log\{p(y \mid x)\}$, and the number of elements in β with absolute value greater than 0.01, $f_2(x) = \sum_{j=1}^p \text{I}[|\beta_j| > 0.01]$. The latter is interpreted as the number of variables chosen by the model. Algorithm 1 produces $m_s = 1$ for all s. This confirms the effectiveness of the inner MCMC algorithm borrowed from Biswas et al. (2021), which is accompanied by favourable theoretical properties as per Biswas et al. (2021, Section 2.2). We compare the estimation of the posterior expectation of $h(x) = \beta_j + \beta_j^2$, $j = 11$, obtained from different algorithms.

Results

We compare the results from coupled particle MCMC (Algorithm 6) with $\rho = 1$ and $\rho = 0.9$ using $N = 25$ particles to the two-scale coupling of Biswas et al. (2021, Section 3.2) which provides a direct coupling of an MCMC algorithm that we use as the inner MCMC step. Algorithm 6 with $\rho = 0.9$ uses this same coupling of the inner MCMC when calling Algorithm 4. We do not present the results for smaller values of ρ as they are not competitive due to too large meeting times. The three coupled MCMCs are run $R = 128$ times

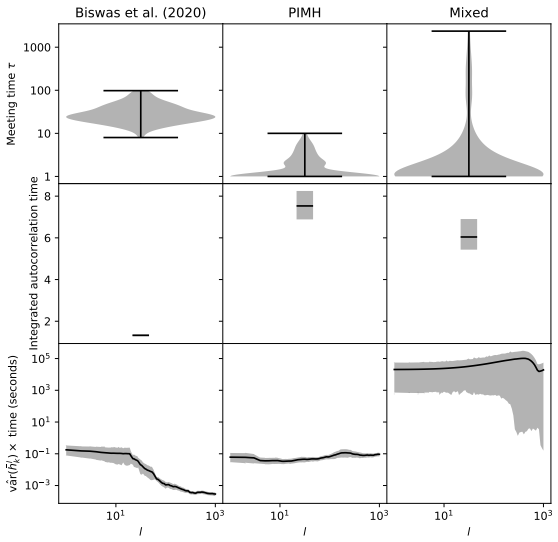


Fig. 2 Results from the coupled MCMC from Biswas et al. (2021) (left), and Algorithm 6 with $\rho = 1$ (PIMH) and $\rho = 0.9$ (mixed) for the horseshoe regression simulation. The top row contains violin plots of $\log(\tau)$. The bottom two rows show the integrated autocorrelation time and $\hat{\text{var}}(\bar{h}_k^l) \times$ time as a function of l with their means and 95% bootstrapped confidence intervals in black and gray, respectively.

each for $l_{\max} = 10^3$ iterations. Figure 2 presents the results analogously to Figure 1.

Coupled particle MCMC meets in fewer iterations than the coupled MCMC from Biswas et al. (2021). Nonetheless, coupled particle MCMC takes more computation time to achieve the same estimation accuracy. One reason for this is the mixing of the coupled MCMC as shown in Figure 2 with a much higher integrated autocorrelation time for coupled particle MCMC. Another reason is that the computational cost of the outer particle MCMC step in Algorithm 6 is much higher than that of a single inner MCMC step. This increased cost of particle methods could be alleviated by parallelizing across particles if the main concern is elapsed real time rather than cumulative CPU time across CPU cores. More generally, effective direct couplings of (inner) MCMC, like the method from Biswas et al. (2021), would often be superior if available, as any improved coupling through SMC methods does not justify their added cost. On the other hand, coupled particle MCMC can be useful when effective direct coupling is not available. The results for '$\hat{\text{var}}(\bar{h}_k^l) \times$ time' show that it is not sufficient to only consider meeting times when assessing coupled MCMC for unbiased estimation if the MCMC algorithms compared differ in mixing or computational cost.

Figure 2 suggests that conditional SMC mixes better than PIMH but results in a worse coupling of the outer MCMC. Here, the improvement over PIMH in mixing does not outweigh the worse coupling of conditional SMC and the associated variance of the bias correction term at $l = 10^3$ outer MCMC iterations. For l sufficiently large, the relative contribution of the bias correction term would vanish resulting in lower '$\hat{\text{var}}(\bar{h}_k^l) \times$ time' for conditional SMC than for PIMH, though such l might not be computationally feasible.

6 Application: Gaussian graphical models

6.1 Model

We consider Gaussian graphical models (Dempster 1972; Lauritzen 1996) as an example of a complex posterior on discrete spaces. Coupling of MCMC kernels on the original non-extended space appears highly challenging. In general, generation of unbiased estimators for posterior expectations in this context is a major undertaking, and it is attempted, to the best of our knowledge, for the first time in this work. We allow the graphs to be non-decomposable, and present methodology that sidesteps approximation of intractable normalising constants.

The object of inference is an undirected graph $G = (V, E)$ defined by a set of nodes $V = \{1, \dots, p\}$, $p \geq 1$, and a set of edges $E \subset V \times V$. As in Lenkoski (2013), we slightly abuse notation and write $(i, j) \in G$ for $(i, j) \in E$, i.e. when vertices i and j are connected in G. We aim to infer G given an $n \times p$ data matrix Y with $n \geq 1$ independent rows distributed according to $\mathcal{N}(0, K^{-1})$ for a precision matrix $K \in M^+$, where M^+ is the set of $p \times p$ symmetric, positive-definite matrices. Graph G constrains K in that $K_{ij} = 0$ if $(i, j) \notin G$. Thus, $K \in M^+(G)$, where $M^+(G) \subset M^+$ is the cone of matrices $K \in M^+$ with $K_{ij} = 0$ for $(i, j) \notin G$.

Under the notation in Section 3, we have $x = (K, G)$, $y = Y$. Notice that we are required to include K into x as only then $p(y \mid x)$ is analytically available. We get that $p(y \mid x)$ is equal to $p(Y \mid K, G) = (2\pi)^{-np/2} |K|^{n/2} \exp(-\frac{1}{2} \langle K, U \rangle)$, where $U = Y^\top Y$ is the scatter matrix and $\langle K, U \rangle = \text{tr}(K^\top U)$ the trace inner product. A conjugate prior for K conditional on G is the G-Wishart distribution $\mathcal{W}_G(\delta, D)$ (Roverato 2002) with density

$$p(K \mid G) = \tfrac{1}{I_G(\delta, D)} |K|^{\delta/2 - 1} \exp\left(-\tfrac{1}{2} \langle K, D \rangle\right), \quad K \in M^+(G),$$

parametrised by $\delta > 2$ and a positive-definite rate matrix D. The normalising constant $I_G(\delta, D)$ does not have an analytical form for general non-decomposable G (Uhler et al. 2018). We henceforth choose $\delta = 3$ in agreement with previous work (Jones et al. 2005; Lenkoski 2013; Tan et al. 2017). Due to conjugacy, $K \mid G, Y \sim \mathcal{W}_G(\delta + n, D^*)$ where $D^* = D + U$. Since the objective is inference on G, one would like

to compute the posterior

$$\begin{aligned} p(G \mid Y) &\propto p(G)\, p(Y \mid G) \\ &= p(G) \int_{M^+(G)} p(Y \mid K, G)\, p(K \mid G)\, dK \\ &\propto p(G)\, \frac{I_G(\delta + n, D^*)}{I_G(\delta, D)}, \end{aligned} \tag{3}$$

with $p(G)$ the prior on G.

As the normalising constant I_G is not analytically available in general, some approximation is required. Standard approaches apply Monte Carlo or Laplace approximation of $p(G \mid Y)$ (Atay-Kayis and Massam 2005; Tan et al. 2017). Recent work avoids approximation of normalising constants via careful MCMC constructions. See Wang and Li (2012); Cheng and Lenkoski (2012); Lenkoski (2013); Hinne et al. (2014). Our methodology falls in this latter direction of research. Compared to Wang and Li (2012); Cheng and Lenkoski (2012), our MCMC samples directly from G-Wishart laws, rather than preserving them, to improve mixing. Our MCMC sampler is very similar to the one in Lenkoski (2013), with the exception of sampling elements of the precision matrices directly from the full conditional distribution, rather that applying a Metropolis-within-Gibbs step.

We use a size-based prior (Armstrong et al. 2009, Section 2.4) for G. That is, $p(G)$ is the induced marginal of a joint prior distribution on G and the number of edges $n_e = |E|$ such that the prior on G given n_e is uniform, $p(G \mid n_e) \propto 1$. For n_e, we choose a truncated geometric distribution with success probability $1/(p+1)$, i.e. $p(n_e) \propto \{p/(p+1)\}^{n_e}$, $n_e = 0, 1, \ldots, p(p-1)/2$. This choice completes the prior specification for G and induces sparsity. The success probability is such that the mean of the non-truncated geometric law equals the number of vertices p. Prior elicitation based on the sparsity constraint that the prior mean of n_e equals p has also been used with independent-edge priors (Jones et al. 2005, Section 2.4).

6.2 Further algorithmic specifications

We derive an inner MCMC step in Appendix E to apply Algorithm 6 to Gaussian graphical models. Here, we further detail how we use Algorithm 6.

Rejection sampling with $\alpha_0 > 0$

We set $\alpha_0 > 0$ and resort to rejection sampling with the prior $p(x)$ as proposal to sample from $\pi_{\alpha_0}(x)$ in Step 1 of Algorithm 2. This reduces the computational cost for the overall SMC algorithm considerably in our applications, due to the prior $p(G)$ being diffuse: the inner MCMC step (Algorithm 10) changes at most one edge of the graph in each step such that a large number of steps m_1 are required for effective diversification of particles for small α_1 for which $\pi_{\alpha_1}(x) \approx p(x)$. This implementation of the m_1 MCMC steps then dominates the computational cost. Setting $\alpha_0 > 0$ induces a larger α_1 and a much smaller m_1.

The maximum likelihood estimate of the precision K is $\arg\max_{K \in M^+} p(Y \mid K) = nU^{-1}$. So, the acceptance probability of the rejection sampler follows as

$$\begin{aligned} \frac{p(y \mid x_0^i)^{\alpha_0}}{\sup_{x \in \mathscr{X}} p(y \mid x)^{\alpha_0}} &= \frac{p(Y \mid K_0^i)^{\alpha_0}}{\sup_{K \in M^+} p(Y \mid K)^{\alpha_0}} \\ &= (|K_0^i|\,|U|)^{n\alpha_0/2} \exp\left[\tfrac{\alpha_0}{2}\left\{np(1 - \log n) - \left\langle K_0^i, U \right\rangle\right\}\right]. \end{aligned}$$

We choose α_0 to achieve an acceptance rate of $a = 1/50$. Specifically, we draw N_0/a particles from the prior $p(x)$ and set α_0 to the largest value such that N_0 of these particles are accepted.

PIMH vs conditional SMC

The main challenge for the class of Gaussian graphical models is the effective coupling of the outer MCMC chain so that the size of the meeting times τ is not too large. Following the discussion in Section 5, PIMH yields smaller τ than conditional SMC. We therefore use Algorithm 6 with $\rho = 1$ here. Even so, some chains fail to couple (Section 6.4), suggesting that lower values of ρ would result in impracticably large meeting times.

6.3 Data description

We apply the Gaussian graphical models on a set of metabonomic data. The sample consists of $n = 471$ six-year-old children from the Growing Up in Singapore Towards healthy Outcomes (GUSTO) study (Soh et al. 2014). Measurements are taken from blood serum samples via nuclear magnetic resonance to obtain metabolic phenotypes. Specifically, the levels of certain metabolites, which are low-molecular-weight molecules involved in metabolism, are measured. Examples of metabolites are cholesterol, fatty acids and glucose. See Soininen et al. (2009) for a description of the measurement process. We focus on a set of 35 metabolites. Thus, the $n \times p$ data matrix Y consists of $p = 35$ metabolite levels. We quantile normalise the metabolites such that they marginally follow a standard Gaussian distribution.

6.4 Results

The statistics used to determine the number of inner MCMC steps m_s are the log-likelihood $f_1(x) = \log\{p(y \mid x)\}$ and the number of edges $f_2(x) = |E|$. The adaptation results in $S = 165$ temperatures with m_s varying from 1 to 14 and a mean of 6 for $s > 1$, whereas $m_1 = 15$. Indeed, having $\alpha_0 > 0$

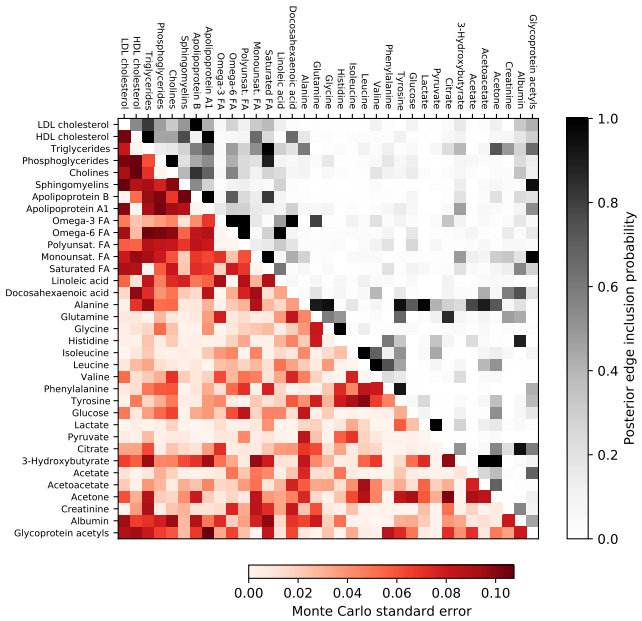


Fig. 3 Posterior edge inclusion probabilities (upper triangle) and their Monte Carlo standard errors (lower triangle) resulting from coupled particle MCMC for the Gaussian graphical model fit on the metabolite data. LDL, HDL and FA stand for low-density lipoprotein, high-density lipoprotein and fatty acids, respectively.

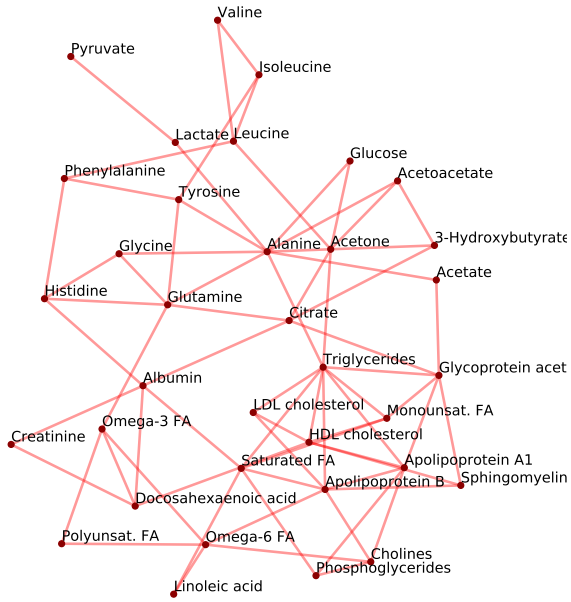


Fig. 4 The median probability graph resulting from coupled particle MCMC for the Gaussian graphical model fit on the metabolite data. LDL, HDL and FA stand for low-density lipoprotein, high-density lipoprotein and fatty acids, respectively.

ensures that the tempered posterior $\pi_{\alpha_1}(x)$ is further away from the uninformative prior $p(x)$ such that the number of MCMC steps for sufficient diversification of particles is limited.

Algorithm 6 is run for $l = 40$ MCMC steps with $N = 10^3$ particles, until the chains meet or when the time budget is up, whichever happens first. We do this repeatedly across 16 CPU cores simultaneously and independently. Our time budget is 72 hours. This results in 30 finished runs of Algorithm 6 and 16 incomplete runs that are cut short due to running out of time. The failure to meet suggests that this posterior is at the limit of what is computationally feasible with the version and numerical execution of coupled particle MCMC used in this work. We discard the incomplete runs which introduces a bias, though this bias decays exponentially with the time budget as discussed in Section 3.3 of Jacob et al. (2020b). Of the 30 finished runs, 19 meet in $\tau = 1$ step and the largest observed τ is 10.

We use these runs to obtain multiple unbiased estimates based on (2) with $k = 7$ and $l = 40$ where $h(x)$ is chosen to estimate the posterior edge inclusion probabilities, i.e. the posterior probability of an edge being included in the graph G. Figure 3 shows the estimates. Since they result from averaging independent and unbiased estimators, we can also estimate their Monte Carlo error. Figure 3 confirms that the chosen time budget is sufficient to achieve a usable posterior approximation with the standard errors maxing out at 0.11. Appendix G compares these probabilities with estimates from running SMC once with a large number of particles. Figure 4 contains the median probability graph, i.e. the graph which includes edges with a posterior inclusion probability greater than 0.5.

7 Discussion

In this work we propose a coupled particle MCMC methodology that can yield effective unbiased posterior approximation and enjoys attractive theoretical properties. Gaussian graphical models provide an example where a good coupling of the MCMC is hard while coupled particle MCMC produces Markov chains that tend to couple quickly. A major contribution of this manuscript is to embed recent developments in the field into a general procedure, building a complete and flexible computational strategy within the area of unbiased MCMC. Moreover, we investigate the performance of the proposed coupled particle MCMC in challenging settings. The paper places itself within an active and recent area of research, and our numerical applications (horseshoe regression, Gaussian graphical models) constitute far more complex examples than those typically considered in the literature on unbiased estimation in state-space models. Furthermore, it is applicable, in an automatic way (in principle), to a broad class of posterior distributions, while existing literature often focuses on particular classes of targets.

The empirical results show that using PIMH instead of conditional SMC in Algorithm 6, as controlled by the pa-

rameter ρ, provides a trade-off between coupling and MCMC mixing with conditional SMC corresponding with better mixing. Thus, conditional SMC results in the most efficient unbiased estimates for number of iterations l sufficiently large, though computing them might be impracticable due to too large meeting times. Then, PIMH provides a computationally more feasible alternative. Whether conditional SMC is feasible can be determined by preliminary and potentially incomplete runs of Algorithm 6. Generally, conditional SMC results in fewer replicate unbiased estimates, which might prevent optimal use of available computing resources (e.g. parallelisation) as well as lead to less reliable estimates of Monte Carlo error. We additionally consider a mixture of PIMH and conditional SMC. We do not find it to be empirically superior in the scenarios considered and existing theoretical results do not immediately apply. Such mixture is similar in spirit to the use of both a coupled Metropolis-Hastings step and a coupled HMC step in Heng and Jacob (2019) where that mixture also provides a trade-off between coupling and MCMC mixing.

Major empirical and theoretical improvements in coupling have been reported for the ancestral and backward sampling modifications to the standard version of conditional SMC (Chopin and Singh 2015; Jacob et al. 2020a; Lee et al. 2020). For instance, Lee et al. (2020, Theorem 11) show that the one-step meeting probability analysed in Proposition 1 does not vanish if $N = \mathcal{O}(S)$ for conditional SMC with backward sampling. Unfortunately, these modifications can usually not be used in our scenario as the transition densities in the Feynman-Kac model resulting from taking m_s inner MCMC steps are often intractable.

The use of SMC adds computational cost compared to running MCMC by itself. For instance, a single step in the outer MCMC defined by PIMH or conditional SMC involves $N\sum_{s=1}^{S} m_s$ inner MCMC steps. This extra computational cost is amenable to parallelization due to the embarrassingly parallel nature of the particles in SMC. Moreover, the extra computational effort results in more accurate Monte Carlo estimates due to Rao-Blackwellization across the N particles. One of the main contributions of this work is to gain further insights into these novel computational strategies. For instance, the numerical experiments in Section 5 indicate that the shape of the posterior at hand determines the efficiency in the performance of SMC-based coupling against alternative approaches.

Acknowledgements We thank the referees for many useful suggestions that helped to greatly improve the content of the paper.

References

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(3):269–342

Andrieu C, Lee A, Vihola M (2018) Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. Bernoulli 24(2):842–872

Armstrong H, Carter CK, Wong KFK, Kohn R (2009) Bayesian covariance matrix estimation using a mixture of decomposable graphical models. Statistics and Computing 19(3):303–316

Atay-Kayis A, Massam H (2005) A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. Biometrika 92(2):317–335

Bhadra A, Datta J, Polson NG, Willard B (2019) Lasso meets horseshoe: A survey. Statistical Science 34(3):405–427

Biswas N, Bhattacharya A, Jacob PE, Johndrow JE (2021) Coupled Markov chain Monte Carlo for high-dimensional regression with Half-t priors. arXiv:2012.04798v2

Carvalho CM, Polson NG, Scott JG (2009) Handling sparsity via the horseshoe. In: van Dyk D, Welling M (eds) Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, Proceedings of Machine Learning Research, vol 5, pp 73–80

Cheng Y, Lenkoski A (2012) Hierarchical Gaussian graphical models: Beyond reversible jump. Electronic Journal of Statistics 6:2309–2331

Chopin N, Papaspiliopoulos O (2020) An Introduction to Sequential Monte Carlo. Springer International Publishing

Chopin N, Singh SS (2015) On particle Gibbs sampling. Bernoulli 21(3):1855–1883

Del Moral P (2004) Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Springer, New York

Del Moral P, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(3):411–436

Dempster AP (1972) Covariance selection. Biometrics 28(1):157

Dobra A, Lenkoski A, Rodriguez A (2011) Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. Journal of the American Statistical Association 106(496):1418–1433

Glynn PW, Rhee CH (2014) Exact estimation for Markov chain equilibrium expectations. Journal of Applied Probability 51(A):377–389

Godsill SJ (2001) On the relationship between Markov chain Monte Carlo methods for model uncertainty. Journal of Computational and Graphical Statistics 10(2):230–248

Heng J, Jacob PE (2019) Unbiased Hamiltonian Monte Carlo with couplings. Biometrika 106(2):287–302

Hinne M, Lenkoski A, Heskes T, van Gerven M (2014) Efficient sampling of Gaussian graphical models using conditional Bayes factors. Stat 3(1):326–336

Jacob PE, Lindsten F, Schön TB (2020a) Smoothing with couplings of conditional particle filters. Journal of the American Statistical Association 115(530):721–729

Jacob PE, O'Leary J, Atchadé YF (2020b) Unbiased Markov chain Monte Carlo methods with couplings. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(3):543–600

Jasra A, Stephens DA, Doucet A, Tsagaris T (2010) Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. Scandinavian Journal of Statistics 38(1):1–22

Jasra A, Kamatani K, Law KJH, Zhou Y (2017) Multilevel particle filters. SIAM Journal on Numerical Analysis 55(6):3068–3096

Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005) Experiments in stochastic computation for high-dimensional graphical models. Statistical Science 20(4):388–400

Kantas N, Beskos A, Jasra A (2014) Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier–Stokes equations. SIAM/ASA Journal on Uncertainty

Quantification 2(1):464–489
Lauritzen SL (1996) Graphical Models. Oxford Statistical Science Series, The Clarendon Press, Oxford University Press, New York
Lee A, Singh SS, Vihola M (2020) Coupled conditional backward sampling particle filter. Annals of Statistics 48(5):3066–3089
Lenkoski A (2013) A direct sampler for G-Wishart variates. Stat 2(1):119–128
Middleton L, Deligiannidis G, Doucet A, Jacob PE (2019) Unbiased smoothing using particle independent Metropolis-Hastings. In: Chaudhuri K, Sugiyama M (eds) Proceedings of Machine Learning Research, PMLR, Proceedings of Machine Learning Research, vol 89, pp 2378–2387
Murray I, Ghahramani Z, MacKay DJC (2006) MCMC for doubly-intractable distributions. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, USA, UAI'06, p 359–366
Rosenthal JS (1997) Faithful couplings of Markov chains: Now equals forever. Advances in Applied Mathematics 18(3):372–381
Roverato A (2002) Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scandinavian Journal of Statistics 29(3):391–411
Soh SE, Tint MT, Gluckman PD, Godfrey KM, Rifkin-Graboi A, Chan YH, Stünkel W, Holbrook JD, Kwek K, Chong YS, Saw SM, the GUSTO Study Group (2014) Cohort profile: Growing up in Singapore towards healthy outcomes (GUSTO) birth cohort study. International Journal of Epidemiology 43(5):1401–1409
Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, Järvelin MR, Kähönen M, Lehtimäki T, Viikari J, Raitakari OT, Savolainen MJ, Ala-Korpela M (2009) High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. The Analyst 134(9):1781
Statisticat, LLC (2020) LaplacesDemon: Complete Environment for Bayesian Inference. R package version 16.1.4
Tan LSL, Jasra A, De Iorio M, Ebbels TMD (2017) Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. The Annals of Applied Statistics 11(4):2222–2251
Uhler C, Lenkoski A, Richards D (2018) Exact formulas for the normalizing constants of Wishart distributions for graphical models. The Annals of Statistics 46(1):90–118
Wang H, Li SZ (2012) Efficient Gaussian graphical model determination under G-Wishart prior distributions. Electronic Journal of Statistics 6(0):168–198

A Systematic resampling

Algorithms 7 through 9 detail the systematic resampling methods used for the empirical results derived from Algorithm 6. They involve the floor function denoted by $\lfloor x \rfloor$, i.e., $\lfloor x \rfloor$ is the largest integer for which $\lfloor x \rfloor \leq x$.

Algorithm 7 Systematic resampling.

Input: Probability vector $p_{1:N}$ and $U \in [0,1]$.

1. Compute the cumulative sums $v_i = N\sum_{j=1}^{i} p_i$, $i = 1,\ldots,N$.
2. Let $j = 1$.
3. For $i = 1,\ldots,N$,
   (a) While $v_j < U$, do $j = j+1$.
   (b) Let $A_i = j$ and $U = U+1$.

Output: A vector of N random indices A such that, if the input U follows $\mathscr{U}(0,1)$, then the expectation of the frequency of any index i equals Np_i.

Algorithm 8 (Chopin and Singh 2015, Algorithm 4) Conditional systematic resampling.

Input: Probability vector $p_{1:N}$.

1. Compute $r = Np_1 - \lfloor Np_1 \rfloor$ and sample
$$U \sim \begin{cases} \mathscr{U}(0, Np_1), & r \leq 0 \\ \frac{r(\lfloor Np_1 \rfloor + 1)}{Np_1}\,\mathscr{U}(0,r) + \frac{Np_1 - r(\lfloor Np_1 \rfloor + 1)}{Np_1}\,\mathscr{U}(r,1), & r > 0 \end{cases}.$$
2. Obtain a vector of indices B by running Algorithm 7 with $p_{1:N}$ and U as input.
3. Draw A uniformly from all cycles of B that yield $A_1 = 1$.

Output: A vector of N random indices A with $A_1 = 1$.

Algorithm 9 Coupled conditional systematic resampling.

Input: Probability vectors $p_{1:N}$, $\bar{p}_{1:N}$.

1. Generate B and $\bar{B}$ by running the first two steps of Algorithm 8 with $p_{1:N}$ and $\bar{p}_{1:N}$ as input, respectively, using the same random numbers for each.
2. Construct a joint probability distribution on the cycles A and $\bar{A}$ of B and $\bar{B}$ for which $A_1 = \bar{A}_1 = 1$ such that the marginal distributions are uniform:
   (a) Calculate the overlap $|\{i \in 1:N \mid A_i = \bar{A}_i\}|$ for each pair of cycles with $A_1 = \bar{A}_1 = 1$.
   (b) Iteratively assign the largest probability afforded by the constraint of uniform marginals to the pair with the highest overlap.
3. Draw A and $\bar{A}$ from this joint distribution on cycles of B and $\bar{B}$.

Output: Vectors $A, \bar{A}$ of N random indices with $A_1 = \bar{A}_1 = 1$.

B Proofs for Section 4

Our results derive from Lee et al. (2020). They consider a smoothing set-up which maps to our context of approximating a general posterior $\pi(x)$ using adaptive SMC. Specifically, their target density is (Lee et al. 2020, Equation 1)

$$\Pi(x_{0:S}) \propto M_0(x_0)\, G_0(x_0) \prod_{s=1}^{S} M_s(x_{s-1}, x_s)\, G_s(x_{s-1}, x_s). \tag{4}$$

In our context, the term $M_0(x) = \pi_{\alpha_0}(x)$ is a tempered posterior, the term $G_0(x) = p(y \mid x)^{\alpha_1 - \alpha_0}$ a tempered likelihood, $M_s(x_{s-1}, x_s)$ the density of the Markov transition starting at x_{s-1} resulting from the m_s MCMC steps which are invariant w.r.t. $\pi_{\alpha_s}(x)$ in Step 2c of Algorithm 2 for $s = 1,\ldots,S$, $G_s(x_{s-1}, x_s) = p(y \mid x_s)^{\alpha_{s+1} - \alpha_s}$ a tempered likelihood for $s = 1,\ldots,S-1$, and $G_S(x_{S-1}, x_S) = p(y \mid x_S)^{1-\alpha_S}$ a tempered likelihood. Then, the coupled conditional particle filter in Algorithm 2 of Lee et al. (2020) reduces to the coupled conditional SMC in our Algorithm 4. Thus, the results in Lee et al. (2020) apply to Algorithm 4.

B.1 Proof of Proposition 1

Since $G_s(x_{s-1}, x_s)$ does not depend on x_{s-1}, we can write $G_s(x_{s-1}, x_s) = G(x_s)$ for $s = 1,\ldots,S$ as in Section 2 of Lee et al. (2020). Assumption 1, that $p(y \mid x)$ is bounded, implies that $G_s(x_s)$ is bounded for $s = 0,\ldots,S$, which is Assumption 1 in Lee et al. (2020). Therefore, Theorem 8 of Lee et al. (2020) provides $\Pr(x'_{0:S} = \bar{x}'_{0:S}) \geq N/(N+c)$.

Part (iii) follows similarly to the proof for Theorem 10(iii) of Lee et al. (2020): We have $\Pr(\tau > t) \leq \{1 - N/(N+c)\}^{t-1}$ for $t \geq 1$. Therefore,

$$E(\tau) = \sum_{t=0}^{\infty} \Pr(\tau > t) \leq 1 + \sum_{t=1}^{\infty} \Pr(\tau > t) \leq 1 + \sum_{t=1}^{\infty} \left(1 - \frac{N}{N+c}\right)^{t-1} = 2 + \frac{c}{N},$$

where the last equality follows from the geometric series formula $\sum_{t=0}^{\infty}(1-r)^t = 1/r$ for $|r| < 1$. Part (iii) implies Part (ii). □

B.2 Proof of Proposition 2

Theorem 10 of Lee et al. (2020) provides results for a statistic that we denote by $h_{0:S} : \mathscr{X}^{S+1} \to \mathbb{R}$. Consider $h_{0:S}$ defined by $h_{0:S}(x_{0:S}) = h(x_S)$ where $h : \mathscr{X} \to \mathbb{R}$ is our statistic of interest. Then, $h_{0:S}$ is bounded by Assumption 2. The marginal distribution of x_S under the density on $x_{0:S}$ in (4) is our posterior of interest $\pi(x)$. Consequently, the results for $h_{0:S}$ in Theorem 10 of Lee et al. (2020) provide the required results for h. □

C Comparison with coupled HMC

The coupled HMC method of Heng and Jacob (2019) provides an alternative to coupled particle MCMC for unbiased posterior approximation if the posterior is amenable to HMC. The latter typically requires $\mathscr{X} = \mathbb{R}^{d_x}$ and that the posterior is continuously differentiable. Here, we apply coupled HMC to the posterior considered in Section 5.1 with a slight modification to make it suitable for HMC: the uniform prior over the hypercube $[-10, 10]^{d_x}$ is replaced by the improper prior $p(x) \propto 1$ for $x \in \mathbb{R}^{d_x}$ to ensure differentiability. The set-up of coupled HMC follows Section 5.2 of Heng and Jacob (2019) with the following differences. The leap-frog step size is set to 0.1 instead of 1 as the resulting MCMC failed to accept with the latter. We do not initialize both chains independently but instead set $\bar{x}(1) = x(0)$ as in Algorithm 6 since we found that this change reduces meeting times. We use code from `https://github.com/pierrejacob/debiasedhmc` to implement the method from Heng and Jacob (2019).

Figure 5 presents the results analogously to Figure 1. In terms of number of iterations, coupled HMC mixes worse and takes longer to meet than coupled particle MCMC. These increases are not offset by a lower computational cost per iteration. An important caveat here is that computation time depends on the implementation, and here coupled HMC is implemented using an `R` package and coupled particle MCMC in Python.

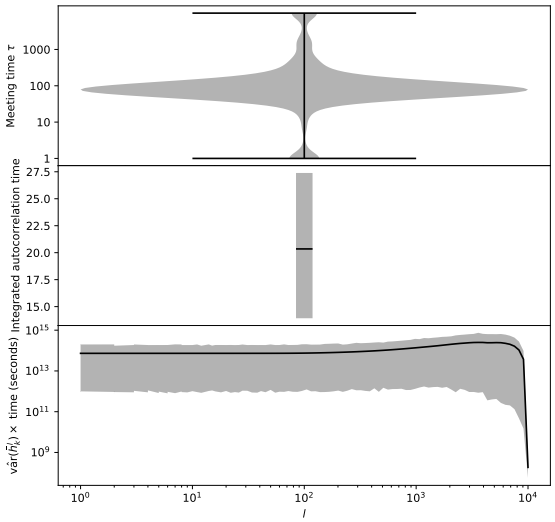


Fig. 5 Results from execution of coupled HMC from Heng and Jacob (2019) for the case of the mixture of Gaussians, under the adjustments described in Appendix C. The top row contains violin plots of $\log(\tau)$. The bottom two rows show the integrated autocorrelation time and $\hat{\text{var}}(\bar{h}_k^l) \times$ time as a function of l, with their means and 95% bootstrapped confidence intervals in black and gray, respectively.

D Additional simulations studies

Here, we provide some further simulation studies where the set-up is the same as in Section 5.1 except for the following. We consider a probability of PIMH of $\rho = 0.05$ in addition to the other values of ρ, the maximum l is $l_{\max} = 2 \cdot 10^3$ and the number of repetitions is $R = 128$. Figure 8 considers different number of particles of N. Figure 9 varies the dimensionality of the parameter d_x where we use the true values $x^* = (-3, 0, 3)^\top$ and $x^* = (-3, 0, 3, 6)^\top$ for $d_x = 3$ and $d_x = 4$, respectively, based on the set-up in Middleton et al. (2019, Appendix B.2). Additionally, Figure 9b uses independent inner MCMC steps across both chains except for that the MCMC step is faithful to any coupling. This contrasts with Section 5.1 which uses a common random number coupling for the Metropolis-Hastings inner MCMC step.

A higher number of particles N results in shorter meeting times. Criterion '$\hat{\text{var}}(\bar{h}_k^l) \times$ time' is lowest for larger N, though beyond a certain N, not much improvement is gained. Jacob et al. (2020a) reach a similar conclusion when varying N for coupled conditional particle filters.

Performance deteriorates with increasing dimensionality d_x, especially for smaller values of ρ. For $d_x = 4$ (Figure 9d), the chains even often fail to meet within the maximum number of iterations of 2,000 considered for $\rho = 0, 0.05$. We also see such lack of coupling in Figure 9b for $\rho = 0$, suggesting that the coupling of the inner MCMC is important for good performance when working with coupled conditional SMC. This is despite the fact that the theoretical results in Section 4 do not depend on the quality of the coupling of the inner MCMC.

For certain values of l, using ρ away from 0 or 1 is competitive with conditional SMC or PIMH in terms of '$\hat{\text{var}}(\bar{h}_k^l) \times$ time' although not notably better than using just one of them. The benefit of a mixture versus using only conditional SMC in terms of coupling is highlighted in Figure 9b where the inner MCMC is uncoupled.

E Inner MCMC step for Gaussian graphical models

We set up an MCMC step with $p(x \mid y) = p(K, G \mid Y)$ as invariant distribution. The corresponding MCMC step for the tempered density $p_\alpha(x \mid y)$, $\alpha \in (0, 1]$, required for Algorithm 6, follows by replacing n and U by αn and αU, respectively, as $p(y \mid x)^\alpha = (2\pi)^{-\alpha np/2} |K|^{\alpha n/2} \exp(-\frac{1}{2} \langle K, \alpha U \rangle)$. We make use of the algorithm for sampling from a G-Wishart law introduced in Lenkoski (2013, Section 2.4). Thus, we can sample from $K \mid G, Y \sim \mathscr{W}_G(\delta + n, D^*)$. It remains to derive an MCMC transition that preserves $p(G \mid Y)$, as samples of G can be extended to $x = (K, G)$ by generating $K \mid G, Y$.

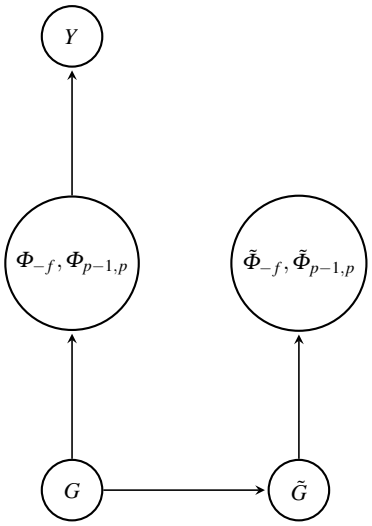


Fig. 6 The enlarged hierarchical model giving rise to a posterior on an extended space that will be preserved by our MCMC kernel. The construction aims at: i) removing the requirement for calculation of intractable normalising constants; ii) avoiding introducing tuning parameters. The main text defines $\Phi = (\Phi_{-f}, \Phi_{p-1,p})$ and $\tilde{\Phi} = (\tilde{\Phi}_{-f}, \tilde{\Phi}_{p-1,p})$.

We consider the double reversible jump approach from Lenkoski (2013) and apply the node reordering from Cheng and Lenkoski (2012, Section 2.2) to obtain an MCMC step with no tuning parameters. The MCMC step is a Metropolis-Hastings algorithm on an enlarged space that bypasses the evaluation of the intractable normalisation constants $I_G(\delta, D)$ and $I_G(\delta+n, D^*)$ in the target distribution (3). It is a combination of ideas from the PAS algorithm of Godsill (2001), which avoids the evaluation of $I_G(\delta+n, D^*)$, and the exchange algorithm of Murray et al. (2006), which sidesteps evaluation of $I_G(\delta, D)$. We will give a brief presentation of the MCMC kernel that we are using as it does not coincide with approaches that have appeared in the literature.

To attain the objective of suppressing the normalising constants in the method, one works with a posterior on an extended space, defined via the directed acyclic graph in Figure 6. The left side of the graph gives rise to the original posterior $p(G)\,p(K \mid G)\,p(Y \mid K)$. Denote by $\tilde{G}$ the proposed graph, with law $q(\tilde{G} \mid G)$. Lenkoski (2013) chooses a pair of vertices (i, j) in G, $i < j$, at random and applies a reversal, i.e. $(i, j) \in \tilde{G}$ if and only if $(i, j) \notin G$. The downside is that the probability of removing an edge is proportional to the number of edges in G, which is typically small. Instead, we consider the method in Dobra et al. (2011, Equation A.1) that also applies the reversal, but chooses (i, j) so that the probabilities of adding and removing an edge are equal.

We reorder the nodes of G and $\tilde{G}$ so that the edge that has been altered is $(p-1, p)$, similarly to Cheng and Lenkoski (2012, Section 2.2). Given $\tilde{G}$, the graph in Figure 6 contains a final node that refers to the conditional distribution of $p(\tilde{K} \mid \tilde{G})$ which coincides with the G-Wishart prior $p(K \mid G)$. Consider the upper triangular Cholesky decomposition Φ of K so that $\Phi^\top \Phi = K$. Let $\Phi_{-f} = \Phi \setminus \Phi_{p-1,p}$. We work with the map $K \leftrightarrow \Phi = (\Phi_{-f}, \Phi_{p-1,p})$. We apply a similar decomposition for $\tilde{K}$, and obtain the map $\tilde{K} \leftrightarrow \tilde{\Phi} = (\tilde{\Phi}_{-f}, \tilde{\Phi}_{p-1,p})$.

We can now define the target posterior on the extended space as

$$p\big(G, \tilde{G}, \Phi_{p-1,p}, \tilde{\Phi}_{p-1,p} \mid \Phi_{-f}, \tilde{\Phi}_{-f}, Y\big) \propto p\big(G\big)\, q(\tilde{G} \mid G)\, p(\Phi \mid G)\, p(\tilde{\Phi} \mid \tilde{G})\, p(Y \mid \Phi). \quad (5)$$

Given a graph G, the current state on the extended space comprises of

$$\big(G, \tilde{G}, \Phi_{-f}, \Phi_{p-1,p}, \tilde{\Phi}_{-f}, \tilde{\Phi}_{p-1,p}\big), \quad (6)$$

with $\tilde{G} \sim q(\tilde{G} \mid G)$, and Φ, $\tilde{\Phi}$ obtained from the Cholesky decomposition of the precision matrices $K \sim \mathcal{W}_G(\delta+n, D^*)$, $\tilde{K} \sim \mathcal{W}_{\tilde{G}}(\delta, D)$, respectively. Note that the rows and columns of D, D^* have been accordingly reordered to agree with the re-arrangement of the nodes we describe above. Consider the scenario with the proposed graph $\tilde{G}$ having one more edge than G. Given the current state in (6), the algorithm proposes a move to the state

$$\big(\tilde{G}, G, \Phi_{-f}, \Phi^{\mathrm{pr}}_{p-1,p}, \tilde{\Phi}_{-f}, \tilde{\Phi}^{\mathrm{pr}}_{p-1,p}\big). \quad (7)$$

The value $\Phi^{\mathrm{pr}}_{p-1,p}$ is sampled from the conditional law of $\Phi_{p-1,p} \mid \Phi_{-f}, Y$.

We provide here some justification for the above construction. The main points are the following: (i) the proposal corresponds to an exchange of $G \leftrightarrow \tilde{G}$, coupled with a suggested value for the newly 'freed' matrix element $\Phi^{\mathrm{pr}}_{p-1,p}$; (ii) from standard properties of the general exchange algorithm, switching the position of $G, \tilde{G}$ will cancel out the normalising constants of the G-Wishart prior from the acceptance probability; (iii) the normalising constants of the G-Wishart posterior never appear, as the precision matrices are not integrated out.

Appendix F derives that

$$\Phi_{p-1,p} \mid \Phi_{-f}, Y \sim \mathcal{N}\left(\frac{-D^*_{p-1,p}\Phi_{p-1,p-1}}{D^*_{p,p}}, \frac{1}{D^*_{p,p}}\right) \quad (8)$$

This avoids the tuning of a step-size parameter arising in the Gaussian proposal of Lenkoski (2013, Section 3.2). The variable $\tilde{\Phi}^{\mathrm{pr}}_{p-1,p}$ is not free, due to the edge $(p-1, p)$ assumed being removed, and is given as (Roverato 2002, Equation 10)

$$\tilde{\Phi}^{\mathrm{pr}}_{p-1,p} = -\sum_{i=1}^{p-2} \tilde{\Phi}_{i,p-1}\tilde{\Phi}_{ip} / \tilde{\Phi}_{p-1,p-1}$$

The acceptance probability of the proposal is given in Step 6 of the complete MCMC transition shown in Algorithm 10 for exponent $\varepsilon = 1$. In the opposite scenario when an edge is removed from G, then, after again re-ordering the nodes, the proposal $\tilde{\Phi}^{\mathrm{pr}}_{p-1,p}$ is sampled from

$$\tilde{\Phi}_{p-1,p} \mid \tilde{\Phi}_{-f} \sim \mathcal{N}\left(\frac{-D_{p-1,p}\tilde{\Phi}_{p-1,p-1}}{D_{p,p}}, \frac{1}{D_{p,p}}\right)$$

whereas we fix $\Phi^{\mathrm{pr}}_{p-1,p} = -\sum_{i=1}^{p-2} \Phi_{i,p-1}\Phi_{ip} / \Phi_{p-1,p-1}$. The corresponding acceptance probability for the proposed move is again as in Step 6 of Algorithm 10, but now for $\varepsilon = -1$.

F Proposal for precision matrices

This derivation is similar to Appendix A of Cheng and Lenkoski (2012). Assume that the edge $(p-1, p)$ is in the proposed graph $\tilde{G}$ but not in G. The prior on $\tilde{\Phi}_{p-1,p} \mid \tilde{\Phi}_{-f}$ follows from Equation 2 of Cheng and Lenkoski (2012) as

$$p(\tilde{\Phi}_{p-1,p} \mid \tilde{\Phi}_{-f}, \tilde{G}) \propto \exp\left(-\frac{1}{2}\langle \tilde{\Phi}^\top \tilde{\Phi}, D\rangle\right).$$

The likelihood is

$$p(Y \mid \tilde{K}) \propto |\tilde{K}|^{n/2} \exp\left(-\frac{1}{2}\langle \tilde{K}, U\rangle\right).$$

Here, $|\tilde{K}|$ does not depend on $\tilde{\Phi}_{p-1,p}$ since $|\tilde{K}| = |\tilde{\Phi}|^2 = (\prod_{i=1}^p \tilde{\Phi}_{ii})^2$. Combining the previous two displays thus yields $p(\tilde{\Phi}_{p-1,p} \mid \tilde{\Phi}_{-f}, Y) \propto \exp(-\langle \tilde{\Phi}^\top \tilde{\Phi}, D^*\rangle/2)$. Dropping terms not involving $\tilde{\Phi}_{p-1,p}$ yields (8).

Algorithm 10 Inner MCMC step for the Gaussian graphical model.

Input: Graph G.

1. (a) If G is complete or empty, sample (i, j), $i < j$, uniformly from the edges in G or the edges not in G, respectively.
   (b) Else, sample (i, j), $i < j$, as follows: w.p. $1/2$, uniformly from the edges in G; w.p. $1/2$, uniformly from the edges not in G.
2. Reorder the nodes in G so that the i-th and j-th nodes become the $(p-1)$-th and p-th nodes, respectively. Rearrange D and D^* accordingly.
3. Let $\tilde{G}$ be as G except for edge $(p-1, p)$, with $(p-1, p) \in \tilde{G}$ if and only if $(p-1, p) \notin G$.
4. Sample $K \sim \mathcal{W}_G(\delta + n, D^*)$, $\tilde{K} \sim \mathcal{W}_{\tilde{G}}(\delta, D)$. Compute the corresponding upper triangular Cholesky decompositions Φ, $\tilde{\Phi}$.
5. Fix $\Phi_{-f} = \Phi \setminus \Phi_{p-1,p}$ and $\tilde{\Phi}_{-f} = \tilde{\Phi} \setminus \tilde{\Phi}_{p-1,p}$.
   (a) If $(p-1, p) \notin G$, sample the proposal
   $$\Phi^{\text{pr}}_{p-1,p} \mid \Phi_{-f}, Y \sim \mathcal{N}(-D^*_{p-1,p}\Phi_{p-1,p-1}/D^*_{p,p},\, 1/D^*_{p,p})$$
   and set $\tilde{\Phi}^{\text{pr}}_{p-1,p} = -\sum_{i=1}^{p-2} \tilde{\Phi}_{i,p-1}\tilde{\Phi}_{ip}/\tilde{\Phi}_{p-1,p-1}$.
   (b) If $(p-1, p) \in G$, sample the proposal
   $$\tilde{\Phi}^{\text{pr}}_{p-1,p} \mid \tilde{\Phi}_{-f} \sim \mathcal{N}(-D_{p-1,p}\tilde{\Phi}_{p-1,p-1}/D_{p,p},\, 1/D_{p,p})$$
   and set $\Phi^{\text{pr}}_{p-1,p} = -\sum_{i=1}^{p-2} \Phi_{i,p-1}\Phi_{ip}/\Phi_{p-1,p-1}$.
6. Accept proposed move from $\big(G, \tilde{G}, \Phi_{-f}, \Phi_{p-1,p}, \tilde{\Phi}_{-f}, \tilde{\Phi}_{p-1,p}\big)$ to state $\big(\tilde{G}, G, \Phi_{-f}, \Phi^{\text{pr}}_{p-1,p}, \tilde{\Phi}_{-f}, \tilde{\Phi}^{\text{pr}}_{p-1,p}\big)$ w.p. $1 \wedge R$, for R equal to
   $$\frac{p(\tilde{G})\,q(G \mid \tilde{G})}{p(G)\,q(\tilde{G} \mid G)} \exp\Big\{-\tfrac{1}{2}\langle K^{\text{pr}} - K, D^*\rangle - \tfrac{1}{2}\langle \tilde{K}^{\text{pr}} - \tilde{K}, D\rangle\Big\}$$
   $$\times \left[\frac{\Phi_{p-1,p-1}\sqrt{D_{p,p}}}{\tilde{\Phi}_{p-1,p-1}\sqrt{D^*_{p,p}}} \exp\Big\{\tfrac{D_{p,p}}{2}(\tilde{\Phi}^{\text{pr}}_{p-1,p} - \tilde{\theta})^2 - \tfrac{D^*_{p,p}}{2}(\Phi^{\text{pr}}_{p-1,p} - \theta)^2\Big\}\right]^{\varepsilon},$$
   for $\tilde{\theta} = -D_{p-1,p}\tilde{\Phi}_{p-1,p-1}/D_{p,p}$ and $\theta = -D^*_{p-1,p}\Phi_{p-1,p-1}/D^*_{p,p}$. Here, K^{pr} (resp. $\tilde{K}^{\text{pr}}$) denotes the precision with upper triangular Cholesky decomposition given by the synthesis of $\Phi_{-f}, \Phi^{\text{pr}}_{p-1,p}$ (resp. $\tilde{\Phi}_{-f}, \tilde{\Phi}^{\text{pr}}_{p-1,p}$), and $\varepsilon = -1$ if $(p-1, p) \in G$, else $\varepsilon = 1$.
7. Revert the reordering from Step 2. Return $\tilde{G}$ if the proposed move at Step 6 is accepted, else return G.

Output: MCMC update for graph G such that the invariant distribution is the target posterior $p(G \mid Y)$.

G Comparison with SMC for the metabolite application

We compare the results in Figure 3 with those from running the SMC in Algorithm 2 with a large number of particles $N = 10^5$. Comparing Figures 3 and 7 shows that the results are largely the same. The edge probabilities for which they differ substantially are harder to estimate according to the Monte Carlo standard errors from coupled particle SMC in Figure 3.

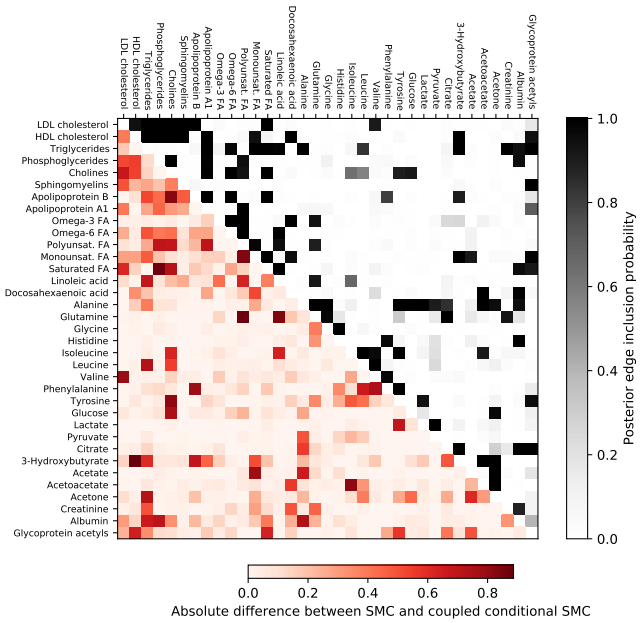
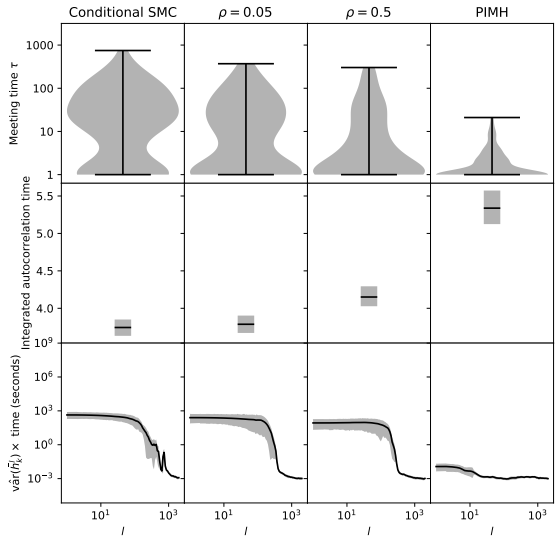
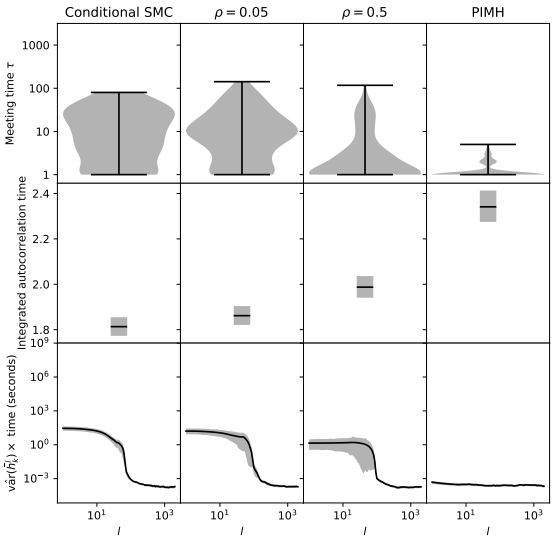


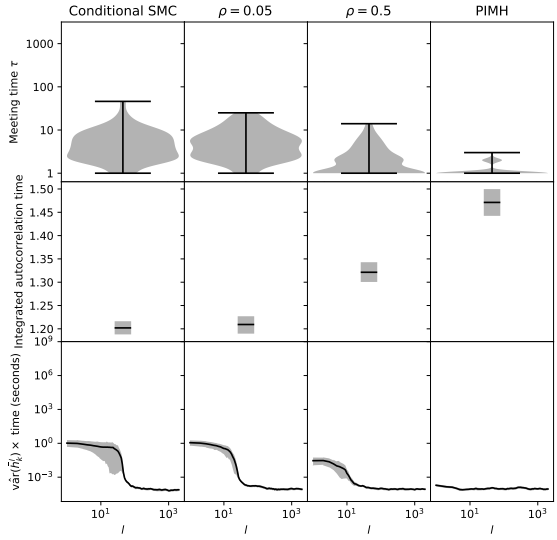
Fig. 7 Posterior edge inclusion probabilities for the Gaussian graphical model fit on the metabolite data from SMC with $N = 10^5$ particles (upper triangle) and their absolute difference from the estimates from coupled particle MCMC in Figure 3 (lower triangle). LDL, HDL and FA stand for low-density lipoprotein, high-density lipoprotein and fatty acids, respectively.



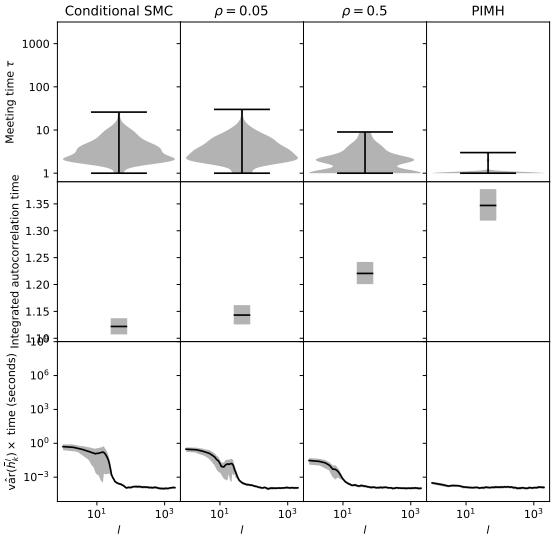
(a) $d_x = 2, N = 25$



(b) $d_x = 2, N = 100$

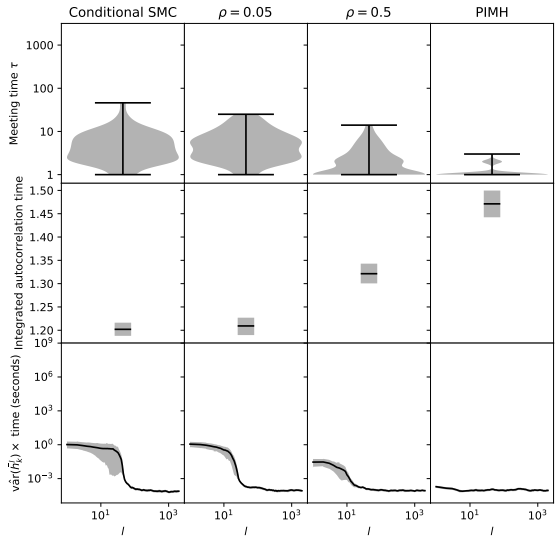


(c) $d_x = 2, N = 500$

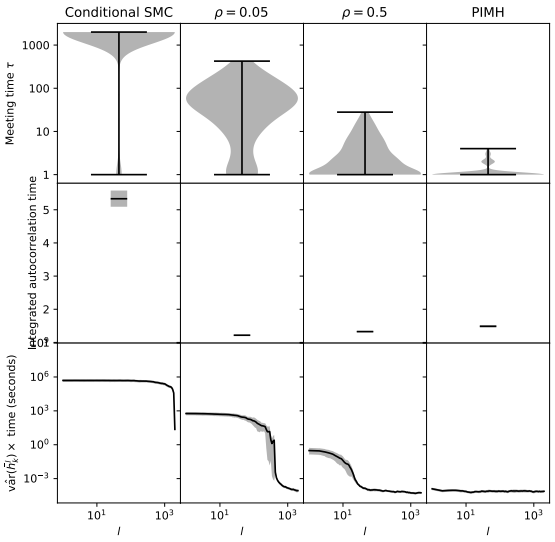


(d) $d_x = 2, N = 1000$

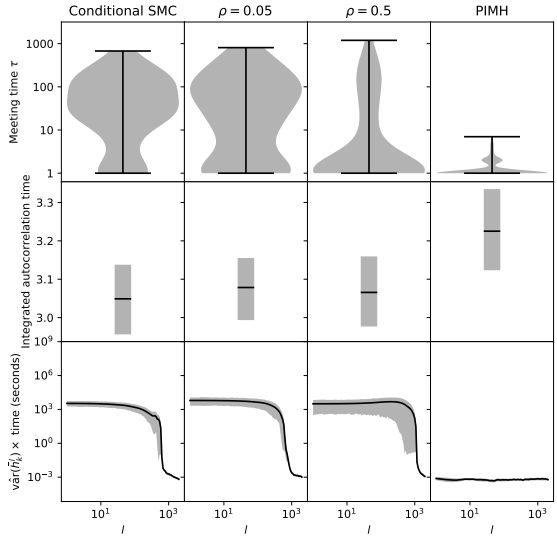
Fig. 8 Results from execution of Algorithm 6 with $\rho = 0$ (conditional SMC), $\rho = 0.05$, $\rho = 0.5$ and $\rho = 1$ (PIMH), for the case of the mixture of Gaussians for various number of particles N. The top row contains violin plots of $\log(\tau)$. The bottom two rows show the integrated autocorrelation time and $\hat{\text{var}}(\bar{h}_k^l) \times$ time as a function of l, with their means and 95% bootstrapped confidence intervals in black and gray, respectively.



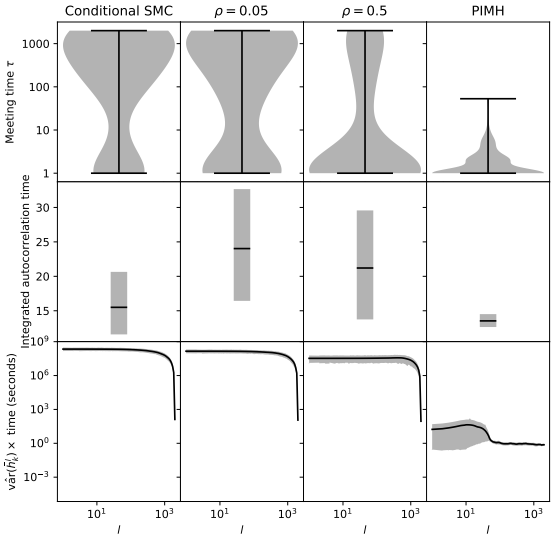
(a) $d_x = 2$, $N = 500$



(b) $d_x = 2$, $N = 500$, uncoupled inner MCMC



(c) $d_x = 3$, $N = 500$



(d) $d_x = 4$, $N = 500$

Fig. 9 Results from execution of Algorithm 6 with $\rho = 0$ (conditional SMC), $\rho = 0.05$, $\rho = 0.5$ and $\rho = 1$ (PIMH), for the case of the mixture of Gaussians for various parameter dimensionalities d_x and, in (b), for when the inner MCMC steps are not coupled beyond being faithful. The top row contains violin plots of $\log(\tau)$. The bottom two rows show the integrated autocorrelation time and $\hat{\text{var}}(\bar{h}_k^l) \times$ time as a function of l, with their means and 95% bootstrapped confidence intervals in black and gray, respectively.