Scalable inference in state-space models



Lawrence Middleton

St. Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2019

Abstract

This thesis provides a set of novel Monte Carlo methods to perform Bayesian inference, with an emphasis on a state-space modelling framework. The thesis comprises three self-contained research papers. The first paper works at the interface of modelling and inference, extending existing work to develop a new model for medium and high-dimensional data possessing step-changes changes in the correlation structure. A bespoke inference routine is developed, with experimental results demonstrating the speed-up of convergence to equilibrium under a novel alteration of a particle Gibbs sampler for change-point detection. The second paper is concerned with the problem of obtaining unbiased estimators for the smoothing problem in state-space models. New methodology is developed leveraging recent developments in unbiased Markov chain Monte Carlo (MCMC) in conjunction with a particle independent Metropolis–Hastings Markov kernel. We show that the resulting coupling time can be precisely characterised. As a result, it is possible to study the large sample limit of the coupling time to help guide the choice in an algorithmic tuning parameter (the number of particles in this case). The third and final paper develops this methodology further, to obtain unbiased estimators for MCMC algorithms in settings where the likelihood may be unavailable to evaluate pointwise. The motivation for this work is to find a simple means by which to parallelise exact Bayesian inference in state-space models. The focus is primarily on pseudo-marginal algorithms and particle MCMC, in which the mixing behaviour of the algorithm is typically polynomially ergodic. In this setting we propose new methodology to allow the obtention of unbiased estimators of expectations of stationary distributions of these Markov chains. In addition, we extend existing results to both prove the validity and characterise the computational efficiency of these estimators, in terms of algorithmic tuning parameters. We evaluate the resulting methodology on otherwise intractable state-space and latent variable models, including a multivariate probit model popular in econometrics and a nonlinear state-space model arising in neuroscience. Additional algorithmic extensions are provided and subsequently applied to a parameter inference problem in a high-dimensional Ising model.

Contents

1 Introduction and thesis structure **1**

2 Bayesian inference of structural breaks in factor copula models **3**

2.1 Introduction 5

2.1.1 Background and Introduction 5

2.1.2 Copula models for time series 6

2.1.3 Copula factor models 8

2.1.4 Pseudo-observations and two stage inference 10

2.2 Changepoint copula factor model 11

2.2.1 Single-factor change-point model 11

2.3 Inference 13

2.3.1 State-space model inference in changepoint problems 14

2.3.2 Particle Gibbs sampler 16

2.4 Applications 19

2.4.1 Synthetic data 19

2.5 Structural breaks during 2008 financial crisis 23

2.5.1 Inference 24

2.5.2 Results 25

2.6 Conclusion 29

2.7 Appendix 31

2.7.1 Particle inference for changepoint problems 31

3 Unbiased Smoothing using Particle Independent Metropolis-Hastings **38**

3.1 Introduction 40

3.1.1 State-space models, problem statement and review 40

3.1.2 Particle methods 41

3.1.3 PIMH method 43

3.2 Methodology 44

3.2.1 Unbiased MCMC via couplings 44

3.2.2 Coupled PIMH 46

3.2.3 Validity and meeting times 47

3.2.4 Unbiased filtering 49

3.3 Analysis . . . 49
3.3.1 Meeting time: large sample approximation . . . 49
3.3.2 On the selection of N for large m . . . 51
3.4 Numerical experiments . . . 52
3.4.1 Stochastic kinetic model . . . 53
3.4.2 Lévy-driven stochastic volatility . . . 54
3.5 Discussion . . . 56
3.6 Appendix . . . 59
3.6.1 Proof of Proposition 9 . . . 59
3.6.2 Integrated autocorrelation time for various test functions . . . 60
3.6.3 Rao-Blackwellisation for stochastic kinetic model . . . 60
3.6.4 Data and parameter estimation . . . 61
3.7 Application of unbiased estimation to SMC samplers . . . 62
3.7.1 Bayesian computation using SMC samplers . . . 62
3.7.2 Numerical example . . . 63

4 Unbiased Markov chain Monte Carlo for intractable target distributions 65
4.1 Introduction . . . 67
4.1.1 Context . . . 67
4.1.2 Unbiased estimators from coupled Markov chains . . . 69
4.1.3 Theoretical validity under polynomial tails . . . 72
4.1.4 Conditions for polynomial tails . . . 73
4.1.5 Efficiency under polynomial tails . . . 75
4.2 Unbiased pseudo-marginal MCMC . . . 78
4.2.1 Pseudo-marginal Metropolis–Hastings . . . 78
4.2.2 Coupled pseudo-marginal Metropolis–Hastings . . . 79
4.2.3 Theoretical guarantees . . . 80
4.2.4 Tails of meeting times in a toy experiment . . . 82
4.3 Experiments in state space models . . . 83
4.3.1 Linear Gaussian state space model . . . 84
4.3.2 Neuroscience experiment . . . 88
4.4 Methodological extensions . . . 94
4.4.1 Block pseudo-marginal method . . . 94
4.4.2 Exchange algorithm . . . 99
4.5 Conclusion . . . 102
4.6 Appendix . . . 103
4.6.1 Proof of Theorem 15 . . . 103
4.6.2 Proof of Theorem 18 . . . 104

4.6.3 Proof of Proposition 19 109
4.6.4 Proof of Proposition 22 113
4.6.5 Description of referendum survey data 117

5 Conclusion **118**
5.1 Bayesian inference of structural breaks in factor copula models . . . 118
5.2 Unbiased Smoothing using Particle Independent Metropolis–Hastings 118
5.3 Unbiased Markov chain Monte Carlo for intractable target distributions 119
5.4 A look to the future 120

Bibliography **121**

Acknowledgements

I would like to thank my supervisors and mentors at the Oxford-Man Institute and AHL, namely Anthony, James, Eric and Philipp, for showing me what exemplary research looks like, as well as what it means to be a good researcher. It has been a pleasure to interact so closely with each of them, and I am truly grateful for their guidance and direction, helping me to grow and develop as a researcher. It has been a humbling, challenging and fruitful experience.

In addition, I would like to extend my gratitude to my collaborator and mentor at Harvard, Prof. Jacob as well as my collaborator Prof. Deligiannidis for his insights and ideas. I am equally grateful for the enthusiastic discussions provided by Prof. Anthony Lee in his supervisory role at various points in the process.

Moreover, I am grateful for the feedback and grilling discussions provided by Prof. Geoff Nicholls within the department. He is without doubt the most insightful, diligent and capable faculty member I have been fortunate enough to interact with, and I am grateful for the thoughtfulness and scaffolding he has provided throughout.

Last but not least, I would like to thank the continuous support of my family, with the inclusion of Anil and Stephen, and friends across Oxford, London, Cambridge, Warwick and Cambridge (Massachusetts) that have all contributed to the success of this colossal undertaking and whose unwavering compassion I am extremely grateful for.

1

Introduction and thesis structure

State-space models form a popular class of models used to capture time-varying dependency in data, characterised typically as noisy sensor readings of an unobserved latent Markov process. Recently, new datasets have provided new modelling challenges in areas ranging from bioinformatics, ecology, epidemiology, finance and population genetics that require new and novel inference routines. Many of these applications present highly nonlinear and potentially high-dimensional modelling dependencies and as such require novel methodology to perform tractable inference. Examples include stochastic volatility models in finance, gene autoregulation models in biochemistry and chaotic weather models in physics. The following thesis provides contributions to the existing body of work aiming to perform tractable inference in these models.

This is an integrated thesis, whereby the contributions are provided in three separate research papers, with each paper constituting the equivalent work of a chapter in a traditional thesis. The common theme linking each paper together is Monte Carlo inference with a particular focus on state-space models, though with some results presented more generally. To avoid unnecessary repetition we leave the details of the more specific definition of these objects to the papers, along with respective literature reviews.

Broadly speaking, Monte Carlo methods provide a popular means by which to perform tractable inference in the presence of highly nonlinear complex models. The following thesis aims to contribute to the existing methodology for performing inference, with a primary focus on particle methods. The thesis comprises three chapters and a conclusion, with each chapter corresponding to a distinct research paper. The papers (in order) are as follows:

1. *Bayesian inference of structural breaks in factor copula models*, Middleton, L., Holmes, C., & Doucet, A. (2019).
2. *Unbiased Smoothing using Particle Independent Metropolis–Hastings*, Middleton, L., Deligiannidis, G., Doucet, A., & Jacob, P. E. (2019).
3. *Unbiased Markov chain Monte Carlo for intractable target distributions*, Middleton, L., Deligiannidis, G., Doucet, A., & Jacob, P. E. (2018).

The first chapter illustrates some of the merits of particle methods for inference in state-space models through the development of a new model and accompanying novel particle Gibbs method for the detection of changepoints in medium- to high-dimensional data. Secondly, the next chapter provides new methodology, designed to perform unbiased smoothing in state-space models, with the resulting methodology applicable to more general settings as well. Finally, new methodology is developed to perform unbiased inference of MCMC algorithms, with a focus on subgeometrically-ergodic chains. We show that in this setting we are still able to obtain valid estimators (in the sense of finite variance, with a finite expected computational cost), examine the computational efficiency of these estimators and provide conditions under which the resulting chains satisfy the technical requirements of these theoretical results in realistic settings.

2

Bayesian inference of structural breaks in factor copula models

Lawrence Middleton, Chris Holmes, Arnaud Doucet

Abstract

Building on modern advancements in medium- to high-dimensional copula modelling and Bayesian inference in multiple changepoints, the following extends the high dimensional dynamic copula model of Creal and Tsay (2015) to infer structural changes in the correlations of sets of time series. A novel particle Markov chain Monte Carlo inference scheme is developed to perform efficient inference in the model, exploiting an extra modelling degrees of freedom present in factor models for computational purposes. The methodology performs well on a historical dataset of stock data taken from the S&P 500 as well as synthetic data. It successfully infers the changepoint correlation structure corresponding to a number of significant macroeconomic events over the 2008 financial credit crisis.

2.1 Introduction

2.1.1 Background and Introduction

Copula modelling provides a popular means to isolate dependency modelling from marginal modelling for a set of dependent random variables. Such models have proved useful for the purposes of portfolio selection, risk management and option pricing Cherubini et al. (2004). We here introduce methodology to identify structural changes in these dependencies through the construction of piecewise constant copula models within a Bayesian framework. For the proposed model, we develop an efficient Gibbs sampler capable of approximating the medium- to high-dimensional integrals arising in posterior inference. The model and accompanying inference routine aim to balance tractability while still capturing the salient features of the data. This work extends a dynamic factor copula model proposed recently in Creal and Tsay (2015) to the setting where the dynamic is piecewise constant. The model is estimated on a set of stocks from the S&P 500.

Related work includes Bayesian inference of Gaussian copula models developed in Pitt et al. (2006). In the machine learning literature, inference for structural breaks in sparse Gaussian graphical models has been developed in Xuan and Murphy (2007) and developed subsequently in Yu et al. (2014) for sparse copula graphical models. Both methods, however, only provide point estimates of the resulting changepoint locations rather than a tractable MCMC routine as developed in the following. In particular, we are able to exploit idiosyncratic aspects of model identifiability in order to improve mixing of the resulting MCMC algorithm. Furthermore, assumptions of sparsity are typically inappropriate in a financial setting when assets in the same market can all exhibit a high degree of correlation.

In the econometrics literature, Manner et al. (2018) recently proposed a frequentist approach for structural breaks in factor copula models, relying on the simulated method of moments. Tests for a structural break in dynamic factor models based on principal component analysis have been established in Breitung and Eickmeier (2011), where it is demonstrated that without accounting for structural changes the

number of latent factors may be over estimated. A more recent review of factor models in this setting is provided in Bai and Wang (2016).[1]

2.1.2 Copula models for time series

In the following we derive the factor copula framework used for dynamic factor copula modelling in Creal and Tsay (2015) and modify this to when structural changes are present in the copulas (they are assumed constant within each segment).

Copula modelling provides a well-established means of modelling the comovement of random quantitites. They admit a flexible framework in which a wide class of multivariate distributions can be expressed, effectively 'gluing together' univariate marginal distributions into entire joint models, with the appealing property from the modelling perspective that there is a clear separation between marginal models and dependency between random variables.

For a set of n scalar random variables $\{X^{(i)}\}_{i=1}^n$ each with marginal cumulative distribution $F_i(x) := \mathbb{P}[X_i \leq x]$ copula models rely on transforming each variable through its CDF to form a random variable $U_i = F_i(X_i)$ uniformly distributed on the unit interval, which we will refer to as the *probability integral transform* (PIT). The support of the PITs will depend on the support of X_i though will be taken to be $[0, 1]$ for continuous X_i and for the purposes of exposition we will focus on X_i with support on the real line. We are able to specify a valid joint distribution over this set of random variables $\{X_i\}_{i=1}^n$ through introducing a joint distribution over the unit hypercube $C : ([0, 1]^n, \mathcal{B}([0, 1]^n)) \to [0, 1]$, known as the copula

$$(U_1, ..., U_n) \sim C(du_{1:n}) \quad \text{with} \quad (X_1, ..., X_n) = (F_1^{-1}(U_1), ..., F_n^{-1}(U_n)) \tag{2.1}$$

where $F^{-1}(u) := \inf\{u : F(x) \geq u\}$ is the inverse CDF. Existence of C for any set of random variables and the existence of random variables for any C is formalised by Sklar's theorem Cherubini et al. (2004). As a result, copulas uniquely define a dependency structure between random variables through consideration of their PITs on the n-dimensional unit hypercube.

[1] Code to reproduce experiments can be found in `https://github.com/particlemontecarlo/copula_changepoint`.

Dynamic copulas provide a more flexible means to capture dependencies between a sequence of random variables. Since the treatment of conditional copulas in Patton (2006), there has been a surge of work modelling dynamic dependency structures in financial and econometric models (see Manner and Reznikova (2012) for a more recent review). For a set of stochastic processes $\{X_{1:t}^{(i)}\}_{i=1}^{n}$ evolving on a measurable space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted to a filtration $\mathcal{F}_t$ (time $t \in \mathbb{N}_0$ will be discrete in the following), the dynamic dependency structure can be specified as

$$
\begin{aligned}
(U_t^{(1)}, ..., U_t^{(n)})|\mathcal{F}_{t-1} &\sim C(du_{1:n}|\mathcal{F}_{t-1}) \\
(X_t^{(1)}, ..., X_t^{(n)})|\mathcal{F}_{t-1} &= (F_1^{-1}(U_t^{(1)}|\mathcal{F}_{t-1}), ..., F_n^{-1}(U_t^{(n)}|\mathcal{F}_{t-1}))
\end{aligned}
$$

where $F_i^{-1}(u|\mathcal{F}_{t-1})$ denotes the distribution of each PIT conditional on the σ-algebra generated by the first $t-1$ random variables. Conditioning both the copula and the marginal distributions on the same σ-algebra, $\mathcal{F}_{t-1}$, ensures these conditional distributions are consistent with the correct joint distribution. For modelling purposes, however, the conditional marginal distributions $F_{i,t}^{-1}(u|\mathcal{F}_{t-1})$ implicitly assume that conditional on $X_{t-1}^{(i)}$ then $X_t^{(i)}$ is independent of all $X_{t-1}^{(j)}$ for $j \neq i$, so that $X_t^{(i)} = F_{i,t}^{-1}(U_t|X_{t-1}^{(i)})$. Practically, this means for example that given an asset's previous value, knowlege of any other assets are uninformative towards its next value, except through the dependency structure implied by the conditional copula $C(du_{1:n}|\mathcal{F}_{t-1})$. Such an approach mirrors the static case where the separation of modelling decisions associated with the marginal evolution of an asset and the depedency betweeen them.

For parametric copulas, $C_\vartheta(du_{1:n})$, dynamics are typically introduced through specifying a dynamic on the parameter ϑ_t. Without additional care it may be possible for the dynamic introduced in ϑ_t to induce an undesirable conditional independence structure on the process $(X_t^{(1:n)})$ such that $(X_t^{(1:n)})$ is, for example, no longer Markovian or even forward looking. This is primarily a concern when ϑ_t is piecewise constant, as per Manner et al. (2018), as the condition inevitably poses restrictions on the jump times of ϑ_t. However, based on the above, we see

that provided ϑ_t is measurable with respect to $\mathcal{F}_t$, the resulting marginal process will be too and in this case we express the model as follows

$$(U_t^{(1)}, ..., U_t^{(n)}) \sim C_{\vartheta_t}(du_{1:n}|\vartheta_{1:t-1})$$
$$(X_t^{(1)}, ..., X_t^{(n)})|\{X_{1:t}^{(1:n)} = x_{1:t}^{(1:n)}\} = (F_1^{-1}(U_t^{(1)}|x_{1:t-1}^{(1)}), ..., F_n^{-1}(U_t^{(n)}|x_{1:t-1}^{(n)}))$$

We discuss the dynamics for ϑ_t in more detail in the following.

2.1.3 Copula factor models

Factor copula models can be seen as a means to reduce the degrees of freedom parameterising a dependency structure in high dimensions, based on an underlying factor model. Factor models reduce an n-dimensional model for correlated observations to a p-dimensional model for uncorrelated observations, $p < n$. Bayesian inference in these models in the static case has been established for Gaussian factor copula models in Murray et al. (2013) and in more generality in Oh and Patton (2017) in an econometric setting. In the static case, where pseudo-observations are assumed i.i.d., factor copulas are constructed through firstly defining an auxiliary random variable $\tilde{\mathbf{y}}$ via a linear factor model

$$\tilde{\mathbf{y}}_t = \Lambda \mathbf{z}_t + \epsilon_t \quad \text{where} \quad \mathbf{z}_t \sim p_z(dz) \quad \text{and} \quad \epsilon_t \sim p_\epsilon(d\epsilon) \tag{2.2}$$

where $\mathbf{z}_t \in \mathbb{R}^p$ are low-dimensional zero-mean factors with $\mathbb{C}[\mathbf{z}_t] = I_p$ (the identity matrix) and $\epsilon_t \in \mathbb{R}^n$ are random noise variables, such that ϵ_t and $\mathbf{z}_t$ are independent of each other with $\Lambda \in \mathbb{R}^{n \times p}$ the factor loadings. Following Creal and Tsay (2015) in the time-varying context, we consider the case where ϵ_t are component-wise independent also. The copula related to this factor model is the unique distribution on the n-dimensional hypercube given by transforming each element of $\tilde{\mathbf{y}}_t$ through its marginal distribution. Let λ_i' denote the i^{th} row of Λ and define

$$G_i(y) := \mathbb{P}[\tilde{Y}_t^{(i)} \leq y] = \int \mathbb{1}\left[\lambda_i' z_t + \epsilon_t^{(i)} \leq y\right] p_z(dz) p_\epsilon(d\epsilon)$$

as the (typically intractable) marginal CDF associated with element $\tilde{Y}_t^{(i)} \in \tilde{\mathbf{y}}_t$. The 'factor copula' C_Λ then refers to the unique distribution satisfying

$$\mathbf{u}_t \sim C_\Lambda(du_{1:n}) \quad \text{where} \quad \mathbf{u}_t := (G_1(\tilde{Y}_t^{(1)}), ..., G_n(\tilde{Y}_t^{(n)}))$$

so that a joint distribution over a set of random variables $\{X_t^{(i)}\}_{i=1}^n$ can be induced as in (2.1). Based on the factor structure imposed on $\tilde{\mathbf{y}}_t$ we see that C_Λ will also exhibit a reduction in the number of degrees of freedom available to capture dependencies, with $\mathbf{u}_t$ being just a monotone transformation of $\tilde{\mathbf{y}}_t$. This is the primary motivation for factor copula models.

As noted in Oh and Patton (2017), except in special cases, the form of C_Λ will depend on the distributions p_z and p_ϵ and may be analytically unavailable. However, such models are typically easy to simulate as noted in Creal and Tsay (2015), where G_i was taken to be normal, Student-t and skewed Student-t distributions. In the following, we give special consideration to Gaussian factor models though note these available generalisations. For the Gaussian copula, without loss of generality, both p_z and p_ϵ are zero-mean multivariate normal with identity covariance. As such we have that marginally $\tilde{\mathbf{y}}_t \sim \mathcal{N}(0, \Lambda\Lambda' + I)$, so that G_i is the CDF of $\mathcal{N}(0, 1 + \lambda_i'\lambda_i)$. Marginally (after integrating out factors $\mathbf{z}_t$) $\mathbf{u}_t$ can be given by rescaling $\tilde{\mathbf{y}}_t$ by $\text{diag}\left((\Lambda\Lambda' + I)^{-\frac{1}{2}}\right)$ and transforming through the standard normal CDF Φ, i.e.

$$\mathbf{u}_t = (\Phi^{-1}\left(\frac{\tilde{Y}_t^{(1)}}{\sqrt{1+\lambda_i'\lambda_i}}\right), ..., \Phi^{-1}\left(\frac{\tilde{Y}_t^{(n)}}{\sqrt{1+\lambda_i'\lambda_i}}\right))$$

Equivalently, we have the form given in Murray et al. (2013), Creal and Tsay (2015) in terms of factors $\mathbf{z}_t$ and errors ϵ_t as

$$\begin{aligned} Y_t^{(i)} &= \frac{1}{\sqrt{1+\lambda_i'\lambda_i}}\lambda_i' z_t + \frac{1}{\sqrt{1+\lambda_i'\lambda_i}}\epsilon_{it}, \quad z_t \sim \mathcal{N}(0, I_p), \quad \epsilon_{it} \sim \mathcal{N}(0,1) \\ U_t^{(i)} &= \Phi^{-1}(Y_t^{(i)}) \end{aligned} \tag{2.3}$$

The resulting factor copula C_Λ is, by definition, the Gaussian copula, i.e. the distribution with density

$$C_\Lambda(du_{1:n}) = \frac{1}{|\sqrt{R}|}\exp\left(-\frac{1}{2}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))(R^{-1} - I)(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n))\right) du_{1:n}$$

where R is a correlation matrix given by $R = \text{diag}(\Lambda\Lambda' + I)^{-\frac{1}{2}}(\Lambda\Lambda' + I)\text{diag}(\Lambda\Lambda' + I)^{-\frac{1}{2}}$. Factor models within a financial context effectively 'count' the number of correlations present in a market. If this is asymptotically quadratic we should expect

p to grow proportionally to n. If, however, the correlations grow only linearly with the size of the market we would expect that the correlations are better described by a set of lower dimensional factors (e.g. $p < \infty$ as n increases). Also, owing to the underlying factor structure, the model is unidentifiable without additional constraints as $\tilde{\mathbf{y}} \overset{d}{=} \tilde{\mathbf{y}}'$ where $\tilde{\mathbf{y}}' = \Lambda H \mathbf{z}_t + \epsilon_t$ for any orthogonal matrix H, i.e. the distribution of $\tilde{\mathbf{y}}$ is invariant to rotations of the factors. This has some consequence in the inference procedure and is discussed in more detail in the following.

The model in Equation (2.3) is presented in some generality where each joint observation is assumed well described by p latent factors with $\lambda_i, z_t \in \mathbb{R}^p$. The following will focus primarily on single-factor models with a discussion of multiple factors in Section 2.6. Furthermore, at times it is convenient to refer to either the unscaled factor loadings, λ_i, or scaled factor loadings, $\rho_i = \frac{\lambda_i}{\sqrt{1+\lambda_i^2}}$ depending on the context, though note for $\rho \in (-1, 1)$ there is a one-to-one correspondence. The dynamic copula factor model described in Creal and Tsay (2015) adopts the above dimensionality reduction of Murray et al. (2013) while imposing a Markovian structure on the latent factor loadings. This allows the model to be considered within a state-space framework thereby admitting sequential Monte Carlo as a suitable inference scheme.

2.1.4 Pseudo-observations and two stage inference

As is conventional in the econometric literature, the following approach treats the parameters in the marginal distributions as fixed and performs inference directly on the resulting PITs. As such, the following assumes a two-stage inference procedure where first of all marginal models are fit to the series $\{X_{1:T}^{(i)}\}_{i=1}^n$ and subsequently 'pseudo-observations' are evaluated by $Y_t^{(i)} = \Phi^{-1}(F_i(X_t^{(i)}|X_{1:t-1}^{(i)}))$. In principle this departs from a fully Bayesian treatment as the uncertainty in the marginal models is not propagated into uncertainty in the copula. This however has been the approach advocated in Creal and Tsay (2015), justified as parametric models for marginals are typically $\sqrt{T}$-consistent as well as for tractability. As noted for the inference routine in Pitt et al. (2006), for fully Bayesian inference, any changes to

the marginal models require re-evaluation of the PITs which itself would require an additional Metropolis–Hastings accept-reject step if full joint inference is required.

2.2 Changepoint copula factor model

We here consider a segmentation model of the data, where it is assumed over any given segment the copula is constant. A segmentation of the data, $\mathcal{I}$, is uniquely determined by a set of changepoints $\tau_{1:K}$. A segment is defined by two consecutive change-points τ_i and τ_{i+1}, corresponding to the time indices $t \in \{\tau_i + 1, ..., \tau_{i+1}\}$. A segmentation refers to a sequence of changepoints $\mathcal{I} = \{\tau_0 < \tau_1 < ... < \tau_K\}$ where for convenience $\tau_0 = 0$ and observations are indexed from $t \in \{1, ..., T\}$, therefore constraining the first segment to start at $t = 1$ by convention, and K is used to denote the random number of changepoints. The following adopts a Bayesian approach to changepoint inference, with the distribution of $\mathcal{I}$ following a product partition model of Barry and Hartigan (1992). Within each segment, pseudo-observations are assumed conditionally independent of all those outside the segment, i.e.

$$p(\mathcal{Y}, \{\phi_j\}_{j=1}^K, \theta, \tau_{1:K}) = \prod_{j=1}^{K} \left(\prod_{t \in \mathcal{I}_j} p(y_t|\phi_j, \theta) p(\phi_j|\theta) \right) p(\theta) p(\tau_{1:K}) \tag{2.4}$$

where $\mathcal{Y} := \{y_t^{(i)}\}_{t=1,i=1}^{T,n}$ denotes a dataset of n series of length T, ϕ_j denotes parameters specific to each segment and θ denotes 'global parameters' common to all segments. The objective of the inference routine is to learn jointly both this segmentation and model parameters.

2.2.1 Single-factor change-point model

Based on the product partition structure in (2.4), we introduce the piecewise constant factor copula model as follows

$$\begin{aligned}
\tau_{1:K} &\sim p(\tau_{1:K}) \\
\Lambda_j &\overset{i.i.d.}{\sim} p(\Lambda) \qquad \text{for} \quad j \in \{1, ..., K\} \\
U_t^{(i)} &\overset{i.i.d.}{\sim} C_{\Lambda_j}(du_{1:n}) \quad \text{for} \quad t \in \{\tau_j + 1, ..., \tau_{j+1}\} \\
Y_t^{(i)} &= \Phi(U_t^{(i)})
\end{aligned}$$

where $p(\Lambda)$ denotes the prior distribution on segment-specific factor loadings. The inference algorithms designed in Fearnhead (2006), Fearnhead and Liu (2007), Whiteley et al. (2011) place some assumptions on the prior distibution of changepoints $p(\tau_{1:K})$ in that they assume the lengths of segments are independent under the prior. i.e. letting $d_j := |\tau_{j+1} - \tau_j|$ denote the length of the j^{th} segment we have that the prior distribution of changepoint locations $p(\tau_{1:K})$ is induced by

$$\tau_j = \sum_{s=1}^{j} d_s \qquad \text{where} \qquad d_j \overset{i.i.d.}{\sim} g(d) \qquad \text{for} \qquad \sum_{j=1}^{K} d_j \leq T$$

for some distribution $g : (\mathbb{N}_0, \mathcal{B}(\mathbb{N}_0)) \to [0, 1]$ over the positive integers. As a result, the number of changepoints K is a random variable counting the number of changepoints occurring before T with a distribution implied by the distribution over segment lengths. A simple choice for g is provided by a geometric distribution $g = \text{Geo}(p_G)$, thereby implying that for each $t \in \{1, ..., T\}$ the probability of observing a new changepoint is independent and equal to the probability of observing a changepoint at any other point in time and implies a binomial prior distribution on the number of changepoints, K.

Within each segment, we place independent priors on elements of Λ so that for $j \in \{1, ..., K\}$ we have that $p(\Lambda_j) = \prod_{i=1}^{n} p(\lambda_j^{(i)})$ a prior $p(\lambda_j^{(i)})$. Based on the expression for the normal copula factor model in (2.3) in terms of the latent factors and errors, we are able to express the piecewise analog of the dynamic model developed in Creal and Tsay (2015), with pseudo-observations $Y_t^{(i)} = \Phi^{-1}(F_i(X_t^{(i)}|X_{1:t-1}^{(i)}))$ as

$$|\tau_{j+1} - \tau_j| \overset{i.i.d.}{\sim} \text{Geo}(\alpha) \qquad \text{for } \sum_{j=1}^{K} |\tau_{j+1} - \tau_j| \leq T \tag{2.5}$$

$$\lambda_j^{(i)} \overset{i.i.d.}{\sim} p(\lambda_j^{(i)}) \qquad \text{for } (i, j) \in \{1, ..., n\} \times \{1, ..., K\} \tag{2.6}$$

$$z_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1) \tag{2.7}$$

$$Y_t^{(i)}|\tau_{1:K}, z_t, \{\lambda_j^{(i)}\} \sim \mathcal{N}(\rho_j^{(i)} z_t, 1 - \rho_j^{(i)2}) \qquad \text{where } \rho_j^{(i)} = \frac{\lambda_j^{(i)}}{\sqrt{1 + \lambda_j^{(i)2}}} \tag{2.8}$$

for $i, t \in \{1, ..., n\} \times \{1, ..., T\}$ and $j \in \{1, ..., K\}$ is used to index segments. A conjugate Beta prior $p(\alpha) = \text{B}(\alpha; a, b)$ is placed on the parameter of $p(\tau_{1:K}|\alpha)$.

The posterior distribution is then

$$\pi(\tau_{1:K}, \{\lambda_j^{(i)}\}, z_{1:T}|\mathcal{Y}) \propto \left(\prod_{j=1}^{K}\prod_{t\in\mathcal{I}_j}\left(\prod_{i=1}^{n} p(y_t^{(i)}|z_t, \Lambda_j)p(z_t)\right)p(\Lambda_j)\right)p(\tau_{1:K}|\alpha)p(\alpha) \tag{2.9}$$

$$\text{where } p(y_t^{(i)}|z_t, \Lambda_j) = \mathcal{N}(y_t^{(i)}; \rho_j^{(i)} z_t, 1-\rho_j^{(i)2}) \quad \text{and} \quad p(z_t) = \mathcal{N}(z_t; 0, 1) \tag{2.10}$$

$y_t^{(i)}$ denotes a realisation of the pseudo-observation $Y_t^{(i)}$ and $\mathcal{I}_j := \{\tau_j + 1, ..., \tau_{j+1}\}$ denotes the times in segment j.

2.3 Inference

The following describes a novel Gibbs sampler to obtain approximate samples from the posterior distribution of changepoints, factors and factor loadings, conditional on the pseudo-observations. By recasting the inference problem for changepoints within a sequential framework we are able to exploit efficient particle methodology developed in Whiteley et al. (2011), Fearnhead (2006), Fearnhead and Liu (2007) to obtain samples from π.

For the product partition model of (2.4) efficient methods for inference of $\tau_{1:K}$, assuming θ is known, have been formulated in Fearnhead (2006), Fearnhead and Liu (2007) when it is possible to evaluate the predictive conditional distribution for any $s \leq t-1$ of

$$p(y_t|y_{s:t-1}, \theta) := \frac{\int p(y_{s:t}|\phi, \theta)p(\phi|\theta)d\phi}{\int p(y_{s:t-1}|\phi, \theta)p(\phi|\theta)d\phi} \tag{2.11}$$

where y_t denotes a realisation of the cross-section of n pseudo-observations $y_t := (y_t^{(1)}, ..., y_t^{(n)})$ and where $p(\phi|\theta)$ denotes a prior distribution over parameters ϕ specific to each segment. When θ is unknown, we are able to perform Bayesian inference using particle MCMC methods (PMCMC) Andrieu et al. (2010) and the bespoke inference routines provided in Whiteley et al. (2011). For the changepoint copula model of (2.9) we see that the factor loadings are segment specific, i.e. $\phi_j = \{\lambda_j^{(i)}\}_{i=1}^n$, while changepoint parameter α is a global parameter. As the factors are i.i.d., they could be either, however we treat them as global parameters, i.e. $\theta = \{z_{1:T}, \alpha\}$, for

the purposes of inference. We see that treating $z_{1:T}$ as a global parameter ensures that evaluation of the predictive likelihood $p(y_t|y_{s:t-1}, \theta)$ in (2.11) can be performed using n 1-dimensional integrals, rather than the required n-dimensional integrals that would be needed if $z_{1:T}$ we treated as segment specific parameters.

Indeed the dimensionality of the state space, $(\tau_{1:K}, \{\rho_j^{(i)}\}, z_{1:T}, \alpha) \in 2^T \times [-1, 1]^{nT} \times \mathbb{R}^T \times [0, 1]$, is exponentially increasing in T suggesting naive enumeration of the state-space is impossible. As a result, we develop a novel Gibbs sampler, iterating through conditionals $p(\tau_{1:K}|\theta, \mathcal{Y})$, $p(\phi_{1:K}|\tau_{1:K}, \theta, \mathcal{Y})$ and $p(\theta|\phi_{1:K}, \tau_{1:K}, \mathcal{Y})$, employing the changepoint routine of Whiteley et al. (2011). Broadly speaking, PMCMC kernels provide a means of using sequential Monte Carlo (SMC) methods to provide efficient high-dimensional proposals within larger Metropolis–Hastings schemes. A particular variant of PMCMC, particle Gibbs, provides a means to exploit SMC ideas to efficiently sample from a kernel leaving the conditional distribution of $p(\tau_{1:K}|\theta, \mathcal{Y})$ invariant after marginalising out $\phi_{1:K}$. We first of all introduce the sequential formulation and the resulting particle methodology assuming θ is fixed, before describing the resulting particle Gibbs sampler targeting π.

2.3.1 State-space model inference in changepoint problems

Following Fearnhead (2006), Fearnhead and Liu (2007), the following demonstrates that we are able to treat changepoint inference as a smoothing problem within a state-space model framework under a known parameter θ. We will then employ the particle Gibbs sampler of Whiteley et al. (2011) to sample from the posterior π, inferring $\tau_{1:K}$, θ and $\{\phi_j\}_{j=1}^K$ jointly.

To ease notation, we omit dependency of mass functions and densities on θ, treating it as fixed. The method relies on reparameterising a sequence of changepoint locations as follows. Defining $C_t = \max\{\tau_j \leq t - 1\}$ for $t \in \mathbb{N}$ as the location of the most recent changepoint prior to time t we see that for any t there is a one-to-one correspondence between $C_{1:T}$ and $\tau_{1:K}$ (an example of this reparameterisation is shown in Figure 2.1), so that we infer $p(C_{1:T}|\mathcal{Y})$.

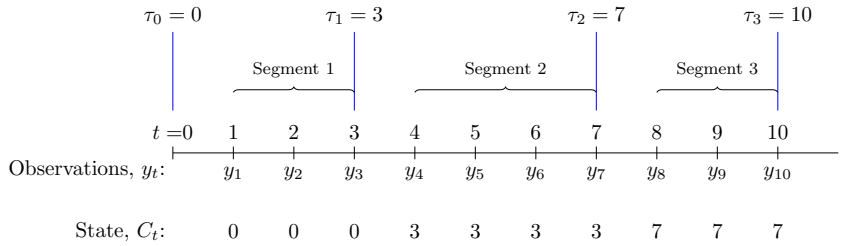


Figure 2.1: Example mapping between $\tau_{1:K}$ and $C_{1:T}$, where $C_{1:T}$ is given by $C_t = \max\{\tau_j \leq t-1\}$ for each t and $\tau_{1:K}$ is given by the ordered set $\tau_{1:K} = \{x : \exists x \in C_{1:T}\}$.

Recursive expressions for the filtered distribution $p(C_t|y_{1:t})$, can then be given in terms of the state space transition and observation distributions, $p(C_{t+1}|C_t)$ and $p(y_{t+1}|C_{t+1}, y_{1:t})$, which we provide as follows. Consider the event $\{C_t = s\}$ for some $s \in \{0, ..., t-1\}$. If we increment t by one then either $\{C_{t+1} = C_t = s\}$, i.e. the most recent changepoint prior to $t+1$ coincides with that prior to time t or $\{C_{t+1} = t\}$, i.e. there is an additional changepoint introduced at time t. Let $\bar{d}_t \geq 1$ denote the size of the 'current' segment, i.e. the size of the unique segment satisfying $t \in \{\tau_{j-1}+1, ..., \tau_j\}$. In this case we then have the following conditional probability of the most recent changepoint prior to time $t+1$ coinciding with that prior to time t

$$\begin{aligned} p(C_{t+1} = C_t|C_t = r) &= \mathbb{P}[\bar{d}_{t+1} \geq t-r+1|\bar{d}_{t+1} \geq t-r] \\ &= \frac{\mathbb{P}[\bar{d}_{t+1} \geq t-r+1]}{\mathbb{P}[\bar{d}_{t+1} \geq t-r]} = \frac{1-G(t-r)}{1-G(t-r-1)} \end{aligned} \tag{2.12}$$

where $G(s) := \sum_{i=1}^{s} g(i) = \mathbb{P}[d \leq s]$ denotes the cumulative distribution associated with g, the prior over segment lengths. Equality on 2.12 follows as $\left\{\bar{d}_{t+1} \geq t-r+1\right\} \supset \left\{\bar{d}_{t+1} \geq t-r\right\}$. The state transition can then be computed according to

$$p(C_{t+1} = C_t|C_t = r) = \begin{cases} p(C_{t+1} = C_t|C_t = r) & \text{if} \quad s = r < t+1 \\ 1 - p(C_{t+1} = C_t|C_t = r) & \text{if} \quad s = t+1 \end{cases} \tag{2.13}$$

where we note that in the case of a geometric distribution such a probability is given trivially as $\frac{1-G(t-r)}{1-G(t-r-1)} = 1 - p_G$, (a fail in a sequence of Bernouilli trials). This allows the filtering recursions for state-space models to be obtained as Doucet

and Johansen (2011)

$$p(C_{t+1} = s|y_{1:t}) = \sum_{r=0}^{t-1} p(C_{t+1} = s|C_t = r)p(C_t = r|y_{1:t}) \tag{2.14}$$

$$p(C_{t+1} = s|y_{1:t+1}) \propto p(y_{t+1}|C_{t+1} = s, y_{1:t})p(C_{t+1} = s|y_{1:t}) \tag{2.15}$$

where $p(y_{t+1}|C_{t+1} = s, y_{1:t}) = p(y_{t+1}|y_{s+1:t})$ denotes the predictive density of observation y_{t+1} conditional on $y_{s+1:t}$. We see that $p(y_t|C_t = s, y_{1:t-1})$ corresponds to a time inhomogeneous observation distribution, conditional on the state at observation t. The constant of proportionality in (2.15) can be computed exactly by summing over $C_{t+1} \in \{0, ..., t\}$.

Computing the filtering recursions exactly in (2.14) and (2.15) requires a summation of $\mathcal{O}(t)$ for each t, as a result of the need to marginalise over the linearly increasing support of C_t. Performing these recursions exactly induces an $\mathcal{O}(T^2)$ computational cost, which can be prohibitively large for long time series, thereby motivating the particle methods Fearnhead (2006), Fearnhead and Liu (2007). Described in more detail in Appendix 2.7.1, Fearnhead (2006) provides an efficient means to sample $\tau_{1:K} \sim p(\tau_{1:K}|\theta, \mathcal{Y})$ while Whiteley et al. (2011) provides an efficient means to sample from a kernel leaving $p(\tau_{1:K}|\theta, \mathcal{Y})$ invariant, known as 'conditional SMC' through considering an extended target distribution including all random variables used to perform the particle methods developed in Fearnhead (2006), Fearnhead and Liu (2007). The resulting resampling scheme is highly optimised, providing a small modification of the optimal resampling scheme developed in Fearnhead (2006). The resulting procedure provides an efficient means to obtain samples from the joint posterior of $\tau_{1:K}$, $\phi_{1:K}$ and θ.

2.3.2 Particle Gibbs sampler

We propose the following novel particle Gibbs inference scheme, exploiting among other properties the model's sequential and factorial structure. The scheme presented here is a form of blocked Gibbs, asymptotically targeting the posterior $\pi(\tau_{1:K}, \{\lambda_j^{(i)}\}, z_{1:T}|\mathcal{Y})$. A single Gibbs sweep targeting π is provided in Algorithm

1, where we use $U\{-1,1\}$ to denote a uniform distribution on $\{-1,+1\}$. We discuss each step as follows.

Algorithm 1 Particle Gibbs for factor copula changepoint inference

Gibbs sweep leaving π invariant

Input: Current values of changepoint locations, factors, and segment length parameter $(\tau_{1:K}, z_{1:T}, \alpha)$

1. Conditional SMC for changepoint locations, collapsing out factor loadings $\tau'_{1:K} \sim$ cSMC$(\cdot|z_{1:T}, \tau_{1:K})$ (Algorithm 3, Appendix 2.7.1)
2. Sample changepoint parameters $\alpha' \sim p(\alpha|\tau'_{1:K})$
3. Sample factor loadings using $\rho'_{1:K} \sim p(\rho_{1:K}|z_{1:T}, \tau'_{1:K})$ (approximate sampling)
4. Sample factors $z'_{1:T} \sim p(z_{1:T}|\rho'_{1:K})$ (closed form expression)
5. (*) For $j \in \{1, ..., K\}$

$$B_j \sim U\{-1,1\} \quad z''_t = B_j z'_t \quad \text{for} \quad t \in \mathcal{I}_j$$

Step 1 requires a cSMC step targeting the full conditional of changepoint locations and changepoint model parameters, as given in Algorithm 3. For the copula factor model this requires marginalising out latent factor loadings $\rho_{1:K}$ as in (2.11) to evaluate posterior predictives for observations y_t conditional on $y_{r:t-1}$ for $r < t$. In this context this can be achieved through

$$p(y_t|y_{r:t-1}, \theta) = \prod_{i=1}^{n} p(y_t^{(i)}|y_{r:t-1}^{(i)}, z_{r:t-1}) \tag{2.16}$$

$$\text{where} \qquad p(y_t^{(i)}|y_{r:t-1}^{(i)}, z_{r:t-1}) = \frac{\int \prod_{s=r}^{t} \mathcal{N}(y_s^{(i)}; \rho^{(i)} z_t, 1-\rho^{(i)2}) \pi_\rho(\rho^{(i)}) d\rho^{(i)}}{\int \prod_{s=r}^{t-1} \mathcal{N}(y_s^{(i)}; \rho^{(i)} z_t, 1-\rho^{(i)2}) \pi_\rho(\rho^{(i)}) d\rho^{(i)}}$$

where $\pi_\rho(\rho^{(i)})$ represents the induced prior density on the scaled factor loadings, i.e. the density of $\rho_j^{(i)} = \frac{\lambda_j^{(i)}}{\sqrt{1+\lambda_j^{(i)2}}}$ for $\lambda_j^{(i)} \sim p(\lambda_j^{(i)})$ which may be given explicitly. We see from (2.16) that conditioning on $z_{1:T}$ results in only a series of n 1-dimensional integrals, allowing standard numerical quadrature techniques to be employed marginalising out each $\rho^{(i)}$ over the interval $[-1,1]$.

Step 2 is a standard Gibbs update for a binomial distribution with conjugate prior for α. In the absence of prior conjugacy on the factor loadings, Step 3 is performed using either slice sampling or approximating the inverse CDF using

quadrature on the full conditional of each $\rho_j^{(i)}$. Due to the conditional independence of each $\rho_j^{(i)}$ given the factors in segment $\mathcal{I}_j$, the problem reduces to a one-dimensional sampling scheme with unnormalised full conditional given by

$$p(\rho_{ij}|z_t : t \in \mathcal{I}_j) \propto \prod_{t \in \mathcal{I}_j} \mathcal{N}(y_t^{(i)}; \rho_j^{(i)} z_t, 1 - \rho_j^{(i)2}) \pi_\rho(\rho_j^{(i)})$$

for $i \in \{1, ..., n\}$. The appearance of $\rho_j^{(i)}$ (and also for $\lambda_j^{(i)}$) in both the variance and mean of the normal density induces a non-logconcave distribution, precluding the use of adaptive rejection sampling. The inference scheme deals trivially with missing data through an additional imputation step. Step 4 is conditionally normal, where conditional on a segmentation, factor loadings and pseudo-observations, the full conditional for factors can be formed through the conjugacy of the standard normal prior on z_t then for $t \in \mathcal{I}_j$ we have that $p(z_t|\rho_j^{(1:n)}) \propto \mathcal{N}(z_t; 0, 1) \prod_{i=1}^n \mathcal{N}(y_t^{(i)}; \rho_j^{(i)} z_t, 1 - \rho_j^{(i)2})$.

For the final step, marked $(*)$, we introduce an additional optional step that can be used to significantly boost convergence to stationarity of the resulting chain for the single factor model. As a result of the factor structure there is an inherent unidentifiability in the model. One option, employed in Murray et al. (2013), Creal and Tsay (2015) is to impose identifiability restrictions on the factor loadings to converge to a unique posterior. Alternatively, we consider inference on the unidentifiable model and note that although $z_{1:T}$ and $\{\rho_j^{(i)}\}_{i=1}^n$ may not be uniquely identifiable, the resulting correlation matrix associated within each segment is. For $j \in \{1, ..., K\}$ let $\rho_j := (\rho_j^{(1)}, ..., \rho_j^{(n)})'$, then we have that $R_j := \rho_j \rho_j' - \text{diag}(\rho_j) + I$. We formalise the validity of this move in the following proposition.

Proposition 1. *For symmetric factor loading priors $\pi_\rho(\rho)$, then conditional on a segmentation and factor loadings $z_t = B_j z_t$ and $\rho_j^{(i)} = B_j \rho_j^{(i)}$ with B_j uniformly distributed on $\{-1, 1\}$ leaves π invariant.*

Proof. For any segment $\mathcal{I}_j$ we introduce the binary random variables $B_j \in \{-1, 1\}$ and define $\bar{z}_t := B_j z_t$ and $\bar{\rho}_j^{(i)} := B \rho_j^{(i)}$ so that unconditionally, for $\mathcal{I} \sim p(\mathcal{I})$ and $B_j \sim \mathcal{U}\{-1, 1\}$, under the prior $\bar{z}_t \overset{d}{=} z_t$ and $\bar{\rho}_j^{(i)} \overset{d}{=} \rho_j^{(i)}$. Let $a := \tau_j + 1$, $b := \tau_{j+1}$

denote the start and end points of segment j. By inspection of the posterior π (2.9) we have that conditional on a segmentation $\mathcal{I}$ that the density over $z_{a:b}$ and $\rho_j^{(1:n)}$ is given by

$$p(z_{a:b}, \rho_j^{(1:n)}|y_{a:b}) \propto \prod_{t=a}^{b}\left[\prod_{i=1}^{n}\mathcal{N}(y_t^{(i)}; \rho_j^{(i)}z_t, 1-\rho_j^{(i)2})\pi_\rho(\rho_j^{(i)})p(z_t)\right]$$

which is marginally (after integrating out $B_{1:K}$) equivalent to

$$\begin{aligned} p(z_{a:b}, \rho_j^{(1:n)}, B_j|y_{a:b}) &\propto \prod_{t=a}^{b}\left[\prod_{i=1}^{n}\mathcal{N}(y_t^{(i)}; \bar{\rho}_j^{(i)}\bar{z}_t, 1-\rho_j^{(i)2})\pi_\rho(\rho_j^{(i)})p(z_t)\right]p(B_j) \\ &= \prod_{t=a}^{b}\left[\prod_{i=1}^{n}\mathcal{N}\left(y_t^{(i)}; \left(B_j\rho_j^{(i)}\right)(B_jz_t), 1-\rho_j^{(i)2}\right)\pi_\rho(\rho_j^{(i)})p(z_t)\right]p(B_j) \end{aligned}$$

As a result, conditional on $z_{a:b}$ and ρ_j we are able to sample $B_j \sim U\{-1, 1\}$ without changing the value of the likelihood $\mathcal{N}\left(y_t^{(i)}; \left(B_j\rho_j^{(i)}\right)(B_jz_t), 1-\rho_j^{(i)2}\right)$ and the resulting move leaves π invariant. □

As a result, we are able to introduce an additional Gibbs move to switch the signs of the factors $z_{1:T}$ within each of the segments. From Algorithm 1, we see that there is no need to flip the signs of $\rho_j^{(1:n)}$ as they are integrated out at the next iteration. We illustrate the importance of this sign switching in the next section.

2.4 Applications

The following experiments illustrate properties of the model and resulting sampler on simulated data and two real datasets. The real data spans the financial crisis of 2008 to provide some notion of a 'ground truth' over which to test the methodology, in that we would typically expect a structural break to have occurred within this time perior. We compare the resulting changepoints to BBC news articles providing an analysis over this period.

2.4.1 Synthetic data

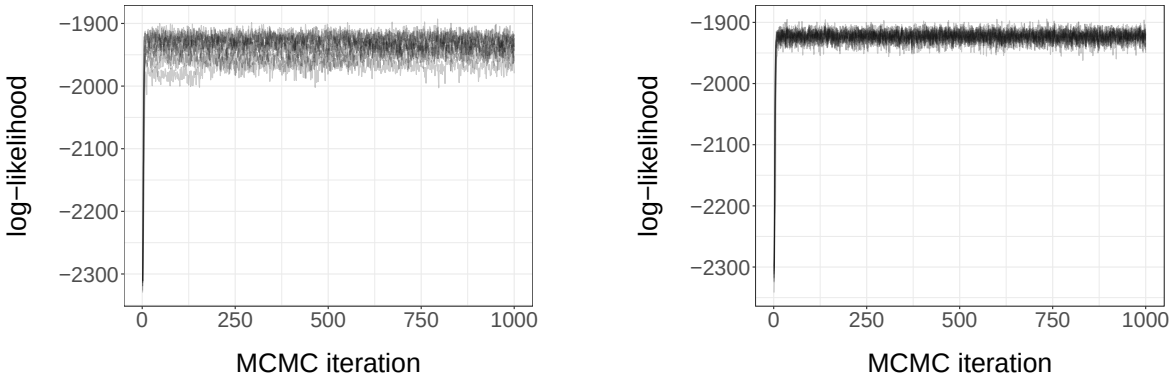
We illustrate aspects of the model and inference using synthetic data demonstrating firstly the speed up afforded by the sign switching and secondly properties of inference arising as the number of time series, n, increases.

2.4.1.1 Comparison of initialisation with and without label switching

We consider convergence of the chains comparing the kernels with and without sign switching of the latent factors $z_{1:T}$, for an $n = 15$ dimensional example, with $T = 100$ and where there is a single changepoint at $t = 50$. . To impose identifiability constraints in the absence of sign switching, the inference algorithm restricts the first component of the n-dimensional ρ_j for each $j \in \{1, ..., K\}$, conditional on a realisation of changepoints $\tau_{1:K}$, to be positive. In particular we focus on the initial convergence of the resulting chain. To help monitor convergence of the high-dimensional space of the resulting sampler we examine the estimate of the marginal log-likelihood $\log \hat{p}(\mathcal{Y}, \tau_{1:K}, z_{1:T}, \theta)$, a quantity that can readily be computed from the output of Algorithm 1 after marginalising out latent factor loadings. It is an estimate in so far as for $N = T$ the estimate coincides with the true marginal log-likelihood (in this case the particle approximation is exact in the sense that the recursions are computed exactly, without Monte Carlo error).

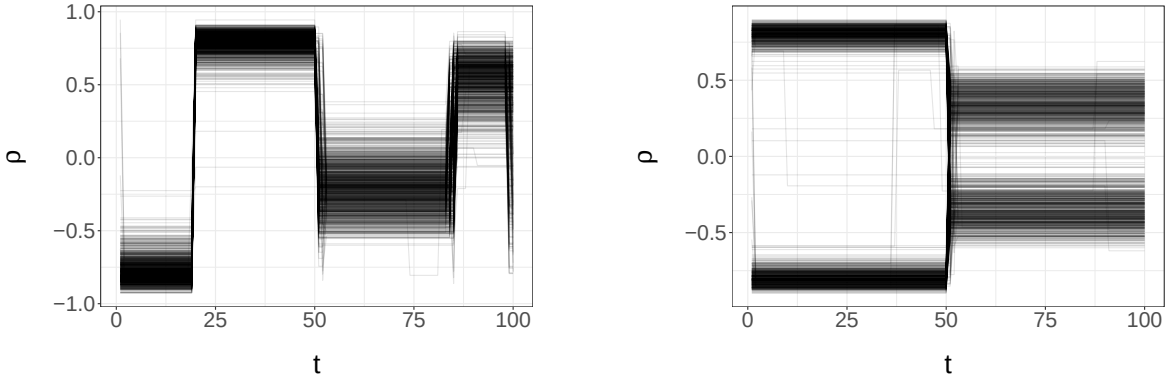
Each chain was initialised according to $z_{1:T} \sim \mathcal{N}(0, 1)$, $p_G = 0.5$ and $\tau_{1:K}$ such that changepoints occur at equally spaced intervals of length 10. We consider the initial convergence when initialised far from stationarity for 10 chains, each of length 1,000 with $N = 10$ particles and plot each chain's marginal log-likelihood towards a limiting value with and without label-switching in Figure 2.2 over this initial period.

We see from Figure 2.2b firstly that when the additional sign switching of $z_{1:T}$ is introduced convergence to a stationary value occurs rapidly within the first 100 iterations. In comparison, Figure 2.2a shows that while for some chains convergence can be rapid, in the worst case scenario convergence to stationarity can be significantly impaired. We examine in more detail one of the chains in which slow convergence occurs. In this case, Figure 2.3 plots posterior samples of ρ_2 (the second component in $\rho_{1:K}$) and Figure 2.4 plots the posterior distribution over changepoint locations $\tau_{1:K}$ with and without switching. We see firstly, the bimodality introduced through the lack of identifiability of the latent factors. Secondly, we see that without the switching, the chain struggles to jump between modes and, as such, the sampler introduces additional changepoints when initialised far from stationarity.



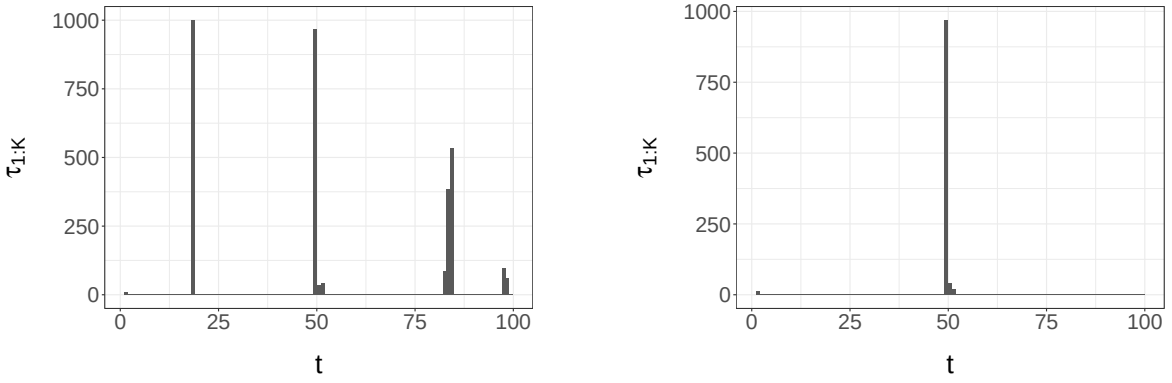
(a) Without switching **(b)** With switching

Figure 2.2: Marginal log-likelihood functions against MCMC iteration. Each plot shows the lowest, median and largest values of $\log p(\mathcal{Y}, \tau_{1:K}, z_{1:T}, \theta)$ over the 10 simulated chains.



(a) Without switching **(b)** With switching

Figure 2.3: Posterior samples of ρ_2.



(a) Without switching **(b)** With switching

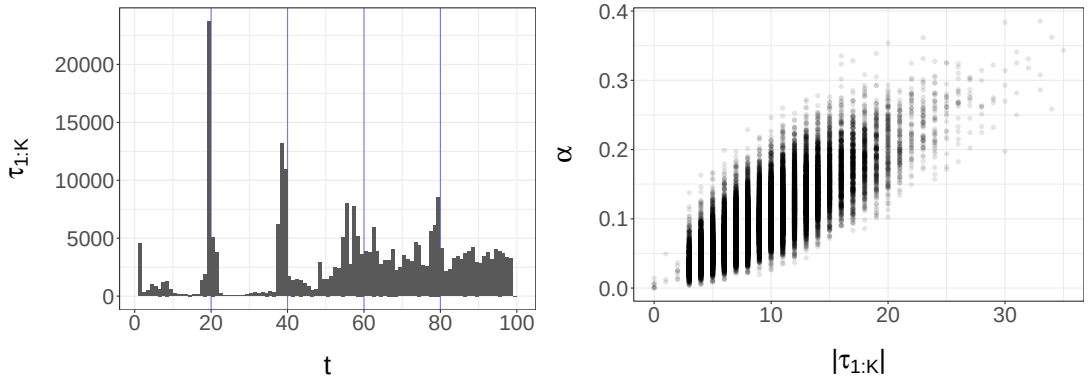
Figure 2.4: Posterior samples of $\tau_{1:K}$ (lower). Rescaled factor loadings were thinned to 100 samples over the chains of length 10,000 though no thinning was performed to estimate the posterior over $\tau_{1:K}$.

2.4.1.2 Posterior concentration for increasing n

We next consider how increasing n can help reduce the uncertainty in the changepoint locations. In general, we expect the uncertainty in changepoint locations to be strongly dependent on the uncertainty in parameter estimates for each segment, in that were each segment's parameters known with certainty, this would imply the changepoint locations are known also. For $T < \infty$ (with $n < \infty$ or $n \to \infty$) we do not expect posterior concentration as uncertainty in the geometric parameter α (see the model description in Equations 2.5-2.8) will only vanish as $T \to \infty$. As a result, we illustrate experimentally how n affects the uncertainty in the changepoint locations. We consider $T = 100$ and vary $n \in \{5, 10, 15\}$, fixing the number of particles $N = 10$. The data was generated such that the true location of changepoints was fixed at $t \in \{20, 40, 60, 80\}$, with a total of 4 changepoints.

We show the posterior over changepoints for each of these values in Figures 2.5 and 2.6, where we see the concentration of the changepoint locations. For each value of n, a chain was run for $30,000$ iterations. Additionally, Figure 2.5b plots the posterior distribution of α against $|\tau_{1:K}|$ (the number of changepoints) for $n = 5$. We see that there is a strong dependency between the two with a high degree of correlation. Such a result provides further justification for a Bayesian analysis for the multiple changepoint problem. The prior on factor loadings was taken to be the standard normal distribution. Under certain conditions we see that for factor models without structural changes that the MLE for factor loadings are consistent with the true factor loadings as the limit as both n and T tend to infinity, when the number of factors is unknown a priori. Finally, it may be possible to develop methods for which the infill asymptotic holds for Du et al. (2016), though we leave this for future research.

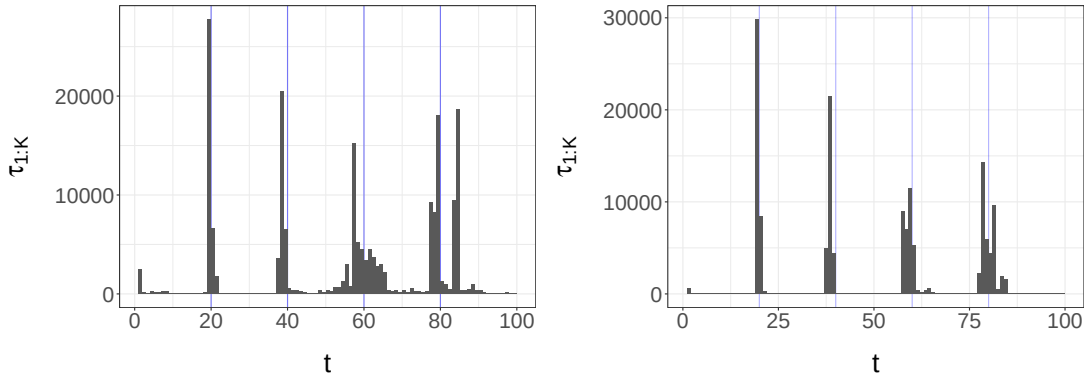
Finally, we plot the autocorrelation plots for two test functions - we note that owing to the medium- to high-dimensional structure of the problem it is not immediately obvious which functions to test. We elect to choose $h_1 = |\tau_{1:K}|$, the number of changepoint locations to summarise the mixing over the $\tau_{1:K}$ co-ordinates and also $h_2 = \sum_{t=1}^{T} |z_t|$ to summarise mixing over the latent factors. The choice



(a) Distribution of $\tau_{1:K}$, $n = 5$

(b) Scatter plot of p_G against $|\tau_{1:K}|$, the number of changepoints, for $n = 5$

Figure 2.5: Results for $\tau_{1:K}$ with $n = 5$



(a) Distribution of $\tau_{1:K}$, $n = 10$

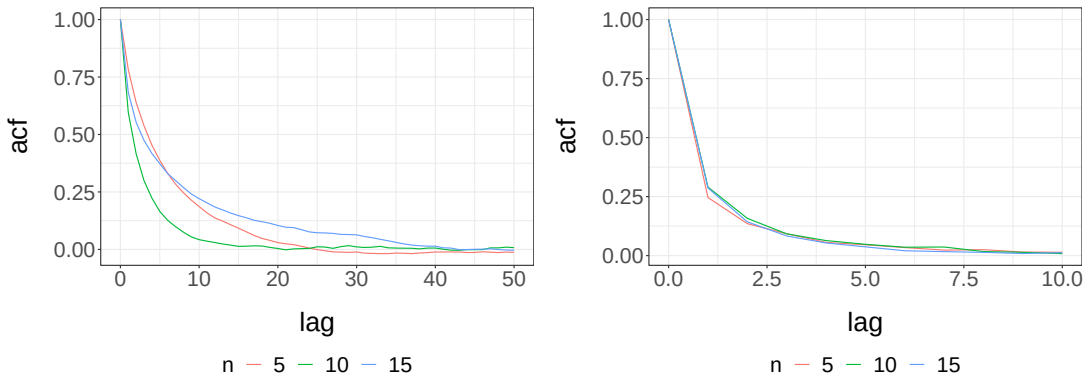
(b) Distribution of $\tau_{1:K}$, $n = 15$

Figure 2.6: Posterior distributions of $\tau_{1:K}$ with $n = 10$ and 15.

to take the absolute value of the latent factors was motivated by the fact that the inference algorithm randomly switches the sign of the latent factors. We see in Figure 2.7a the autocorrelations are typically dying out after around 50 iterations for the h_1 and after only a few iterations for h_2 suggesting the algorithm mixes more quickly in the latent factor co-ordinate than it does over the changepoint locations.

2.5 Structural breaks during 2008 financial crisis

We examine the short-term structural changes apparent during the 2008 financial crisis. In particular, we consider the 10 largest (as measured by initial market capitalisation) of the S&P 500 from the perior 1^{st} January 2007 to 1^{st} January



(a) Autocorrelation for the number of changepoints $h_1 = |\tau_{1:K}|$

(b) Autocorrelation for the sum $h_2 = \sum_{t=1}^{T} |z_t|$

Figure 2.7: Autocorrelation plots for two test functions, h_1 and h_2, of the posterior distribution

2009, leading to in total 503 observations. Stocks pressent in the S&P 500 were obtained from the Yahoo database.

2.5.1 Inference

For each stock we fit a parametric GARCH model to each of the stocks using the `rugarch` statistical package for R. In the following we consider two conditional distributions, i.e. let $X_t^{(i)}$ denote the log-returns of stock i at time t, then we are able to specify the conditional distribution of $X_t^{(i)}|X_{1:t-1}^{(i)}$. We refer the reader to Ghalanos (2017) for more details. Using the flexibility afforded by the `rugarch` package we set this to 1) a skew Student-T distribution and 2) a generalised hyperbolic distribution. For each of the 10 stocks considered the we perform a Shapiro-Wilks test on pseudo-observations, testing to ensure they are approximately standard normal. In this case the median p-value was 0.61 (min. 0.13 and max. 0.96) and 0.77 (min. 0.053 and max. 0.97) for the two distributions respectively, suggesting an ample fit for the marginal models.

The sampler was run with $N = 10$ particles and 100 quadrature points for the numerical integration and run for 100,000 iterations, discarding the first 20,000 iterations as burn-in. The choice in burn-in was informed by trace plots of the number of changepoints showing a certain stickiness for the number of changepoints

up until a large number of iterations. It is noted that the naive implementation did not take account the recursive nature of the quadrature which would provide an additional point of speed up.

2.5.1.1 Priors on factor loadings

Experiments on stock data considered two prior distributions on segment factor loadings corresponding to the two priors detailed in Murray et al. (2013). Both priors were parameterised in terms of the unscaled factor loadings, $\lambda_j^{(i)} \sim p(\lambda_j^{(i)})$, with density on scaled factor loadings given by a change of variables $\rho_j^{(i)} = \frac{\lambda_j^{(i)}}{\sqrt{1+\lambda_j^{(i)2}}}$. The two considered were

1. Standard normal $p(\lambda_j^{(i)}) = \mathcal{N}(\lambda_j^{(i)}; 0, 1)$
2. Generalised double Pareto (shrinkage prior) $p(\lambda_j^{(i)}) = \text{GDP}(\lambda_j^{(i)}; 3, 1)$ where the GDP density is given as

$$\text{GDP}(\lambda_j^{(i)}; \alpha, \beta) := \frac{\alpha}{2\beta}\left(1 + \frac{\left|\lambda_j^{(i)}\right|}{\beta}\right)^{-(1+\alpha)}$$

The former is considered due to its routine employment in normal factor models and additionally more closely resembles the latent Gaussian process employed in Creal and Tsay (2015). The latter prior is considered due to its recommendation in Murray et al. (2013) as a means to ensure parsimony, with the resulting model favouring fewer heavily loaded factors. Furthermore, for a single factor normal copula model, there is a tendency for the probability mass of scaled factors to distributed largely around ± 1, whereas GDP priors provide a more even spread. The factor structure makes explicit specification of uninformative priors non-trivial. Both satisfy the conditions of Proposition 1.

2.5.2 Results

The resulting posterior distribution was compared against a news article provided by The Guardian online Kingsley (2012). Such a procedure enables a certain ground truth to be ascertained, in that we may expect structural changes to occur around the period of major news events. The results are provided in Figure 2.8 where we plot the

posterior distribution of changepoint locations for each of the chain. As well as this, the posterior distribution of the number of changepoint locations is shown also. We see immediately that the number of changepoints is generally larger for the normal prior over changepoint locations. Furthermore, we see that there is a relatively small difference between the posterior distribution over the number of changepoints for the two marginal models, suggesting a degree of robustness to this choice.

The bottom line of the figure shows key events in over the financial crisis as suggested by The Guardian news article. The events labelled 1-7 correspond to perhaps the most significant events in the financial crisis, as follows

1. 7^{th} September 2008 - Bailout Fannie Mae and Freddie Mac
2. 15^{th} September 2008 - Lehmann brothers files for bankruptcy
3. 17^{th} September 2008 - HBOS rescued by LLoyds TSB
4. 21^{st}- 29^{th} September 2008 - Two US banks collapse
5. 30^{th} September 2008 - Ireland pledges to underwrite banking system
6. 7-8^{th} October 2008 - Three of Icelands banks collapse
7. 13^{th} October 2008 - UK bailout of major banks including RBS, LLoyds TSB and HBOS

We see that over this period, ranging approximately from the 7^{th} September 2008 to 13^{th} October 2008 that for each of the settings, i.e. each prior distribution and each conditional distribution, there are significant spikes in the distribution of changepoint locations over this period. More specifically, for the sparser GDP prior, there is a significant spike towards the middle of this perior, and with the normal prior there is approximately three spikes in changepoint locations.

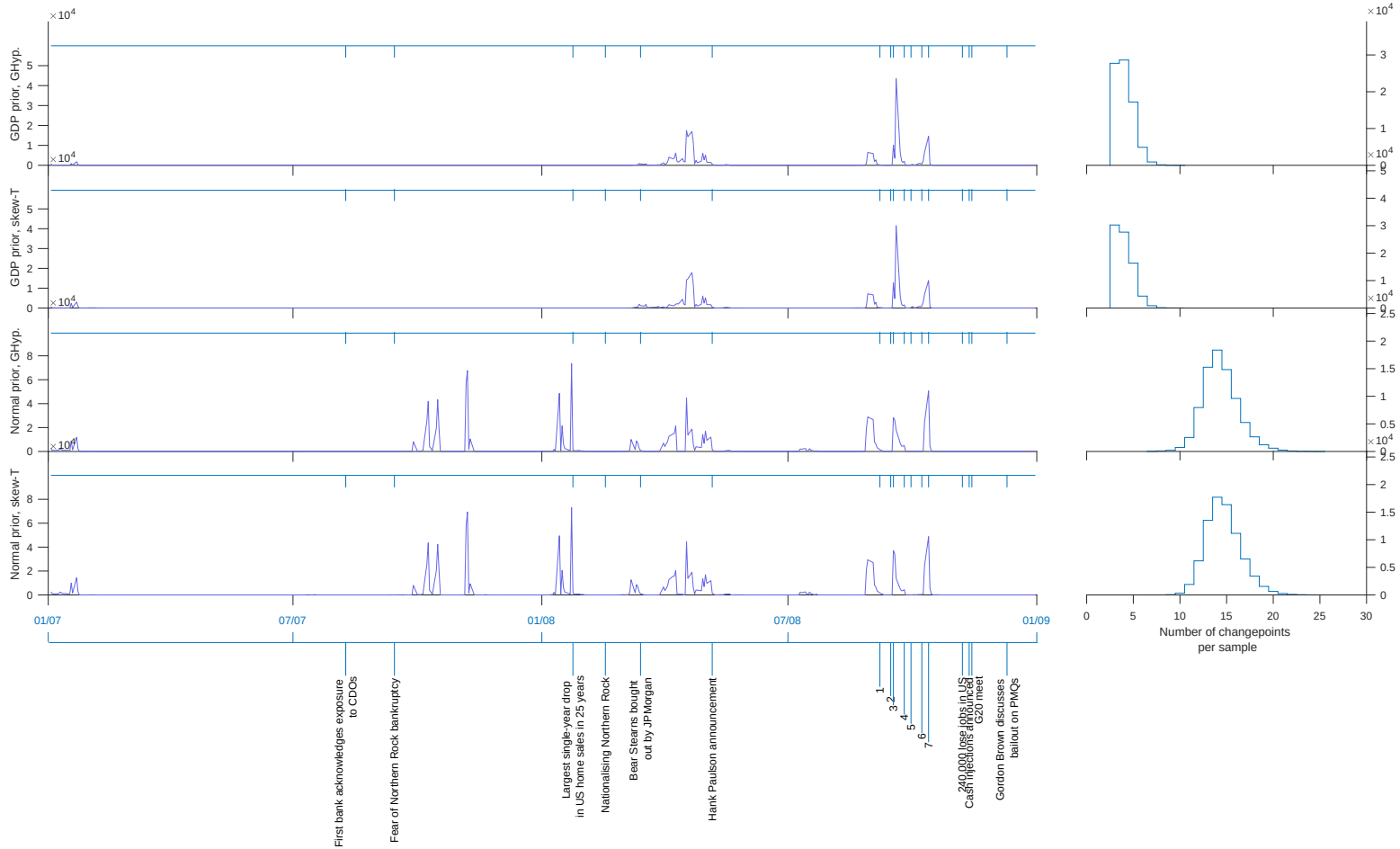
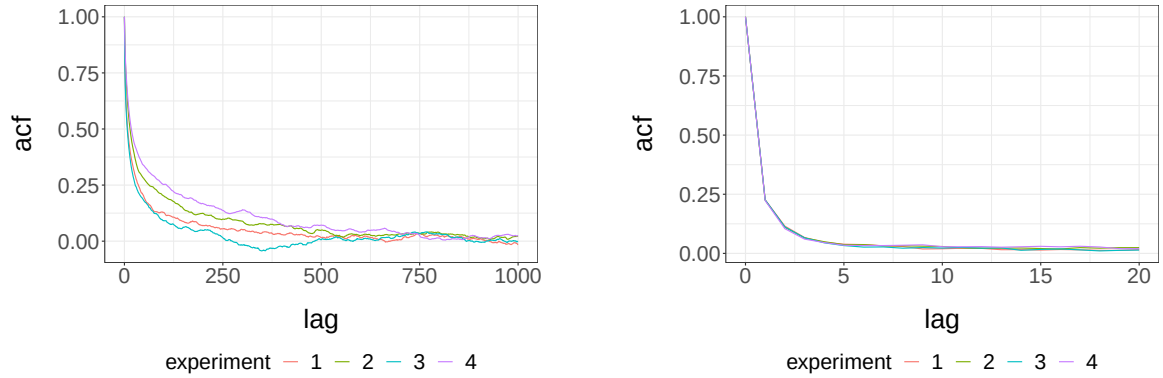


Figure 2.8: Posterior samples of changepoint locations over the 2008 financial crisis.

Finally, we examine the computational performance of the resulting sampler. We see that depending on the test function, mixing can be relatively slow as demonstrated by the autocorrelation plots in Figure 2.9. We examine the four experiments above, numbering them as follows

1. Normal prior, generalised hyperbolic distribution for marginal model
2. GDP prior, generalised hyperbolic distribution for marginal model
3. Normal prior, skew Student-T distribution for marginal model
4. GDP Prior, skew Student-T distribution for marginal model.



(a) Autocorrelation plots for number of changepoints for test function $|\tau_{1:K}|$.

(b) Autocorrelation plots for number of changepoints for test function $\sum_{t=1}^{T} |z_t|$.

Figure 2.9: Experiment 1) Normal prior, GHYP 2) GDP prior, GHYP 3) Normal prior SSTD 4) GDP Prior SSTD.

We plot two test functions of the high-dimensional posterior. Firstly, we plot the number of changepoints present in each sample (Figure 2.9a) and secondly, a test function of the latent factors $\sum_{t=1}^{T} |z_t|$ where as before, the absolute value of the latent state is selected to account for the fact that at each sampling iteration the sign of the latent factor may be flipped while still preserving the correct stationary distribution (Figure 2.9b). We see that in Figure 2.9a there is a high-degree of autocorrelation present in the number of changepoints suggesting the chain may exhibit a certain stickiness when moving between changepoint configurations of different size. We compare this to the autocorrelation of Figure 2.9a whereby the autocorrelations die out relatively quickly. Plausibly, the number of particles used in the experiment could be increased to reduce the autocorrelation, though we

do not expect the autocorrelation to become arbitrarily small as N increases and this would inevitably come at the cost of fewer MCMC iterations, so potentially higher Monte Carlo variance.

2.6 Conclusion

The above model and inference routine provides novel methodology for estimating the structural changes in correlations between sets of time series. The methodology builds on dynamic copula work in medium- to high-dimensions resulting in a flexible changepoint inference routine admitting a wide class of marginal distributions for each series. Positive results were obtained on a set of stocks, identifying with accuracy significant events within the 2008 global financial crisis as reported by the Guardian newspaper.

A number of possible extensions exist both for the model and inference routine. Firstly, normal copula models have been criticised in the financial community for their light tails and do not necessarily represent the true correlation structure between assets. Furthermore, the single-factor nature of the model employed above may be overly restrictive for large portfolios, though it is noted that the dynamic copula model of uses a single factor on a portfolio of 200 credit default swaps and equities.

One of the limitations of such an approach is that within any one segment, the posterior density of ρ_i will collapse to a delta function as the length of the segment increases. Therefore a revision would be needed to employ an adaptive quadrature rule, however this was unneeded for the above implementation. Secondly, extending the current model to multiple factor models for the normal copula has significant implications on the inference scheme. While not explored in detail a possible approach would be to approximate the PMCMC scheme for changepoints developed above using the approximate Rao-Blackwellised particle filter for non-conjugate priors developed in Johansen et al. (2012). Difficulties, however, would involve the adaptation of the optimal sampling procedure in Fearnhead and Liu (2007) to such an application. In terms of the inference, a procedure to circumvent the two stage

inference procedure (fitting each marginal and then fitting the joint copula) may also be valuable. The key difficulty in this, however, is having to evaluate PITs using models for marginals while simultaneously estimating those models.

2.7 Appendix

2.7.1 Particle inference for changepoint problems

The following describes in moderate detail the particle methods used to construct the conditional SMC kernel leaving $p(\tau_{1:K}|\theta, \mathcal{Y})$ invariant. We begin through a description of the particle methods used for inference with fixed θ then describe in detail the conditional SMC update.

2.7.1.1 Filtering recursions and backward sampling

Originating in Fearnhead (2006), a computationally efficient approximation of these recursions is provided by a particle approximation to $p(C_t|y_{1:t})$, where at time t the probability mass function $p(C_t|y_{1:t}) : \{1, .., t\} \to \mathbb{R}^+$ is approximated by N weighted particles. At each iteration of the algorithm, operations are performed on each support point of the filtering distribution. Without the particle approximation, this is of $\mathcal{O}(t)$ for each t. As a result, a complete sweep of the recursions induces a computational cost that is $\mathcal{O}(T^2)$. For large datasets this is prohibitively expensive. As such, a particle approximation of the filtering distribution incurs a cost of $\mathcal{O}(N)$ for each t, with a complete cost that is $\mathcal{O}(NT)$, where the user is able to control the quality of the approximation N. One of the appeals of this methodology is that for $N = T$ then the exact recursions are recovered and no approximation is introduced - though we are back into the quadratic cost with T.

The particle approximation is propagated as follows, though we refer the reader to Whiteley et al. (2011) for more details. At each iteration t, $p(C_t|y_{1:t})$ is approximated by an empirical probability distribution $\hat{p}(x_t|y_{1:t}) := \sum_{x_t \in \mathbf{C}_t} W_t(x_t)\delta_{x_t}$ where $\mathbf{C}_t$ is used to denote $\{x_t^i : W_t^i > 0\}$, the random support of $\hat{p}(x_t|y_{1:t})$. Conditional on a realisation of $\mathbf{C}_t$, the elements are therefore deterministic and following Whiteley et al. (2011) we denote them in the lower case. Additionally, we have that $W_t(x_t) = \frac{\bar{W}_t(x_t)}{\sum_{x_t \in \mathbf{C}_t} \bar{W}_t(x_t)}$ where $\bar{W}_t(x_t)$ denotes an unnormalised weighting function of particle x_t.

Combining this approximation with (2.14) and (2.15) suggests the following expression for the filtering approximation $\hat{p}(x_{t+1}|y_{1:t+1})$ given the filtering approximation $\hat{p}(x_t|y_{1:t})$Fearnhead and Liu (2007)

$$\bar{W}_{t+1}(x_{t+1}) = p(y_{t+1}|y_{s+1:t})\frac{1-G(t-x_t)}{1-G(t-x_t-1)}W_t(x_t) \qquad \text{if } x_{t+1} = x_t < t \tag{2.17}$$

$$\bar{W}_{t+1}(x_{t+1}) = p(y_{t+1})\sum_{x_t^r \in \mathbf{C}_t}\frac{G(t-x_t^r)-G(t-x_t^r-1)}{1-G(t-x_t^r-1)}W_t(x_t^r) \qquad \text{if } x_{t+1} = t \tag{2.18}$$

$$\hat{p}(x_{t+1}|y_{1:t+1}) \propto \sum_{x_t \in \mathbf{C}_t \cup \{t\}} \bar{W}_{t+1}(x_{t+1})\delta_{x_{t+1}}$$

We have that (2.17) corresponds to the propagation of each existing particle and (2.18) corresponds to the introduction of a new particle with value $x_{t+1}^i = t$. After reweighting each particle, a resampling step is employed to remove particles with low weights. Reducing the support of $\hat{p}(x_t|y_{1t})$ as t increases ensures the number of required summations does not grow with t and $|\mathbf{C}_t| \leq N$. It is stressed however, that the reweighting is a purely deterministic transformation of the previous weighted particle set. As such, the only source of randomness occurs as a result of the resampling at each iteration.

As the space is discrete, the highly efficient stratified optimal resampling scheme of Fearnhead (2006) ensures particles are not duplicated after resampling, unlike standard particle filter resampling schemes. The resampling thereby provides a means to approximate $\hat{p}(x_t|y_{1t})$ with a new distribution $\hat{p}'(x_t|y_{1t})$. Importantly, if $N_t := |\mathbf{C}_t|$ denotes the size of the support at iteration t, the resampling provides a means to ensure that $\hat{p}'(x_t|y_{1t})$ is supported on R_t support points, with $R_t < N_t$. We refer the reader to Fearnhead (2006) and Whiteley et al. (2011) for more details on the stratified optimal resampling scheme which we adopt in the following.

As a result, even if t is increasing we are able to mitigate the introduction of a new particle at $x_{t+1} = t$ by resampling $N-1$ particles from N. The technique can provide large computational savings if, for example, the number of changepoints is relatively small - so that in this case $p(C_t|y_{1:t})$ will typically have mass concentrated on only a few values of $\{1, ..., t\}$ for any t in which case $\hat{p}(x_t|y_{1:t})$ need only be supported on a few points to provide a good approximation. Furthermore, we have for any $t \leq N$ that $\hat{p}(x_t|y_{1:t}) = p(x_t|y_{1:t})$, i.e. resampling is unnecessary for

$t \leq N$, in which case the algorithm is exact. An approximation at $t > N$ is only incurred as a result of the resampling.

Based on the sequence of filtering recursions $\hat{p}(x_t|y_{1:t})$ we are able to obtain an approximate sample of $\tau_{1:K}$ from $p(\tau_{1:K}|\theta)$ using the backward sampling scheme of Fearnhead (2006). The procedure is described in Algorithm 2, the final sorting step corrects for the generation of changepoint locations in descending order.

Algorithm 2 SMC backwards sampling

1. Sample $\tau_1 \sim \hat{p}(x_T|y_{1:T})$
2. While $\tau_k > 0$

$$\text{Sample} \quad \tau_{k+1}|\tau_k \sim p(x_k = \tau_k|x^i_{k+1} = \tau_{k+1})w_{\tau_k}(x^i_{k+1}) \quad \text{for} \quad x^i_{k+1} \in \mathbf{C}_{\tau_k}$$
$$\text{Set} \quad k = k + 1$$

3. Return $\tau_{1:k}$ (sorted in ascending order)

2.7.1.2 Conditional SMC for changepoints inference

Sampling from a kernel that is invariant with respect to $p(\tau_{1:K}|\theta)$ can be performed via a version of SMC known as conditional Sequential Monte Carlo (cSMC). At each iteration cSMC runs a particle filter to sample a new value of $\tau_{1:K}$ conditional on an existing configuration $\tau^*_{1:K}$ and any global parameters θ. Validity follows by considering the kernel as sampling from a conditional distribution from an extended target, including all random variables used to generate the particle filter Andrieu et al. (2010), Whiteley et al. (2011). Importantly, the procedure requires that the existing configuration $\tau^*_{1:K}$ survives at each resampling stage. This is enforced by the resampling procedure introduced in Whiteley et al. (2011) as a modification to the stratified optimal resampling.

The resulting cSMC algorithm is provided, for completeness, in Algorithm 3 and is essentially a rewriting of Algorithm (Whiteley et al. 2011; Conditional SMC forward filtering) and as such is stated without any claims of originality. The statement of the algorithm relies on the introduction of binary indicator variables $S_{t-1}(x_{t-1})$ which at each iteration characterise whether particle x_{t-1} has survived the resampling step or not - in which case it has zero weight. Finally, for the class

of models considered here, we have introduce the function providing the predictive likelihood of the next observation conditional on the most recent segment, i.e. $g_t(x_t) = p(y_t|y_{x_t+1:t-1})$ as well as the state-space model transition

$$f_t(x_t|x_{t-1}) = \begin{cases} \frac{1-G(t-x_{t-1}-1)}{1-G(t-x_{t-1}-2)} & \text{if} \quad x_t = x_{t-1} \\ \frac{G(t-x_{t-1}-1)-G(t-x_{t-1}-2)}{1-G(t-x_{t-1}-2)} & \text{if} \quad x_t = t-1 \\ 0 & \text{else} \end{cases}$$

The algorithm relies crucially in partitioning the particles into a set with low weight in step 2c of Algorithm 3. The partitioning intuitively separates those particles into 'low' and 'high' weights based on a constant, α_{t-1} . The high weight particles certainly survive the resulting iteration and are not at risk of being killed off in the resampling. For the 'low' weight particles however, each one can potentially have zero weight after resampling enabling the reduction of the number of support points of the filtering approximation at that algorithmic iteration.

Algorithm 3 Conditional SMC for changepoint inference

Inputs: $\tau_{1:K}$, N number of particles
Outputs: $\tau_{1:K}$

1. Initialise at $t = 1$, $\bar{W}_1(0) = g_1(0)$, $W_1(0) = 1$, $\mathbf{C}_1 = \{0\}$, $S_1(0) = 1$ and $\kappa = 0$.
2. For $t \in \{2, ..., T\}$
   (a) If $t - 1 \leq N$ set $\alpha_{t-1} = \infty$, otherwise set α_{t-1} to the unique solution of $\sum_{x_{t-1} \in \mathbf{C}_{t-1}} 1 \wedge \alpha_{t-1} W_{t-1}(x_{t-1}) = N$
   (b) If $\tau_{\kappa+1} < t - 1$ set $\kappa = \kappa + 1$.
   (c) Set $\mathcal{I}_{t-1} = \{x_{t-1} \in \mathbf{C}_{t-1} : W(x_{t-1}) \leq \frac{1}{\alpha_{t-1}}\}$
   (d) If $\tau_\kappa \notin \mathcal{I}_{t-1}$
      i. Keep $L_{t-1} := |\mathbf{C}_{t-1}| - |\mathcal{I}_{t-1}|$ support points, i.e. those in $x_{t-1} \in \mathbf{C}_{t-1} \backslash \mathcal{I}_{t-1}$ by setting $S_{t-1}(x_{t-1}) = 1$.
      ii. If $|\mathcal{I}_{t-1}| > 0$ resample $N - L_{t-1}$ number of particles from $\mathcal{I}_{t-1}$ using stratified resampling
   (e) If $\tau_\kappa \in \mathcal{I}_{t-1}$
      i. Keep $L_{t-1} := |\mathbf{C}_{t-1}| - |\mathcal{I}_{t-1}|$ support points, i.e. those in $x_{t-1} \in \mathbf{C}_{t-1} \backslash \mathcal{I}_{t-1}$ by setting $S_{t-1}(x_{t-1}) = 1$.
      ii. Resample $N - L_{t-1}$ number of particles from $\mathcal{I}_{t-1}$ using conditional stratified resampling
   (f) For each x_{t-1} surviving the previous resampling step set $S_{t-1}(x_{t-1}) = 1$ otherwise set $S_{t-1}(x_{t-1}) = 0$
   (g) Set $\bar{W}_t(t-1) = g_t(t-1) \sum_{x_{t-1} \in \mathbf{C}_{t-1}} f_t(t-1|x_{t-1}) W_{t-1}(x_{t-1})$
   (h) For $x_{t-1} \in \mathbf{C}_{t-1}$ set $\bar{W}_t(x_{t-1}) = 0$ if $S_{t-1}(x_{t-1}) = 0$ and otherwise

$$\bar{W}_t(x_{t-1}) = g_t(x_{t-1}) f_t(x_{t-1}|x_{t-1}) \frac{W_{t-1}(x_{t-1})}{1 \wedge \alpha_{t-1} W_{t-1}(x_{t-1})}$$

   (i) Update support $\mathbf{C}_t = \{x_{t-1} \in \mathbf{C}_{t-1} : S_{t-1}(x_{t-1}) = 1\} \cup \{t-1\}$
   (j) For $x_{t-1} \in \mathbf{C}_{t-1}$ set $W_t(x_t) \propto \bar{W}_t(x_t)$, $\sum_{x_t \in \mathbf{C}_t} W_t(x_t) = 1$.
3. $\tau_{1:K} \sim$ Backwards sample() (according to Algorithm 2)

The validity of the resulting algorithm is provided in (Whiteley et al. 2011; Theorem 2). After having run forwards for $t \in \{1, ..., T\}$, a backwards sampling step is performed as in Algorithm 2 to sample a new set of changepoint locations $\tau_{1:K}$ and mitigate the issues otherwise associated with particle degeneracy. Without it, as with other particle Gibbs algorithms, the conditional resampling typically results in the particle approximation of $p(\tau_{1:K}|\theta)$ degenerating to a point mass as $\tau^*_{1:K}$ with high probability, thereby slowing the mixing of the resulting chain. Adopting backwards sampling techniques as developed in Godsill et al. (2004) (also known

as 'ancestor sampling' Lindsten et al. (2014)) can be used to mitigate this. The resulting backward simulation method given in Fearnhead and Liu (2007) applies these techniques to the changepoint context.

2.7.1.3 Conditional SMC resampling

The stratified optimal resampling centres around obtaining a thresholding value, α, which can be interpreted an upperbound on the maximum Kolmogorov-Smirnov distance between the distributions before and after resampling Fearnhead (2006). Having obtained α, the resampling mechanism ensures any particle with weight superior to α is resampled with its original weight. As a reult, if $W_t^* > \alpha$ then C_t^* will necessarily have non-zero weight after resampling. For weights inferior to α, guaranteeing the resampling of C_t^* can be performed by an adaptation to the stratified resampling of Carpenter et al. (1999) as proposed in Whiteley et al. (2011). In the case of standard stratified optimal resampling, the empirical cumulative distribution function is constructed and a uniformly spaced grid constructed in the interval $[0, 1]$, such that any variable coinciding with the grid is resampled. The grid is chosen by fixing the number of gridlines equal to the number of particles to be resampled, L, and then choosing a variable uniformly $U^* \in [0, W_t^*]$ and constructing the grid so that at least one of the grid points intersects with the cumulative distribution at C_t^*. Each gridline is then equal to $\frac{k}{L} + U^* + \sum_{i:C_t^i<C_t^*} W_t^i$ for some integer $k \in \{0, ..., L\}$. This procedure is shown in more detail in Figure 2.10.

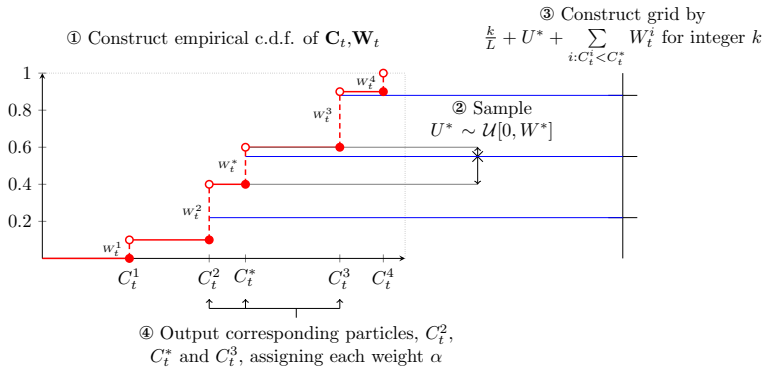


Figure 2.10: Diagram of example cSMC stratified resampling procedure. After constructing the particles' empirical c.d.f., a uniform random variable is sampled and the grid chosen such that the conditioning particle will always be resampled. Resampled particles are then output and assigned weight α.

3

Unbiased Smoothing using Particle Independent Metropolis-Hastings

Lawrence Middleton, George Deligiannidis, Arnaud Doucet, Pierre E. Jacob

Abstract

We consider the approximation of expectations with respect to the distribution of a latent Markov process given noisy measurements. This is known as the smoothing problem and is often approached with particle and Markov chain Monte Carlo (MCMC) methods. These methods provide consistent but biased estimators when run for a finite time. We propose a simple way of coupling two MCMC chains built using Particle Independent Metropolis–Hastings (PIMH) to produce unbiased smoothing estimators. Unbiased estimators are appealing in the context of parallel computing, and facilitate the construction of confidence intervals. The proposed scheme only requires access to off-the-shelf Particle Filters (PF) and is thus easier to implement than recently proposed unbiased smoothers. The approach is demonstrated on a Lévy-driven stochastic volatility model and a stochastic kinetic model.

3.1 Introduction

3.1.1 State-space models, problem statement and review

Let t denote a discrete-time index. State-space models are defined by a latent Markov process $(X_t)_{t\geq 1}$ and observation process $(Y_t)_{t\geq 1}$, (X_t, Y_t) taking values in a measurable space $(\mathsf{X}\times\mathsf{Y}, \mathcal{B}(\mathsf{X})\otimes\mathcal{B}(\mathsf{Y}))$ and satisfying

$$X_{t+1}|\{X_t = x\} \sim f(\cdot|x),\ \ Y_t|\{X_t = x\} \sim g(\cdot|x),$$

for $t \geq 1$ with $X_1 \sim \mu(\cdot)$. In the following we assume that $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\mathsf{Y} \subseteq \mathbb{R}^{d_y}$, and use $f(\cdot|x)$, $g(\cdot|x)$ and $\mu(\cdot)$ to denote densities with respect to the corresponding Lebesgue measure. State inference given a realization of the observations $Y_{1:T} = y_{1:T}$ for some fixed $T \in \mathbb{N}$ requires the posterior density

$$\begin{aligned}\pi\left(x_{1:T}\right) &:= p(x_{1:T}|y_{1:T}) \\ &\propto \mu(x_1)g(y_1|x_1)\prod_{t=2}^{T} f(x_t|x_{t-1})g(y_t|x_t),\end{aligned}$$

and expectations w.r.t. to this density. This is known as smoothing in the literature. For non-linear non-Gaussian state-space models, this problem is complex as this posterior and its normalizing constant $p(y_{1:T})$, often called somewhat abusively likelihood, are intractable. We provide here a means to obtain unbiased estimators of $\pi(h) := \int h(x_{1:T})\pi(x_{1:T})\mathrm{d}x_{1:T}$ for some function $h : \mathsf{X}^T \to \mathbb{R}$.

Particle methods return asymptotically consistent estimators which are however biased for a finite number of particles. Similarly MCMC kernels, such as the iterated Conditional Particle Filter (i-CPF) and PIMH Andrieu et al. (2010), can be used for approximating smoothing expectations consistently but are also biased for a finite number of iterations. Additionally, although theoretical bounds on the bias are available for particle Del Moral et al. (2007), PIMH Andrieu et al. (2010) and i-CPF Chopin and Singh (2015), Andrieu et al. (2018b), Lee et al. (2018) estimators, these bounds are usually not sharp and/or rely on strong mixing assumptions which are not satisfied by most realistic models. Unbiased estimators of $\pi(h)$ computed on parallel machines can be combined into asymptotically valid

confidence intervals as either the number of machines, the time budget, or both go to infinity Glynn and Heidelberger (1991).

Recently, it has been shown in Jacob et al. (2018), Lee et al. (2018) that it is possible to obtain such unbiased estimators by combining the i-CPF algorithm with a debiasing scheme for MCMC algorithms proposed initially in Glynn and Rhee (2014) and further developed in Jacob et al. (2017). These unbiased smoothing schemes couple two i-CPF kernels using common random numbers and a coupled resampling scheme. After a brief review of particle methods and of PIMH, we propose in Section 3.2 an alternative methodology relying on coupling two PIMH kernels. The method is easily implementable as it does not require any modification of the PF algorithm. It can also be used in scenarios where simulation from the Markov transition kernel of the latent process involves a random number of random variables, whereas the methods proposed in Jacob et al. (2018), Lee et al. (2018) would not be directly applicable to these settings. Additionally it does not require being able to evaluate pointwise the transition density contrary to the coupled conditional backward sampling PF scheme of Lee et al. (2018). Section 3.3 presents an analysis of the methodology when T is large. In Section 3.4, the method is demonstrated on a Lévy-driven stochastic volatility model and a stochastic kinetic model.[1]

3.1.2 Particle methods

Particle methods are often used to approximate smoothing expectations Doucet and Johansen (2011), Kantas et al. (2015). Such methods rely on sampling, weighting and resampling a set of N weighted particles (X_t^i, W_t^i), where $X_t^i \in \mathsf{X}$ denotes the value of the i^{th} particle at iteration t and W_t^i its corresponding normalized weight, i.e. $\sum_{i=1}^N W_t^i = 1$. Letting $q_1(x_1)$, $q_t(x_t|x_{t-1})$, denote the proposal density at time $t = 1$ and at time $t \geq 2$ respectively, weighting occurs according to the

[1] Code to reproduce figures is provided at `https://github.com/lolmid/coupled_pimh`

Algorithm 4 Particle Filter

For $i \in [N]$, sample $X_1^i \overset{i.i.d.}{\sim} q_1(\cdot)$, compute weights $W_1^i \propto w_1(X_1^i)$ and set $p_N(y_1) = \frac{1}{N}\sum_{i=1}^N w_1(X_1^i)$. For $2 \leq t \leq T$:

1. Sample particle ancestors $A_{t-1}^{1:N} \sim r(\cdot|\mathbf{W}_{t-1})$.
2. For $i \in [N]$, sample $X_t^i \sim q_t(\cdot|X_{t-1}^{A_{t-1}^i})$ and set $X_{1:t}^i = \{X_{1:t-1}^{A_{t-1}^i}, X_t^i\}$.
3. Compute weights $W_t^i \propto w_t(X_{t-1}^{A_{t-1}^i}, X_t^i)$ and set $p_N(y_{1:t}) = p_N(y_{1:t-1}) \cdot \frac{1}{N}\sum_{i=1}^N w_t(X_{t-1}^{A_{t-1}^i}, X_t^i)$.

Output $\pi_N(\mathrm{d}x_{1:T}) := \sum_{i=1}^N W_T^i \delta_{X_{1:T}^i}(\mathrm{d}x_{1:T})$ and $p_N(y_{1:T})$.

following 'incremental weights':

$$w_1(x_1) := \frac{g(y_1|x_1)\mu(x_1)}{q_1(x_1)} \quad \text{for} \quad t = 1,$$
$$w_t(x_{t-1}, x_t) := \frac{g(y_t|x_t)f(x_t|x_{t-1})}{q_t(x_t|x_{t-1})} \quad \text{for} \quad t \geq 2.$$

We assume that $w_1(x_1) > 0$ and $w_t(x_{t-1}, x_t) > 0$ for $t = 2, ..., T$ and all $x_{1:T}$. Pseudo-code for a standard PF is presented in Algorithm 4 where we let $r(\cdot|\mathbf{W}_t)$, with $\mathbf{W}_t := \left(W_t^1, ..., W_t^N\right)$, denote the resampling distribution, a probability distribution on $[N]^N$ where $[N] := \{1, ..., N\}$. We say that a resampling scheme is unbiased if $\sum_{i=1}^N r(A_t^i = k|\mathbf{W}_t) = NW_t^k$. All standard resampling schemes -multinomial, residual and systematic- are unbiased Doucet and Johansen (2011). This PF procedure outputs an approximation $p_N(y_{1:T})$ of $p(y_{1:T})$ and an approximation $\pi_N(\mathrm{d}x_{1:T})$ of the smoothing distribution $\pi(\mathrm{d}x_{1:T})$. Under weak assumptions, it can be shown that $p_N(y_{1:T})$, resp. $\pi_N(h) := \sum_{i=1}^N W_T^i h\left(X_{1:T}^i\right)$, is an asymptotically consistent (in N) estimator of $p(y_{1:T})$, resp. of $\pi(h)$. However, whereas $p_N(y_{1:T})$ is unbiased (Del Moral (2004), Section 7.4.1), $\pi_N(h)$ admits an asymptotic bias of order C/N for a constant C which is typically impossible to evaluate and for which only loose bounds are available under realistic assumptions Del Moral et al. (2007). In the following by a call to the PF, $(X_{1:T}, p_N) \sim \mathrm{PF}$, we mean a procedure which runs Algorithm 4 and returns $p_N := p_N(y_{1:T})$ (dependence on observations is notationally omitted) and a sample from the approximate smoothing distribution $X_{1:T} \sim \pi_N(\cdot)$, i.e. output $X_{1:T}^i$ with probability W_T^i.

Algorithm 5 PIMH kernel $P((X_{1:T}^{(n)}, p_N^{(n)}), (\cdot, \cdot))$

1. Sample $(X_{1:T}^*, p_N^*) \sim \mathsf{PF}$.
2. Set $(X_{1:T}^{(n+1)}, p_N^{(n+1)}) = (X_{1:T}^*, p_N^*)$ with probability $\alpha(p_N^{(n)}, p_N^*) := 1 \wedge \frac{p_N^*}{p_N^{(n)}}$.
3. Otherwise set $(X_{1:T}^{(n+1)}, p_N^{(n+1)}) = (X_{1:T}^{(n)}, p_N^{(n)})$.

3.1.3 PIMH method

An alternative way to estimate $\pi(h)$ consists of using an MCMC scheme targeting π. The PIMH algorithm achieves this by building a Markov chain on an extended space admitting a stationary distribution $\bar{\pi}$ with marginal π using the PF described in Algorithm 4 as proposal distribution Andrieu et al. (2010). Algorithm 5 provides pseudo-code for sampling the PIMH kernel, where $(X_{1:T}^{(n)}, p_N^{(n)})$ denotes the current state of this Markov chain and $a \wedge b$ means $\min(a, b)$.

Validity and convergence properties of PIMH rely on viewing it as an Independent Metropolis–Hastings (IMH) sampler on an extended space. For ease of presentation, we detail this construction for an unbiased resampling scheme satisfying additionally $r(A_t^i = k|\mathbf{W}_t) = W_t^k$ for all $i, k \in [N]$. The PF of Algorithm 4 implicitly defines a distribution ψ over $N \times T$ particle coordinates and $N \times (T-1)$ ancestors. We use ζ to denote a sample from ψ, where $\zeta \in \mathcal{X} := \mathsf{X}^{NT} \times \{1, ..., N\}^{N(T-1)}$, and the density of ψ is given by

$$\psi(\zeta) = \left(\prod_{i=1}^{N} q_1(x_1^i)\right) \prod_{t=2}^{T} \left(r(a_{t-1}^{1:N}|\mathbf{w}_{t-1}) \prod_{i=1}^{N} q(x_t^i|x_{t-1}^{a_{t-1}^i})\right).$$

We let b_t^j denote the index of the ancestor particle of $x_{1:T}^j$ at generation t, which may be obtained deterministically from the ancestry, using $b_t^j = a_t^{b_{t+1}^j}$ with $b_T^j = j$. From Andrieu et al. (2010), for $(j, \zeta) \in \{1, ..., N\} \times \mathcal{X}$, we express the target $\bar{\pi}(j, \zeta)$ of the resulting IMH sampler as

$$\bar{\pi}(j, \zeta) = \frac{\pi(x_{1:T}^j)}{N^T} \frac{\psi(\zeta)}{q_1(x_1^{b_1^j}) \prod_{t=2}^{T} r(b_{t-1}^j|\mathbf{w}_{t-1}) q_t(x_t^{b_t^j}|x_{t-1}^{b_{t-1}^j})}$$

Running a PF and sampling from π_N corresponds to sampling from the proposal $\bar{q}(j, \zeta) = \psi(\zeta) W_T^j$ and (Andrieu et al. 2010; Theorem 2) shows that $\bar{\pi}(j, \zeta)/\bar{q}(j, \zeta) = p_N(y_{1:T})/p(y_{1:T})$ where for a given (j, ζ) we understand $p_N(y_{1:T})$ as a deterministic

map from $\{1, ..., N\} \times \mathcal{X}$ to $\mathbb{R}^+$. As a result, samples from $\bar{\pi}$ can be obtained asymptotically through accepting proposals with probability $\alpha(p_N^{(n)}, p_N^*)$. We can thus estimate $\pi(h)$ by averaging over iterations $h(X_{1:T}^n)$. As shown in (Andrieu et al. 2010; Theorem 6), we can also estimate $\pi(h)$ by averaging over iterations $\pi_N^{(n)}(h) = \sum_{i=1}^N W_T^{i,n} h(X_{1:T}^{i,n})$, hence reusing all the particle system used to generate the *accepted* proposal at iteration n as $\bar{\pi}(\pi_N(h)) = \pi(h)$. As such, although Algorithms 5 and 6 are stated in terms of proposing and accepting $(X_{1:T}, p_N)$, it is possible to consider them instead as proposing and accepting (J, ζ). In Lee et al. (2019), it is shown that $r(A_t^i = k|\mathbf{W}_t) = W_t^k$ is unnecessary. We only need to use an unbiased resampling scheme to obtain a valid PIMH scheme. This is achieved by defining an alternative target $\bar{\pi}$ on $\{1, ..., N\} \times \mathcal{X}$ such that $X_{1:T}^J \sim \pi$ under $\bar{\pi}$ and $\bar{\pi}(j, \zeta)/\bar{q}(j, \zeta) = p_N(y_{1:T})/p(y_{1:T})$ also hold. It is also established in Lee et al. (2019) that one can even used adaptive resampling procedures Doucet and Johansen (2011), Del Moral et al. (2012).

If we denote by Z the error of the log-likelihood estimator, i.e. $Z := \log\{p_N(y_{1:T})/p(y_{1:T})\}$, the PIMH algorithm induces a Markov chain with transition kernel $Q(z, \mathrm{d}z')$ given by

$$\{1 \wedge \exp(z' - z)\}\, g(\mathrm{d}z') + \{1 - \alpha(z)\}\, \delta_z(\mathrm{d}z'), \tag{3.1}$$

where $g(\mathrm{d}z)$ is the distribution of Z under the law of the particle filter and $\alpha(z) := \int \{1 \wedge \exp(z' - z)\}\, g(\mathrm{d}z')$ is the average acceptance probability from state z. Although not emphasized notationally, both $g(z)$ and $\alpha(z)$ are functions of N. Through an abuse of notation, we denote the invariant density of the above chain with $\pi(z) = g\,(z) \exp\,(z)$. Such a reparameterization has also been used in previous work to analyze the PIMH and the related particle marginal MH algorithm Pitt et al. (2012), Doucet et al. (2015).

3.2 Methodology

3.2.1 Unbiased MCMC via couplings

'Exact estimation' methods provide unbiased estimators of expectations with respect to the stationary distribution of a Markov chain using coupling techniques Glynn

and Rhee (2014), Jacob et al. (2017). In the following we use these tools to couple PIMH kernels and estimate smoothing expectations, after briefly reviewing the general approach.

Consider two $\mathsf{U}-$valued Markov chains, $U := (U^{(n)})_{n\geq 0}$ and $\tilde{U} := (\tilde{U}^{(n)})_{n\geq 0}$, each evolving marginally according to a kernel $K(u, \mathrm{d}u')$ with stationary distribution λ and initialized from η so that $U^{(n)} \overset{d}{=} \tilde{U}^{(n)}$ for all $n \geq 1$, where $\overset{d}{=}$ denotes equality in distribution. We couple these two chains so that $U^{(n)} = \tilde{U}^{(n-1)}$ for $n \geq \tau$, where τ is an almost surely finite meeting time. In this case, we see that for non-negative integers $k < t$ we have the following telescoping-sum decomposition

$$\begin{aligned}\mathbb{E}[h(U^{(t)})] =& \mathbb{E}[h(U^{(k)}) + \sum_{l=k+1}^{t} (h(U^{(l)}) - h(U^{(l-1)}))] \\ =& \mathbb{E}[h(U^{(k)}) + \sum_{l=k+1}^{t\wedge(\tau-1)} (h(U^{(l)}) - h(\tilde{U}^{(l-1)}))].\end{aligned}$$

As a result, under integrability conditions which will be made precise in the sequel, taking the limit as $t \to \infty$ suggests an estimator $H_k(U, \tilde{U}) := h(U^{(k)}) + \sum_{l=k+1}^{\tau-1}(h(U^{(l)}) - h(\tilde{U}^{(l-1)}))$ with expectation $\lambda(h)$. A way to construct such chains considered in Glynn and Rhee (2014), Jacob et al. (2017), Heng and Jacob (2017), Middleton et al. (2018) relies on sampling independently $(U^{(0)}, U^{(1)}) \sim \eta(\mathrm{d}u_0)K(u_0, \mathrm{d}u_1)$ and $\tilde{U}^{(0)} \sim \eta(\mathrm{d}u_0)$, and then successively sampling $(U^{(n+1)}, \tilde{U}^{(n)})$ from $\bar{K}((U^{(n)}, \tilde{U}^{(n-1)}), (\cdot, \cdot))$ such that $U^{(n+1)} = \tilde{U}^{(n)}$ with positive probability, ensuring that both chains evolve marginally according to K. Furthermore, averaging $H_l(U, \tilde{U})$ over a range of values $l \in \{k, k+1, ..., m\}$, for some $m \geq k$, preserves unbiasedness, suggesting the following 'time-averaged' unbiased estimator of $\mu(h)$

$$\begin{aligned}H_{k:m} =& \frac{1}{m-k+1}\sum_{l=k}^{m} h(U^{(l)}) + \\ &\sum_{l=k+1}^{\tau-1} \min\left(1, \frac{l-k}{m-k+1}\right)(h(U^{(l)}) - h(\tilde{U}^{(l-1)})) \\ &:= \mathrm{MCMC}_{k:m}(h) + \mathrm{BC}_{k:m}(h). \end{aligned} \tag{3.2}$$

We view $\mathrm{MCMC}_{k:m}(h)$ as the standard MCMC sample average up to time m, discarding the first k iterates as 'burn-in'. The second term, $\mathrm{BC}_{k:m}(h)$, is the 'bias

correction' term with $\mathrm{BC}_{k:m}(h) := 0$ if $\tau - 1 < k + 1$. The estimator requires only that two chains be simulated until meeting at time τ, after which, if $\tau < m$, only one chain must be simulated up to m. By (Jacob et al. 2017; Proposition 3.1), the validity of the resulting estimators is guaranteed under the following assumptions. We discuss these assumptions for PIMH in Section 3.2.3.

Assumption 2. *Each chain is initialized marginally from a distribution η, evolves according to a kernel K, and is such that $\mathbb{E}[h(U^{(n)})] \to \lambda(h)$ as $n \to \infty$. Furthermore, there exist constants $\eta > 0$ and $D < \infty$ such that $\mathbb{E}[\left|h(U^{(n)})\right|^{2+\eta}] < D$ for all $n \geq 0$.*

Assumption 3. *The two chains are such that the meeting time $\tau = \inf\{n \geq 1 : U^{(n)} = \tilde{U}^{(n-1)}\}$ satisfies $\mathbb{P}[\tau > n] \leq D\delta^n$, for some constants $0 < D < \infty$ and $\delta \in (0, 1)$. The chains stay together after meeting, i.e. $U^{(n)} = \tilde{U}^{(n-1)}$ for all $n \geq \tau$.*

Proposition 4. *Jacob et al. (2017) Under Assumptions 2 and 13, the estimator $H_{k:m}$ obtained by these coupled chains is unbiased, has finite variance and finite expected cost.*

3.2.2 Coupled PIMH

To obtain unbiased estimators of $\pi(h)$, we use the framework detailed in Section 3.2.1 for $K = P$ the transition kernel of PIMH and $\lambda = \bar{\pi}$ the corresponding invariant distribution of PIMH defined on $\mathsf{U} := \{1, ..., N\} \times \mathcal{X}$. This requires introducing a coupling of PIMH kernels. For IMH chains, a natural choice of initialization of the chain is from the proposal distribution. We adopt this in the following, sampling both $(X_{1:T}^{(0)}, p_N^{(0)}) \sim \mathsf{PF}$ and $(\tilde{X}_{1:T}^{(0)}, \tilde{p}_N^{(0)}) \sim \mathsf{PF}$. However, in contrast to previous constructions Jacob et al. (2017), Heng and Jacob (2017), Middleton et al. (2018), our method allows the chains to couple at time $\tau = 1$ through re-using the initial value $(\tilde{X}_{1:T}^{(0)}, \tilde{p}_N^{(0)})$ as a proposal for the first chain.

We summarize the resulting procedure in Algorithm 6. To simplify presentation of the algorithm, we use the convention $\alpha(\tilde{p}_N^{(-1)}, p_N^*) := 1$. Finally, although the coupling scheme is framed in terms of PIMH, it is clear that this algorithm and estimators apply equally to any IMH algorithm.

Algorithm 6 Coupled PIMH

1. Sample $(X_{1:T}^{(0)}, p_N^{(0)}) \sim \mathsf{PF}$, set $n = 1$ and $\tau = \infty$.
2. While $n < \max(m, \tau)$
   (a) Sample $(X_{1:T}^{*}, p_N^{*}) \sim \mathsf{PF}$.
   (b) Sample $\mathfrak{u} \sim \mathcal{U}[0, 1]$.
   (c) If $\mathfrak{u} \leq \alpha\left(p_N^{(n-1)}, p_N^{*}\right)$ set
   $$(X_{1:T}^{(n)}, p_N^{(n)}) = (X_{1:T}^{*}, p_N^{*})$$
   else $(X_{1:T}^{(n)}, p_N^{(n)}) = (X_{1:T}^{(n-1)}, p_N^{(n-1)})$.
   (d) If $\mathfrak{u} \leq \alpha(\tilde{p}_N^{(n-2)}, p_N^{*})$ set
   $$(\tilde{X}_{1:T}^{(n-1)}, \tilde{p}_N^{(n-1)}) = (X_{1:T}^{*}, p_N^{*})$$
   else $(\tilde{X}_{1:T}^{(n-1)}, \tilde{p}_N^{(n-1)}) = (\tilde{X}_{1:T}^{(n-2)}, \tilde{p}_N^{(n-2)})$.
   (e) If $\mathfrak{u} \leq \alpha\left(p_N^{(n-1)}, p_N^{*}\right) \wedge \alpha(\tilde{p}_N^{(n-2)}, p_N^{*})$ set $\tau = n$.
   (f) Set $n \leftarrow n + 1$.
3. Return $H_{k:m}$ as in (3.2).

At each iteration n of Algorithm 6 the proposal $(X_{1:T}^{*}, p_N^{*})$ can be accepted by one, both or neither chains with positive probability. The meeting time τ corresponds to the first time the proposal $(X_{1:T}^{*}, p_N^{*})$ is accepted by both chains. We can also return another unbiased estimator $\bar{H}_{k:m}$ of the form (3.2) with $h(X_{1:T})$ replaced by $\pi_N(h)$ as we have previously seen that $\bar{\pi}(\pi_N(h)) = \pi(h)$. The Rao-Blackwellised estimator $\bar{H}_{k:m}$ will typically outperform significantly $H_{k:m}$ when we are interested in smoothing expectations of functions of states close to T. For functions of states close to the origin, e.g. $h(x_{1:T}) = x_1$, we have $\bar{H}_{k:m} = H_{k:m}$ with high probability if N is moderate because of the particle degeneracy problem Doucet and Johansen (2011), Jacob et al. (2015), Kantas et al. (2015). This is illustrated in Appendix 3.6.3.

3.2.3 Validity and meeting times

The validity of our unbiased estimators is ensured if the following weak assumptions are satisfied.

Assumption 5. *There exist constants $\eta > 0$ and $D < \infty$ such that $\mathbb{E}\left[|h(X_{1:T}^{(n)})|^{2+\eta}\right] < D$ for all $n \geq 0$.*

Assumption 6. *There exist constants $\eta > 0$ and $D < \infty$ such that*

$$\mathbb{E}\left[|\sum_{i=1}^{N} W_T^{i,(n)} h(X_{1:T}^{i,(n)})|^{2+\eta}\right] < D$$

for all $n \geq 0$.

Assumption 7. *The resampling scheme is unbiased and there exist finite constants $(\bar{w}_t)_{t=1}^T$ such that $\sup_{x \in \mathsf{X}} w_1(x) \leq \bar{w}_1$ and $\sup_{(x,x') \in \mathsf{X} \times \mathsf{X}} w_t(x, x') \leq \bar{w}_t$ for $t \in \{2, ..., T\}$.*

Proposition 8. *Under Assumptions 5 and 7, resp. Assumptions 6 and 7, the estimator $H_{k:m}$, resp. $\bar{H}_{k:m}$, of $\pi(h)$ obtained from Algorithm 6 is unbiased and has finite variance and finite expected cost.*

To establish the result for $H_{k:m}$, note that Assumption 5 and our construction implies Assumption 2 for $\lambda = \bar{\pi}$. Assumption 7 provides verifiable and sufficient conditions to ensure uniform ergodicity of PIMH (Andrieu et al. 2010; Theorem 3). Hence it follows from (Jacob et al. 2017; Proposition 3.4) that the geometric bound on the tails of τ is satisfied. Additionally, Algorithm 6 ensures that the chains stay together for all $n \geq \tau$ so Assumption 13 is satisfied. A similar reasoning provides the result for $\bar{H}_{k:m}$.

We have the following precise description of the distribution of the meeting time τ for Algorithm 6. Let $\text{Geo}(\gamma)$ denote the geometric distribution on the strictly positive integers with success probability γ.

Proposition 9. *The meeting time satisfies $\tau | Z^{(0)} \sim \text{Geo}(\alpha(Z^{(0)}))$ and $\mathbb{P}[\tau = 1] \geq \frac{1}{2}$. Additionally, we have $\lim_{N \to \infty} \mathbb{P}[\tau = 1] = 1$ under Assumption 7.*

We recall here that $\alpha(z)$ is the average acceptance probability from state z. The geometric distribution result follows from Proposition 11 in Appendix. The rest of the proposition follows from

$$\begin{aligned}\mathbb{P}[\tau = 1] &= \int \alpha(z) g(\mathrm{d}z) \\ &= \iint \{1 \wedge \exp(z' - z)\} \, g(\mathrm{d}z) g(\mathrm{d}z').\end{aligned}$$

It entails trivially that $\mathbb{P}[\tau = 1] \geq \frac{1}{2}$. Noting that $\lim_{N\to\infty} Z = 0$ a.s. under g (which is dependent on N) under Assumption 7, see e.g. Del Moral (2004), we obtain $\lim_{N\to\infty} \mathbb{P}[\tau = 1] = 1$ by dominated convergence, thus $\lim_{N\to\infty} \mathrm{BC}_{k:m}(h) = 0$.

3.2.4 Unbiased filtering

A PF generates estimates $p_N(y_{1:t})$ of $p(y_{1:t})$ at all times $t = 1, \ldots, T$. These estimates can be used to perform one step of coupled PIMH for T pairs of Markov chains, each pair corresponding to one of the smoothing distributions $p(x_{1:t}|y_{1:t})$. This requires some additional bookkeeping to keep track of the meeting times and unbiased estimators associated with each $p(x_{1:t}|y_{1:t})$ but can be used to unbiasedly estimate expectations with respect to all filtering distributions $p(x_t|y_{1:t})$. This was not directly feasible with coupled i-CPF in Jacob et al. (2018). In particular, this allows us to estimate unbiasedly the predictive likelihood terms $p(y_t|y_{1:t-1})$ which can be used for a goodness-of-fit test Gerlach et al. (1999).

3.3 Analysis

3.3.1 Meeting time: large sample approximation

We investigate here the distribution of the meeting time in the large sample regime, i.e. in the interesting scenarios where T is large. Under strong mixing assumptions, Bérard et al. (2014) showed that letting $\frac{T}{N} = \gamma$ for some $\gamma > 0$ then the following Central Limit Theorem (CLT) holds: $Z \xrightarrow{d} \mathcal{N}\left(-\frac{1}{2}\sigma^2, \sigma^2\right)$ as $T \to \infty$ where $\sigma^2 = \gamma\bar{\sigma}^2$ for some $\bar{\sigma}^2 > 0$. Empirically, this CLT appears to hold for many realistic models not satisfying these strong mixing assumptions Pitt et al. (2012), Doucet et al. (2015). Under this CLT, using similar regularity conditions as in Schmon et al. (2018), it can be shown that the Markov kernel Q defined in (3.1) converges in some suitable sense towards the Markov kernel Q_σ given by

$$\{1 \wedge \exp(z' - z)\}\, g_\sigma(z')\mathrm{d}z' + \{1 - \alpha_\sigma(z)\}\, \delta_z(\mathrm{d}z'),$$

where $g_\sigma(z) = \mathcal{N}\left(z; -\frac{1}{2}\gamma\bar{\sigma}^2, \gamma\bar{\sigma}^2\right)$ and $\alpha_\sigma(z) := \int \{1 \wedge \exp(z' - z)\}\, g_\sigma(z')\mathrm{d}z'$ denotes the average acceptance probability in z. For this limiting kernel[2], the following result holds.

Proposition 10. *(Doucet et al. 2015; Corollary 3) The invariant density of Q_σ is $\pi_\sigma(z) = \mathcal{N}\left(z; \frac{1}{2}\sigma^2, \sigma^2\right)$ and its average acceptance probability is given by*

$$\alpha_\sigma(z) := 1 - \Phi\left(\frac{z + \frac{\sigma^2}{2}}{\sigma}\right) + e^{-z}\Phi\left(\frac{z - \frac{\sigma^2}{2}}{\sigma}\right),$$

where Φ denotes the cumulative distribution function of the standard Normal distribution.

Under this large sample approximation, the probability $\mathbb{P}[\tau = n]$ can be written as

$$\mathbb{E}\left[\mathbb{P}[\tau = n | Z^{(0)}]\right] = \int \alpha_\sigma(z)(1 - \alpha_\sigma(z))^{n-1} g_\sigma(z)\mathrm{d}z, \tag{3.3}$$

thus $\mathbb{P}[\tau = 1] = \frac{1}{2}\{1 + \exp(\sigma^2)\mathrm{Erfc}(\sigma)\}$, Erfc denoting the complementary error function. Similarly we see that the expected meeting time is given by $\mathbb{E}[\tau] = \mathbb{E}_{g_\sigma}[\alpha_\sigma(Z)^{-1}]$. This allows us to approximate numerically expectations, quantiles and probabilities of τ as a function of σ, the standard deviation of $\log p_N(y_{1:T})$ under the law of the particle filter. Figure 3.1 displays $\mathbb{P}[\tau = 1]$ and $\mathbb{E}[\tau]$ as a function of σ. We see that $\mathbb{P}[\tau = 1] = 0.71$ for $\sigma = 1$ rising to $\mathbb{P}[\tau = 1] = 0.95$ for $\sigma = 0.1$. The expected meeting time $\mathbb{E}[\tau]$ is also the expected number of iterations to return an unbiased estimator for $m = 0$ and grows relatively benignly with σ.

We compare the distribution of the meeting times with its large sample approximation (3.3) on a stationary auto-regressive (AR) model $X_t \sim \mathcal{N}(aX_{t-1}, 1)$ and $Y_t \sim \mathcal{N}(X_t, \sigma_y^2)$ with $a = 0.5$ and $\sigma_y^2 = 10$ on a simulated dataset of $T = 100$. Estimates of $\mathbb{P}[\tau \geq n]$ were obtained empirically for a range of values of N and compared with the predicted values based on the estimated variance σ^2 of $\log p_N(y_{1:T})$ using 10^5 runs of coupled PIMH. Varying N between 10 and 110, σ^2 was between 0.2 and 3.0 in this range. The tail probabilities of the meeting time over this range are shown in Figure 4.2a. Confidence intervals for the estimates of $\mathbb{P}[\tau \geq n]$

[2]Note that Q_σ is not uniformly ergodic whereas Q is under Assumption 7.



Figure 3.1: $\mathbb{P}[\tau = 1]$ (left) and $\mathbb{E}[\tau]$ (right) as a function of the standard deviation σ of $\log p_N(y_{1:T})$.

are shown in Figure 4.2a with error bars indicating ± 2 standard deviations. We see that there is a satisfactory agreement, with predicted tail probabilities closely matching confidence intervals for each value, and larger values of N leading as expected to shorter meeting times on average.

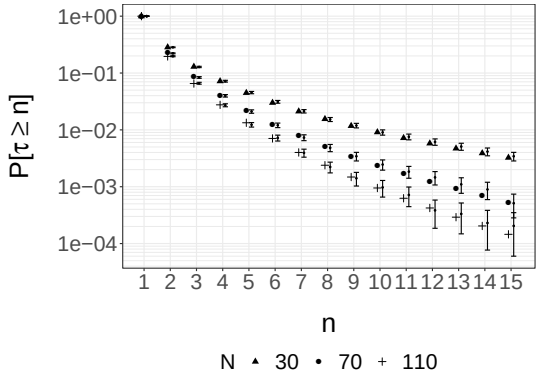


Figure 3.2: Empirical estimates of $\mathbb{P}[\tau \geq n]$ (with ± 2 standard errors) and comparison with the large sample approximation for toy example.

3.3.2 On the selection of N for large m

We provide here a heuristic for selecting N to minimize the variance of $H_{k:m}$ at fixed computational budget when both m and T are large. We will minimize the computational inefficiency defined by

$$\mathrm{C}[H] := \mathbb{V}[H_{k:m}] \times N$$

as the running time is proportional to N. For m sufficiently large, we expect that $\tau < m$ with very high probability and so the time to obtain a single unbiased estimator is m. We note that increasing N typically leads to a decreasing asymptotic variance of the ergodic averages associated with the PIMH chain and from Proposition 9 a reduction in the bias correction, however at the cost of more computation. For large m, we also expect that the dominant term of $\mathbb{V}[H_{k:m}]$ will arise from $\text{MCMC}_{k:m}(h)$ and will be essentially the asymptotic variance of h given by $\mathbb{V}_\pi[h]\text{IF}(h)$ divided by $(m-k+1)$, where $\text{IF}(h)$ is the Integrated Autocorrelation Time (IACT) of h for the PIMH kernel. For N sufficiently large, we expect that $X_{1:T}$ and p_N are approximately independent under the PF proposal. By a reasoning similar to the proof of (Pitt et al. 2012; Lemma 4), $\text{IF}(h)$ will then be approximately proportional to $\text{IF}(\sigma)$ defined in Eq. (11) in Pitt et al. (2012). This is illustrated in Appendix 3.6.2 where we plot $\text{IF}(h)$ for a variety of test functions for a range of N against $\text{IF}(\sigma)$ over the corresponding range of σ and show that they are indeed approximately proportional. Minimizing $\text{IF}[H]$ w.r.t. N is then approximately equivalent to minimizing $\text{IF}(\sigma)/\sigma^2$ as σ^2 is typically inversely proportional to N. This minimization has already been carried out in Pitt et al. (2012) where it was found that the minimizing argument is $\sigma = 0.92$. Practically, this means that one should select N to ensure that the standard deviation of $\log p_N(y_{1:T})$ is equal approximately to this value. The resulting value of N is expected to be close to the value of N minimizing $\text{C}[H]$, which is approximately proportional to $\text{IF}(h)/\sigma^2$. This is verified in Figure 3.3 on the AR example of Section 3.3.1. Note that these guidelines do not apply to $\bar{H}_{k:m}$ for h a function of states close to T as it is not true that $\pi_N(h)$ and p_N are approximately independent.

3.4 Numerical experiments

We apply the methodology to two models where the transition density of the latent process is analytically intractable but simulation from it is possible. However, this involves sampling a random number of random variables. The unbiased smoother proposed in Jacob et al. (2018) relies on common random numbers and it is unclear

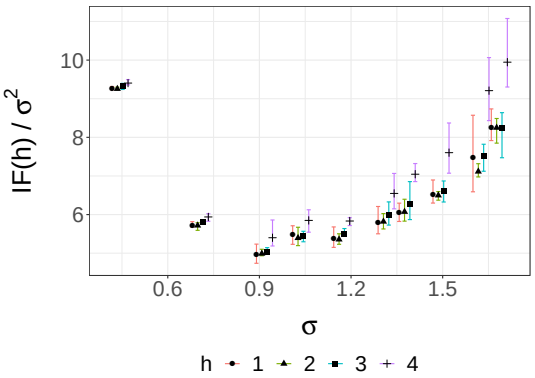


Figure 3.3: IF$(h_i)/\sigma^2$ for $h_1 = x_1$, $h_2 = x_T$, $h_3 = \sum_t x_t$ and $h_4 = \sum_t x_t^2$ as a function of σ^2.

how one could implement it in this context. The coupled conditional backward sampling PF scheme proposed in Lee et al. (2018) does not apply as we cannot evaluate the transition density pointwise. In both scenarios, we use the bootstrap PF with multinomial resampling, that is $q_1(x_1) = \mu(x_1)$ and $q_t(x_t|x_{t-1}) = f(x_t|x_{t-1})$, and Assumption 7 is satisfied.

3.4.1 Stochastic kinetic model

We consider a stochastic kinetic model represented by a jump Markov process introduced in Golightly and Wilkinson (2005). Such models describe a system of chemical reactions in continuous time, with a reaction occurring under the collision of two species at random times. The discrete number of each species describes the state with jumps representing the change in a particular species. The discrete valued state $(X_{t,q})_{t\geq 0, 1\leq q\leq Q}$ comprises a Q-vector of species at each time, where one of R reactions may occur at any random time, given by hazard functions f_r for $r \in \{1, ..., R\}$. The effect of such a reaction is described by a stoichiometry matrix S, where the instantaneous change in the number of species q for a certain reaction r out of a possible R different reactions is encoded in element $S_{q,r}$. For the prokaryotic autoregulation model and parameterisation considered in Golightly and Kypraios (2018), the state is a four dimensional vector evolving according to

8 possible reactions, for which the stoichiometry matrix is given by

$$S = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 2 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$f(X, c) = (c_1 X_4 X_3, c_2(k - X_4), c_3 X_4, c_4 X_1,$$

$$c_5 X_2(X_2 - 1)/2, c_6 X_3, c_7 X_1, c_8 X_2)'.$$

Coefficients $c_{1:8}$ and k are parameters of the model, given by $c = (0.1, 0.7, 0.35, 0.2, 0.1, 0.9, 0.3, 0.1)$ and $k = 10$. We collect $T = 100$ noisy observations of the latent process at regular intervals of length $\Delta = 0.1$, i.e.

$$Y_t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{pmatrix} X_{\Delta t} + \epsilon_t, \quad \epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, I_2).$$

We fix the initial condition $X_0 = (8, 8, 8, 5)$.

To simulate synthetic data and to run the bootstrap PF, we sample the latent process X_t using Gillespie's direct method Gillespie (1977), whereby the time to the next event is exponential with rate $\sum_{r=1}^{R} f_r(X, c)$ and reaction r occurs with probability $f_r(X, c)/\sum_{r=1}^{R} f_r(X, c)$. The estimated survival probabilities of the meeting time τ, with ± 2 standard errors, computed using 500 independent runs of coupled PIMH are plotted in Figure 3.4, along with the probabilities obtained from the large sample approximation in Section 3.3.1, showing good agreement between the two. In Figure 3.5 we display the unbiased smoothing estimators obtained by averaging the unbiased estimators obtained over 500 independent runs for $N = 1,000, k = m = 0$ and the corresponding confidence intervals. Alternative choices of k, m can lead to improved performance at fixed computational budget Jacob et al. (2017), Middleton et al. (2018).

3.4.2 Lévy-driven stochastic volatility

Introduced in Barndorff-Nielsen and Shephard (2001), Lévy-driven stochastic volatility models provide a flexible model for the log-returns of a financial asset. Letting (Y_t) denote the log-return process, we have

$$Y_t = \mu + \beta V_t + V_t^{1/2} \epsilon_t, \qquad \epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),$$

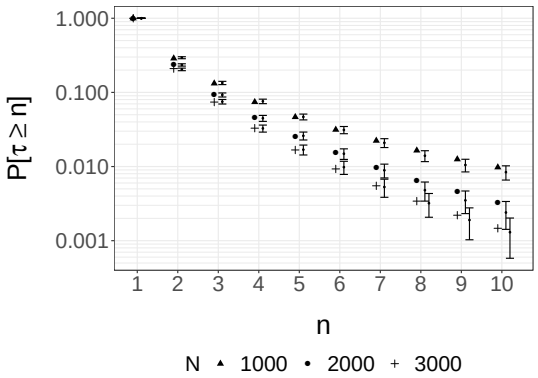


Figure 3.4: Empirical estimates of $\mathbb{P}[\tau \geq n]$ (with ± 2 standard errors) and comparison with those implied by the large sample approximation for stochastic kinetic model.

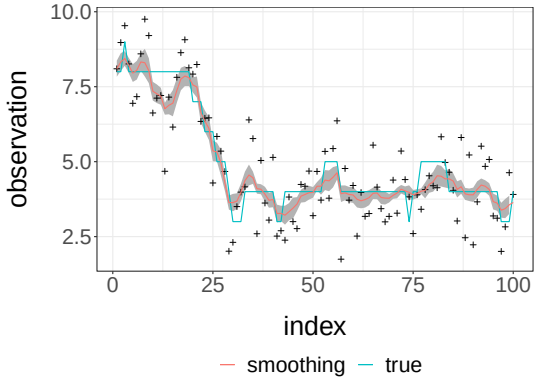


Figure 3.5: Unbiased estimates of $\mathbb{E}(X_{1,t}|y_{1:T})$ (red) (with ± 3 pointwise standard errors) and $X_{1,t}$ (blue) for Markov jump process.

where V_t (termed the 'actual volatility') is treated as a stationary stochastic process. The latent state comprises the pair of actual and spot volatility $X_t = (V_t, W_t)$. In the terminology of Barndorff-Nielsen and Shephard (2002), the integrated volatility is the integral of the spot volatility and the actual volatility is an increment of the integrated volatility over some unit time. Initializing $Z_0 \sim \Gamma(\xi^2/\omega^2, \xi^2/\omega^2)$, the process evolves through sampling the following random variables and recursing the state $X_t = (V_t, W_t)$ according to

$$K \sim \text{Poisson}(\lambda\xi^2/\omega^2), \quad C_{1:K} \overset{i.i.d.}{\sim} \mathcal{U}[t-1, t],$$

$$E_{1:K} \overset{i.i.d.}{\sim} \text{Exp}(\xi/\omega^2), \quad W_t = e^{-\lambda} W_{t-1} + \sum_{j=1}^{K} e^{-\lambda(t-C_j)} E_j,$$

$$V_t = \frac{1}{\lambda}(W_{t-1} - W_t + \sum_{j=1}^{K} E_j).$$

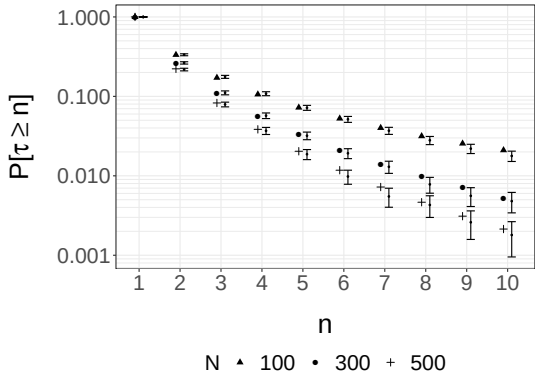
In particular, conditionally on X_{t-1}, simulation of X_t requires a random number of random numbers sampled at each iteration. The parameters (ξ, ω^2) denote the stationary mean and variance of the spot volatility respectively, λ describes the exponential decay of autocorrelations, β denotes the risk premium for excess volatility and μ the drift of the log-return. In the following we perform unbiased smoothing of X_t using $T = 500$ data from the S&P 500 index used in Chopin et al. (2013). A summary of the data and parameter inference is included in Appendix 3.6.4.

The empirical distributions of the meeting time for for $N \in \{100, ..., 500\}$ obtained using $1,000$ runs are shown in Figure 3.6, again showing good agreement with the large-sample approximation. We then obtain unbiased estimators $H_{k:m}$ for $h(x_{1:T}) = \sum_{t=1}^{T} v_t$ using $k = 20, m = 512$ over this grid of N. Figure 3.6b plots $(m - k + 1)\mathbb{V}[H_{k:m}]/(\mathbb{V}_\pi[h]\sigma^2)$. We expect this function to be close to $\mathrm{IF}(h)/\sigma^2$ as $m \to \infty$ and $\mathrm{IF}(h)/\sigma^2$ to be minimized around 0.92. The experiments are consistent with this result. Figure 3.6c presents the unbiased estimates of the spot volatility W_t obtained by averaging the unbiased estimates obtained over 1,000 runs and the corresponding confidence intervals for $k = m = 0$ and $N = 100$. Different choices of k, m could lead to improved performance at fixed computational budget.

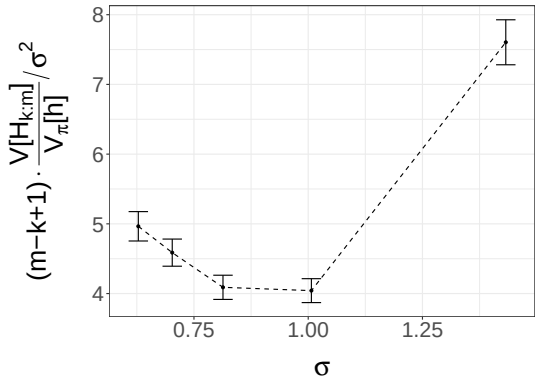
3.5 Discussion

We have introduced a simple approach to perform unbiased smoothing in state-space models and we have provided guidance on the choice of tuning parameters through appealing to a large sample approximation.

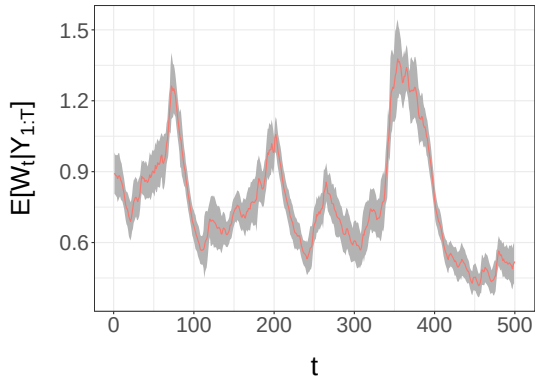
We have established the validity of the estimators when the incremental weights are bounded (Assumption 7) which ensures uniform ergodicity of the PIMH. Rejection sampling is possible under a similar assumption and would provide exact samples from the smoothing distribution. However the expected number of trials before acceptance of such a rejection scheme increases typically exponentially fast with T. If we are only interested in obtaining unbiased smoothing estimators and if the CLT discussed in Section 3.3.1 holds, we expect our coupling scheme to



(a) Empirical (± 2 standard errors) and predicted tail probabilities implied by large-sample approximation $\mathbb{P}[\tau \geq n]$.



(b) Estimates of $(m-k+1)\mathbb{V}[H_{k:m}]/(\mathbb{V}_\pi[h]\sigma^2)$ (± 3 standard errors) computed using 10,000 runs.



(c) Unbiased estimates of $\mathbb{E}(W_t|y_{1:T})$ of the S&P 500 over 2005-2007 and confidence intervals at ± 3 standard errors.

Figure 3.6: Distribution of meeting times (top), computational inefficiency (bottom) and smoothing state estimators for Lévy-driven stochastic volatility applied to real data (bottom).

only require increasing N linearly with T to control σ and thus the corresponding expectation of the meeting time.

Finally, the scheme proposed here can be extended to obtain unbiased estimators of expectations with respect to any posterior distribution by replacing the particle filter proposal within the IMH by Annealed Importance Sampling Neal (2001) or a sequential Monte Carlo sampler Del Moral et al. (2006). This is illustrated in Appendix 3.7.

3.6 Appendix

3.6.1 Proof of Proposition 9

We establish a monotonicity property of the proposed coupling for PIMH. Monotonicity properties of IMH samplers were exploited in an exact simulation context in Corcoran and Tweedie (2002) without this explicit construction.

Proposition 11. *Under the proposed coupling scheme the sequence of likelihood estimates* $(p_N^{(n+1)})_{n\geq 0}$ *stochastically dominates* $(\tilde{p}_N^{(n)})_{n\geq 0}$ *in the sense that for any* $n \geq 1$, $s \geq 0$

$$p_N^{(n)} \geq \tilde{p}_N^{(n-1)} \Rightarrow p_N^{(n+s)} \geq \tilde{p}_N^{(n+s-1)} \quad a.s.$$

Proof. The coupling procedure in Algorithm 6 uses a single proposal p_N^* and samples $\mathfrak{u} \sim \mathcal{U}[0,1]$, with proposals being accepted according to:

$$\text{if } \mathfrak{u} \leq 1 \wedge \frac{p_N^*}{p_N^{(n)}} \text{ then } p_N^{(n+1)} = p_N^*, \text{ else } p_N^{(n+1)} = p_N^{(n)},$$
$$\text{if } \mathfrak{u} \leq 1 \wedge \frac{p_N^*}{\tilde{p}_N^{(n-1)}} \text{ then } \tilde{p}_N^{(n)} = p_N^*, \text{ else } \tilde{p}_N^{(n)} = \tilde{p}_N^{(n-1)}.$$

We see that if $p_N^{(n)} \geq \tilde{p}_N^{(n-1)}$ then either

1. $p_N^{(n+1)} = p_N^*$ in which case

$$\mathfrak{u} \leq 1 \wedge \frac{p_N^*}{p_N^{(n)}} \implies \mathfrak{u} \leq 1 \wedge \frac{p_N^*}{\tilde{p}_N^{(n-1)}}$$

   as $p_N^{(n)} \geq \tilde{p}_N^{(n-1)}$ so that $p_N^{(n+1)} = \tilde{p}_N^{(n)} = p_N^*$ (i.e. the chains meet).
2. $p_N^{(n+1)} = p_N^{(n)}$ and so $p_N^* \leq p_N^{(n)}$. In this case, either $\tilde{p}_N^{(n)} = p_N^*$, and so $p_N^{(n+1)} \geq \tilde{p}_N^{(n)}$, or $\tilde{p}_N^{(n)} = \tilde{p}_N^{(n-1)}$ in which case both chains have rejected p_N^* and the ordering is preserved.

Finally, from the initialization of the procedure, we have $p_N^{(1)} \geq \tilde{p}_N^{(0)}$ because the initial state of the second chain is used as a proposal in the first iteration of the first chain. □

From the above reasoning we see that the chains meet when the first chain accepts its proposal for the first time, as the second chain then necessarily accepts the same proposal.

From the initial state with likelihood estimate $p_N^{(0)}$, the acceptance probability of the first chain is $\int 1 \wedge (p_N/p_N^{(0)})\bar{g}(p_N)\mathrm{d}p_N$, with $\bar{g}$ denoting the density of the PF likelihood estimator p_N. Thus, the time to the first acceptance follows a Geometric distribution with success probability $\int 1 \wedge (p_N/p_N^{(0)})\bar{g}(p_N)\mathrm{d}p_N$. The result stated in Proposition 9 follows when rewriting the problem using the error of the log-likelihood estimator $\log\{p_N(y_{1:T})/p(y_{1:T})\}$.

3.6.2 Integrated autocorrelation time for various test functions

We show here experimentally that $\mathrm{IF}(h)$ is approximately proportional to $\mathrm{IF}(\sigma)$ for various test functions: $h_1 : x_{1:T} \mapsto x_1$, $h_2 : x_{1:T} \mapsto x_T$, $h_3 : x_{1:T} \mapsto \sum_t x_t$ and $h_4 : x_{1:T} \mapsto \sum_t x_t^2$. This is illustrated in Figure 3.7 where $\mathrm{IF}(h)$ is displayed for a range of N against $\mathrm{IF}(\sigma)$ over the corresponding range of σ.

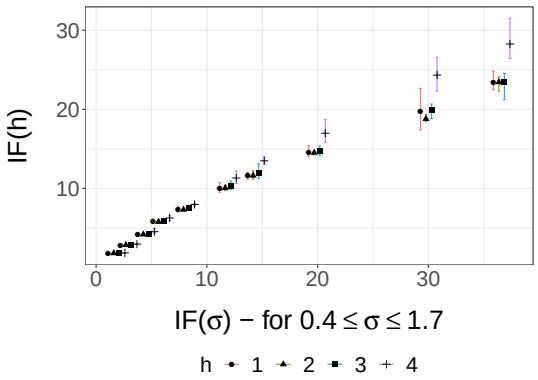


Figure 3.7: Inefficiency $\mathrm{IF}[h_i]$ versus $\mathrm{IF}[\sigma]$, with markers indicating the test functions h_1, h_2, h_3, h_4. The vertical axis scale is relative, depending on the test function.

3.6.3 Rao-Blackwellisation for stochastic kinetic model

We demonstrate here the gains arising from the use of a Rao-Blackwellized estimator detailed in Section 3.2.2. We display in Figure 3.8 the variance of the two unbiased estimators of $\mathbb{E}(X_{1,\Delta t}|y_{1:T})$ for $t = \Delta, ..., T\Delta$ and $T = 100$ for the latent Markov jump process and two different values of N, 200 and 3,000. In both cases we set $k = m = 0$. As expected $\bar{H}_{0:0}$ outperforms $H_{0:0}$ but the benefits are much higher for t close to T than when t is close to 1. For example, we see that for $N = 200$

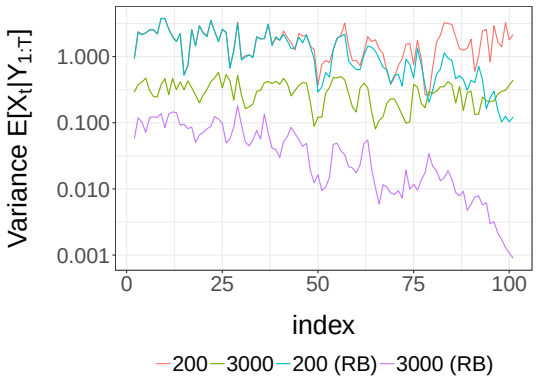
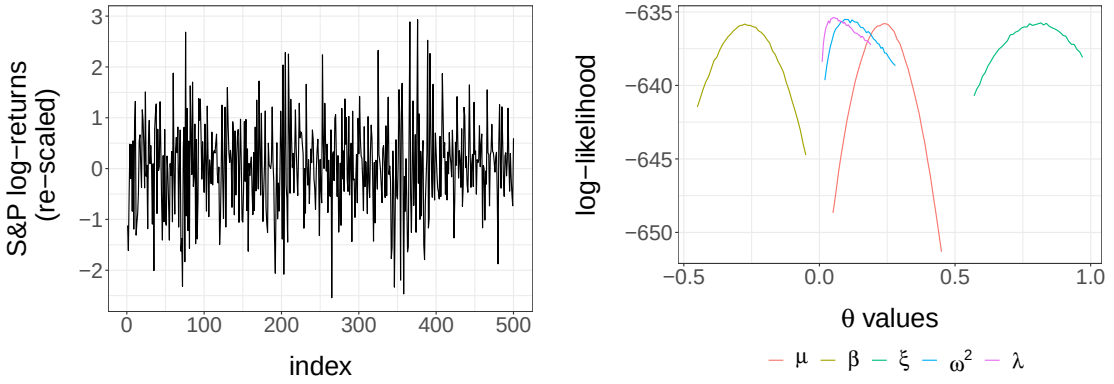


Figure 3.8: Empirical variance of unbiased estimators of $\mathbb{E}(X_{1,\Delta t}|y_{1:T})$: $H_{0:0}$ and Rao-Blackwellised (RB) estimator $\bar{H}_{0:0}$ for stochastic kinetic model.



(a) Daily returns of S&P 500 data

(b) Particle estimates of the log-likelihood function

Figure 3.9: Data and parameter estimation for Lévy-driven stochastic volatility model.

the estimators $H_{0:0}$ and $\bar{H}_{0:0}$ coincide for $t \leq 35$. This is an expected consequence of the particle path degeneracy problem Doucet and Johansen (2011), Jacob et al. (2015), Kantas et al. (2015), with many particles $(X^i_{1:T})_{i\in[N]}$ obtained by the PF at time T sharing common ancestors for t close to 1 when N is too small; see Jacob et al. (2015) for results on the corresponding coalescent time.

3.6.4 Data and parameter estimation

The data used comprised of $T = 500$ log-returns of the S&P 500, starting from the 3^{rd} January 2005, with the scaled data taken from the stochastic volatility example used to demonstrate SMC^2 in Chopin et al. (2013). We plot the raw data in Figure 3.9a. Parameters were estimated using a two-stage procedure,

with SMC^2 used to find a region of high marginal likelihood under the model. Further refinement was performed to compute the maximum likelihood estimator (MLE) $\hat{\theta}$ of $\theta = (\mu, \beta, \xi, \omega^2, \lambda)$ using a grid search around values close to the optimum using $N = 10,000$ particles. We obtained $\hat{\theta} = (0.24, -0.28, 0.82, 0.09, 0.05)$. Likelihood curves around the optimal values are shown in Figure 3.9b, where for each parameter component the log-likelihood was varied while keeping the other parameters fixed at $\hat{\theta}$.

3.7 Application of unbiased estimation to SMC samplers

SMC samplers are a class of SMC algorithms that can be used in Bayesian inference to approximate expectations w.r.t. complex posteriors for static models Del Moral et al. (2006). We show here how we can directly use the methodology proposed in this paper to obtain unbiased estimators of these expectations.

3.7.1 Bayesian computation using SMC samplers

Assume one is interested in sampling from the posterior density $\pi(x) \propto \nu(x)L(x)$ where $\nu(x)$ the prior density w.r.t. a suitable dominating measure and $L(x)$ is the likelihood. We also assume that one can sample from ν. To approximate π, a specific version of SMC samplers introduces a sequence of $T-1$ intermediate densities π_t for $t = 2, ..., T$ bridging ν to π using

$$\gamma_t(x) = \nu(x)L^{\beta_t}(x), \qquad \pi_t(x) = \frac{\gamma_t(x)}{\mathcal{Z}_t},$$

where $\beta_1 = 0 < \beta_2 < ... < \beta_T = 1$. The choice of the sequence $\{\beta_t : t = 2, ..., T-1\}$ can be guided using a preliminary adaptive SMC scheme, as in Zhou et al. (2016), which should subsequently be fixed to preserve unbiasedness of the normalizing constant estimate and validity of the resulting PIMH. In SMC samplers, particles are initialized at time $t = 1$ by sampling from the prior ensuring $w_1(x_1) = 1$.

At time $t \geq 2$, particles are sampled according to an MCMC kernel leaving π_{t-1} invariant and are then weighted according to

$$w_t(x_{t-1}, x_t) = \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})} = L^{\beta_t - \beta_{t-1}}(x_{t-1}).$$

Particles are resampled according to these weights and we set

$$\mathcal{Z}_{t,N} = \mathcal{Z}_{t-1,N} \cdot \frac{1}{N} \sum_{i=1}^{N} w_t(X_{t-1}^{A_{t-1}^i}, X_t^i),$$

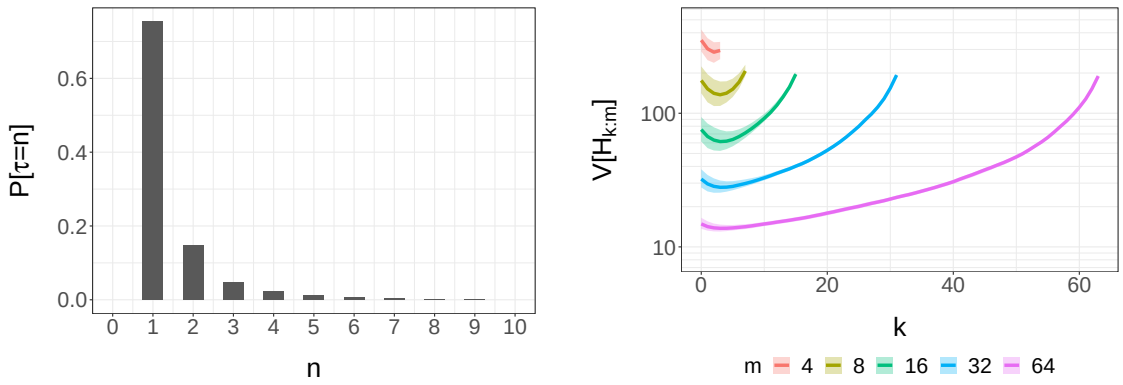
with $\mathcal{Z}_{1,N} = 1$. At time T, $\pi_N(\mathrm{d}x_T) := \sum_{i=1}^{N} W_T^i \delta_{X_T^i}(\mathrm{d}x_T)$ provides a Monte Carlo approximation of the distribution π and $\mathcal{Z}_{T,N}$ plays the role of $p_N(y_{1:T})$, approximating the normalizing constant $\mathcal{Z}_T$ of $\pi = \pi_T$. If no resampling is used, this specific version of SMC samplers coincides with AIS Neal (2001) in which case $\mathcal{Z}_{T,N}$ is given by the average of the product of the incremental weights from time $t = 1$ to $t = T$ instead of the product of the averaged incremental weights. We can use this SMC sampler algorithm or AIS directly within the coupled PIMH scheme, replacing $p_N(y_{1:T})$ by $\mathcal{Z}_{T,N}$ in the acceptance probabilities. We see that $\sup_{(x,x') \in \mathsf{X}^2} w_t(x, x') < \infty$ provided that $\sup_{x \in \mathsf{X}} L(x) < \infty$. Under this condition, if Assumption 6 is satisfied then the estimator $\bar{H}_{k:m}$ of $\pi(h)$ is unbiased and has finite variance and finite expected cost.

3.7.2 Numerical example

We use here coupled PIMH to debias expectations w.r.t. the posterior distribution for a Bayesian mixture model discussed in Lee et al. (2010). We have

$$L(x) = \prod_{n=1}^{M} \left(\frac{1}{D} \sum_{i=1}^{D} \mathcal{N}(y_n; x^i, \sigma^2) \right)$$

with $x := (x^1, ..., x^D) \in \mathbb{R}^D$ constituting the unknown mean components. We consider here $D = 4$ mixture components and $M = 100$ observations. A uniform prior distribution is placed on x over the hypercube $[-10, 10]^D$. The resulting posterior distribution is multimodal. We set $\sigma = 1$ and simulate observations from the model with true values $x^* = (-3, 0, 3, 6)$. We adopt a symmetric random



(a) Empirical distribution of meeting times for SMC sampler with $N = 100$ over 10,000 independent runs. The estimated 95^{th} and 99^{th} percentiles were 6 and 13 respectively.

(b) $\mathbb{V}[\bar{H}_{k:m}]$ of Rao-Blackwellised unbiased estimators of $\pi(h)$ for SMC sampler as a function of k for a range of values of m. The shaded regions correspond to the 1^{st} and 99^{th} percentiles of the variance estimator.

Figure 3.10: SMC sampler unbiased estimators

walk for the Metropolis–Hastings proposals with identity covariance and pick $\beta_t = \left(\frac{t-1}{T-1}\right)^2$ for $T = 200$.

We simulate $10,000$ estimators with $m = 64$ and $N = 100$, after which we are able to estimate variance of test functions for a range of values of k and m noting that there will be some correlation introduced between estimators. The results are shown in Figure 3.10 where we plot the meeting times of the unbiased estimators in Figure 3.10a and the variance of the estimators for a range of values of m in Figure 3.10b using $h : x \mapsto x^1 + x^2 + (x^1)^2 + (x^2)^2$. Uncertainty in the estimated values of $\mathbb{V}[\bar{H}_{k:m}(h)]$ was obtained using $1,000$ bootstrap samples, resampling $10,000$ of the $10,000$ unbiased estimators with replacement. The figure shows the variance of these estimators for a range of values of $m \in \{4, 8, ..., 64\}$ while varying $k \in \{0, ..., m-1\}$. We see, firstly, as expected that as m increases the variance of the estimators reduces. Secondly, for each value of m we see that there exists an optimal value of k, however, as m increases the optimum becomes less pronounced, suggesting that as m increases there is a degree of insensitivity to the choice of k. Finally, for the range of m considered, the optimal values of k appear in a comparatively small interval close to the origin, suggesting that it is not necessary to use large values of k to reduce the variance contribution arising from the bias correction.

4

Unbiased Markov chain Monte Carlo for intractable target distributions

Lawrence Middleton, George Deligiannidis, Arnaud Doucet, Pierre E. Jacob

Abstract

Performing numerical integration when the integrand itself cannot be evaluated point-wise is a challenging task that arises in statistical analysis, notably in Bayesian inference for models with intractable likelihood functions. Markov chain Monte Carlo (MCMC) algorithms have been proposed for this setting, such as the pseudo-marginal method for latent variable models and the exchange algorithm for a class of undirected graphical models. As with any MCMC algorithm, the resulting estimators are justified asymptotically in the limit of the number of iterations, but exhibit a bias for any fixed number of iterations due to the Markov chains starting outside of stationarity. This "burn-in" bias is known to complicate the use of parallel processors for MCMC computations. We show how to use coupling techniques to generate unbiased estimators in finite time, building on recent advances for generic MCMC algorithms. We establish the theoretical validity of some of these procedures, by extending existing results to cover the case of polynomially ergodic Markov chains. The efficiency of the proposed estimators is compared with that of standard MCMC estimators, with theoretical arguments and numerical experiments including state space models and Ising models.

4.1 Introduction

4.1.1 Context

For various statistical models the likelihood function cannot be computed pointwise, which prevents the use of standard Markov chain Monte Carlo (MCMC) algorithms such as Metropolis–Hastings (MH) for Bayesian inference. For example, the likelihood of latent variable models typically involves an intractable integral over the latent space. Classically, one can address this problem by designing MCMC algorithms on the joint space of parameters and latent variables. However, these samplers can mix poorly when latent variables and parameters are strongly correlated under the joint posterior distribution. Furthermore these schemes cannot be implemented if we can only simulate the latent variables and not evaluate their probability density function (Andrieu et al. 2010; Section 2.3). Similarly, in the context of undirected graphical models, the likelihood function might involve an intractable integral over the observation space; see Møller et al. (2006) with examples from spatial statistics.

Pseudo-marginal methods have been proposed for these situations (Lin et al. 2000, Beaumont 2003, Andrieu and Roberts 2009), whereby unbiased Monte Carlo estimators of the likelihood are used within an MH acceptance mechanism while still producing chains that are ergodic with respect to the exact posterior distribution of interest, denoted by π. Pseudo-marginal algorithms and their extensions (Deligiannidis et al. 2015, Tran et al. 2016) are particularly adapted to latent variable models, such as random effects models and state space models, where the likelihood can be estimated without bias using importance sampling or particle filters (Beaumont 2003, Andrieu and Roberts 2009, Andrieu et al. 2010). Related schemes include the exchange algorithm (Murray et al. 2006, Andrieu et al. 2018a), which applies to scenarios where the likelihood involves an intractable, parameter-dependent normalizing constant. Exchange algorithms rely on simulation of synthetic observations to cancel out intractable terms in the MH acceptance ratio. As with any MCMC algorithm, the computation of each iteration requires the

completion of the previous ones, which hinders the potential for parallel computation. Running independent chains in parallel is always possible, and averaging over independent chains leads to a linear decrease of the resulting variance. However, the inherent bias that comes from starting the chains outside of stationarity, also called the "burn-in bias", remains (Rosenthal 2000).

This burn-in bias has motivated various methodological developments in the MCMC literature; among these, some rely on coupling techniques, such as the circularly-coupled Markov chains of Neal (2017), regeneration techniques described in Mykland et al. (1995), Brockwell and Kadane (2005), and "coupling from the past" as proposed in Propp and Wilson (1996). Coupling methods have also been proposed for diagnosing convergence in Johnson (1996; 1998) and as a means to assess the approximation error for approximate MCMC kernels in Nicholls et al. (2012). Recently, a method has been proposed to completely remove the bias of Markov chain ergodic averages (Glynn and Rhee 2014). An extension of this approach using coupling ideas was proposed by Jacob et al. (2017) and applied to a variety of MCMC algorithms. This methodology involves the construction of a pair of Markov chains, which are simulated until an event occurs. At this point, a certain function of the chains is returned, with the guarantee that its expectation is exactly the integral of interest. The output is thus an unbiased estimator of that integral. Averaging over i.i.d. copies of such estimators we obtain consistent estimators in the limit of the number of copies, which can be generated independently in parallel. Relevant limit theorems have been established in Glynn and Heidelberger (1990), Glynn and Whitt (1992), enabling the construction of valid confidence intervals. The methodology has already been demonstrated for various MCMC algorithms (Jacob et al. 2017, Heng and Jacob 2017, Jacob et al. 2018), which were instances of geometrically ergodic Markov chain samplers under typical conditions. However, in the case of intractable likelihoods and pseudo-marginal samplers, the associated Markov chains are typically sub-geometrically ergodic (Andrieu and Vihola 2015).

We show here that unbiased estimators of $\pi(h)$, with finite variance and finite computational cost, can also be derived from polynomially ergodic Markov chains

such as those generated by pseudo-marginal methods. We provide results on the associated efficiency in comparison with standard MCMC estimators. We apply the methodology to particle MCMC algorithms for inference in generic state space models, with an application to a time series of neuron activation counts. We also consider a variant of the pseudo-marginal approach known as the block pseudo-marginal approach (Tran et al. 2016) as well as the exchange algorithm (Murray et al. 2006).

Accompanying code used for simulations and to generate the figures are provided at `https://github.com/lolmid/unbiased_intractable_targets`.

4.1.2 Unbiased estimators from coupled Markov chains

Let π be a probability measure on a topological space $\mathcal{Z}$ equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{Z})$. In this section we recall how two coupled chains that are marginally converging to π can be used to produce unbiased estimators of expectations $\pi(h) := \int h(z)\,\pi(dz)$ for any π-integrable test function $h : \mathcal{Z} \to \mathbb{R}$. Following Glynn and Rhee (2014), Jacob et al. (2017), we consider the following coupling of two Markov chains $(Z_n)_{n\geq 0}$ and $(\tilde{Z}_n)_{n\geq 0}$. First, $Z_0, \tilde{Z}_0$ are drawn independently from an initial distribution π_0. Then, Z_1 is drawn from a Markov kernel P given Z_0, which is denoted $Z_1|Z_0 \sim P(Z_0, \cdot)$. Subsequently, at step $n \geq 1$, a pair $(Z_{n+1}, \tilde{Z}_n)$ is drawn from a Markov kernel $\bar{P}$ given $(Z_n, \tilde{Z}_{n-1})$, which is denoted $(Z_{n+1}, \tilde{Z}_n)|(Z_n, \tilde{Z}_{n-1}) \sim \bar{P}((Z_n, \tilde{Z}_{n-1}), \cdot)$. The kernel $\bar{P}$ is such that, marginally, $Z_{n+1}|Z_n \sim P(Z_n, \cdot)$ and $\tilde{Z}_n|\tilde{Z}_{n-1} \sim P(\tilde{Z}_{n-1}, \cdot)$. This implies that, marginally for all $n \geq 0$, Z_n and $\tilde{Z}_n$ have the same distribution. Furthermore, the kernel $\bar{P}$ is constructed so that there exists a random variable τ termed the meeting time, such that for all $n \geq \tau$, $Z_n = \tilde{Z}_{n-1}$ almost surely (a.s.). Then, for any integer k, the following informal telescoping sum argument informally suggests an unbiased

estimator of $\pi(h)$. We start from $\pi(h) = \lim_{n\to\infty} \mathbb{E}[h(Z_n)]$ and write

$$\begin{aligned}
\pi(h) &= \mathbb{E}[h(Z_k)] + \sum_{n=k+1}^{\infty} \mathbb{E}[h(Z_n)] - \mathbb{E}[h(\tilde{Z}_{n-1})] \quad \text{(write as telescoping sum)}, \\
&= \mathbb{E}[h(Z_k) + \sum_{n=k+1}^{\infty} h(Z_n) - h(\tilde{Z}_{n-1})] \qquad \text{(swap expectation \& limit)}, \\
&= \mathbb{E}[h(Z_k) + \sum_{n=k+1}^{\tau-1} h(Z_n) - h(\tilde{Z}_{n-1})] \qquad (Z_n = \tilde{Z}_{n-1} \text{ for } n \geq \tau).
\end{aligned}$$

The sum $\sum_{n=k+1}^{\tau-1}$ is treated as zero if $\tau - 1 < k + 1$. The suggested estimator is thus defined as

$$H_k(Z, \tilde{Z}) = h(Z_k) + \sum_{n=k+1}^{\tau-1} \{h(Z_n) - h(\tilde{Z}_{n-1})\}, \tag{4.1}$$

with Z and $\tilde{Z}$ denoting the chains $(Z_n)_{n\geq 0}$ and $(\tilde{Z}_n)_{n\geq 0}$ respectively. As in Jacob et al. (2017), we average $H_l(Z, \tilde{Z})$ over a range of values of l, $l \in \{k, k+1, ..., m\}$ for an integer $m \geq k$, resulting in the estimator

$$H_{k:m}(Z, \tilde{Z}) = \frac{1}{m-k+1} \sum_{l=k}^{m} h(Z_l) + \sum_{n=k+1}^{\tau-1} \min\left(1, \frac{n-k}{m-k+1}\right) (h(Z_n) - h(\tilde{Z}_{n-1})). \tag{4.2}$$

Intuitively, $H_{k:m}$ can be understood as a standard Markov chain average after m steps using a burn-in period of $k-1$ steps (which would be in general biased for $\pi(h)$), plus a second term that can be shown to remove the burn-in bias. That "bias correction" term is a weighted sum of differences of the chains between step k and the meeting time $\tau = \inf\{n \geq 1 : \ Z_n = \tilde{Z}_{n-1}\}$. In the following, we will write $H_{k:m} := H_{k:m}(Z, \tilde{Z})$ for brevity. The construction of $H_{k:m}$ is summarized in Algorithm 7, where the initial distribution of the chains is denoted by π_0, and the Markov kernels by P and $\bar{P}$ as above. Standard MCMC estimators require the specification of π_0 and P, while the proposed method requires the additional specification of the coupled kernel $\bar{P}$. We will propose coupled kernels for the setting of intractable likelihoods, and study the estimator $H_{k:m}$ under conditions which cover pseudo-marginal methods.

Algorithm 7 Unbiased MCMC estimator $H_{k:m}$ for any choice of k and m with $0 \leq k \leq m$.

1. Initialization:
   (a) Sample $Z_0, \tilde{Z}_0 \sim \pi_0(\cdot)$.
   (b) Sample $Z_1|\{Z_0 = z_0\} \sim P(z_0, \cdot)$.
   (c) Set $n = 1$ and $\tau = \infty$.
2. While $n < \max(m, \tau)$:
   (a) Sample $(Z_{n+1}, \tilde{Z}_n)|\{Z_n = z_n, \tilde{Z}_{n-1} = \tilde{z}_{n-1}\} \sim \bar{P}\left((z_n, \tilde{z}_{n-1}), \cdot\right)$.
   (b) If $Z_{n+1} = \tilde{Z}_n$ and $\tau = \infty$, set $\tau = n$.
   (c) Increment n by 1.
3. Return $H_{k:m}$ as described in Equation (4.2).

To see how coupled kernels can be constructed, we first recall a construction for simple MH kernels. Assume that $\pi(dz) = \pi(z)\,dz$ where dz is the Lebesgue measure. The standard MH algorithm relies on a proposal distribution $q(dz'|z) = q(z'|z)dz'$, for instance chosen as a Gaussian distribution centered at z. At iteration $n-1$, a proposal $Z' \sim q(\cdot|Z_{n-1})$ is accepted as the new state Z_n with probability $\alpha_{\mathrm{MH}}(Z_{n-1}, Z') := \min\left(1, \pi(Z')q(Z_{n-1}|Z')/\pi(Z_{n-1})q(Z'|Z_{n-1})\right)$, known as the MH acceptance probability. If Z' is rejected, then Z_n is assigned the value of Z_{n-1}. This defines the kernel P. To construct $\bar{P}$, following Jacob et al. (2017) we can consider a maximal coupling of the proposal distributions. This is described in Algorithm 8 for completeness; see also Johnson (1998). Here $\mathcal{U}[a, b]$ refers to the uniform distribution on the interval $[a, b]$. The algorithm relies on draws from a maximal coupling (or γ-coupling) of the two proposal distributions $q(\cdot|Z_n)$ and $q(\cdot|\tilde{Z}_{n-1})$ at step $n \geq 1$. Draws $(Z', \tilde{Z}')$ from maximal couplings are such that the probability of the event $\{Z' = \tilde{Z}'\}$ is maximal over all couplings of $Z' \sim q(\cdot|Z_n)$ and $\tilde{Z}' \sim q(\cdot|\tilde{Z}_{n-1})$. Sampling from maximal couplings can be done with rejection sampling techniques as described in Jacob et al. (2017), in Section 4.5 of Chapter 1 of Thorisson (2000) and in Johnson (1998). On the event $\{Z' = \tilde{Z}'\}$, the two chains are given identical proposals, which are then accepted or not based on $\alpha_{\mathrm{MH}}(Z_n, Z')$ and $\alpha_{\mathrm{MH}}(\tilde{Z}_{n-1}, \tilde{Z}')$ using a common uniform random number. In the event that both proposals are identical and accepted, then the chains meet: $Z_{n+1} = \tilde{Z}_n$. One can then check that the chains remain identical from that iteration onwards.

Algorithm 8 Sampling from the coupled MH kernel given $(Z_n, \tilde{Z}_{n-1})$**.**

1. Sample Z' and $\tilde{Z}'$ from a maximal coupling of $q(\cdot|Z_n)$ and $q(\cdot|\tilde{Z}_{n-1})$.
2. Sample $\mathfrak{u} \sim \mathcal{U}[0,1]$.
3. If $\mathfrak{u} < \alpha_{\mathrm{MH}}(Z_n, Z')$ set $Z_{n+1} = Z'$. Otherwise set $Z_{n+1} = Z_n$.
4. If $\mathfrak{u} < \alpha_{\mathrm{MH}}(\tilde{Z}_{n-1}, \tilde{Z}')$ set $\tilde{Z}_n = \tilde{Z}'$. Otherwise set $\tilde{Z}_n = \tilde{Z}_{n-1}$.
5. Return $(Z_{n+1}, \tilde{Z}_n)$.

The unbiased property of $H_{k:m}$ has an important consequence for parallel computation. Consider R independent copies, denoted by $(H_{k:m}^{(r)})$ for $r = 1, \ldots, R$, and the average $\bar{H}_{k:m}^R = R^{-1} \sum_{r=1}^R H_{k:m}^{(r)}$. Then $\bar{H}_{k:m}^R$ is a consistent estimator of $\pi(h)$ as $R \to \infty$, for any fixed (k, m), and a central limit theorem holds provided that $\mathbb{V}[H_{k:m}] < \infty$; sufficient conditions are given in Section 4.1.3. Since τ is a random variable, the cost of generating $H_{k:m}$ is random. Neglecting the cost of drawing from π_0, the cost amounts to that of one draw from the kernel P, $\tau - 1$ draws from the kernel $\bar{P}$, and then $(m - \tau)$ draws from P if $\tau < m$. Overall that leads to a cost of $T_m := 2(\tau - 1) + \max(1, m - \tau + 1)$ units, where each unit is the cost of drawing from P, and assuming that one sample from $\bar{P}$ costs two units. Theoretical considerations on variance and cost will be useful to guide the choice of the parameters k and m as discussed in Section 4.1.5.

4.1.3 Theoretical validity under polynomial tails

We provide here sufficient conditions under which the estimator $H_{k:m}$ is unbiased, has finite expected cost and finite variance. Below, Assumptions 12 and 14 are identical to Assumptions 2.1 and 2.3 in Jacob et al. (2017) whereas Assumption 13 is a polynomial tail assumption on the meeting time weaker than the geometric tail assumption, namely, $\mathbb{P}(\tau > n) \leq K\rho^n$ for all $n \geq 1$, for some constants $K < \infty$ and $\rho \in (0, 1)$, used in Jacob et al. (2017). Relaxing this assumption is useful in our context as the pseudo-marginal algorithm is polynomially ergodic under realistic assumptions (Andrieu and Vihola 2015) and, as demonstrated in Section 4.1.4, this allows the verification of the polynomial tail assumption.

Assumption 12. *Each of the two chains marginally starts from a distribution* π_0*, evolves according to a transition kernel* P *and is such that* $\mathbb{E}[h(Z_n)] \to \pi(h)$ *as*

$n \to \infty$ for a real-valued function h. Furthermore, there exists constants $\eta > 0$ and $D < \infty$ such that $\mathbb{E}[|h(Z_n)|^{2+\eta}] < D$ for all $n \geq 0$.

Assumption 13. *The two chains are such that there exists an almost surely finite meeting time $\tau = \inf\{n \geq 1 : \ Z_n = \tilde{Z}_{n-1}\}$ such that $\mathbb{P}(\tau > n) \leq Kn^{-\kappa}$ for some constants $0 < K < \infty$ and $\kappa > 2\,(2\eta^{-1} + 1)$, where η is as in Assumption 12.*

Assumption 14. *The chains stay together after meeting, i.e. $Z_n = \tilde{Z}_{n-1}$ for all $n \geq \tau$.*

Under Assumption 13, $\mathbb{E}[\tau^p] \leq Kp \sum_{n \geq 0} n^{-\kappa+p-1}$ for all $p \geq 1$ and thus $\mathbb{E}[\tau^p] < \infty$ if $\kappa > p$. As it is assumed that $\kappa > 2\,(2\eta^{-1} + 1)$, this implies that $\mathbb{E}[\tau^p] < \infty$ for $p < 2(2\eta^{-1} + 1)$. In particular, one has $\mathbb{E}[\tau] < \infty$ and thus the computational cost associated with $H_{k:m}$ has a finite expectation. It also implies that τ has a finite second moment.

The following result states that $H_{k:m}$ has not only a finite expected cost but also has a finite variance and that its expectation is indeed $\pi(h)$ under the above assumptions. The proof is provided in Appendix 4.6.1.

Theorem 15. *Under Assumptions 12-13-14, for all $k \geq 0$ and $m \geq k$, the estimator $H_{k:m}$ defined in (4.2) has expectation $\pi(h)$, has a finite expected computing time and admits a finite variance.*

4.1.4 Conditions for polynomial tails

We now proceed to establishing conditions that imply Assumption 13. To state the main result, we put assumptions on the probability of meeting at each iteration. We write $\mathcal{D}$ for the diagonal of the joint space $\mathcal{Z} \times \mathcal{Z}$, that is $\mathcal{D} := \{(z, \tilde{z}) \in \mathcal{Z} \times \mathcal{Z} : z = \tilde{z}\}$ and introduce the measure $\pi_{\mathcal{D}}(dz, d\tilde{z}) := \pi(dz)\delta_z(d\tilde{z})$. In this case, we identify the meeting time τ with the hitting time of the diagonal, $\tau = \tau_{\mathcal{D}} := \inf\left\{n \geq 1 : \left(Z_n, \tilde{Z}_{n-1}\right) \in \mathcal{D}\right\}$. The first assumption is on the ability of the pair of chains to hit the diagonal when it enters a certain subset of $\mathcal{Z} \times \mathcal{Z}$.

Assumption 16. *The kernel $\bar{P}$ is $\pi_{\mathcal{D}}$-irreducible: for any set $A \subset \mathcal{D}$ such that $\pi_{\mathcal{D}}(A) > 0$ and all $(z, \tilde{z}) \in \mathcal{Z} \times \mathcal{Z}$ there exists some $n \geq 0$ such that $\bar{P}^n((z, \tilde{z}), A) > 0$. The kernel $\bar{P}$ is also aperiodic. Finally, there exist $\epsilon \in (0, 1)$, $n_0 \geq 0$ and a set $C \subset \mathcal{Z}$ such that*

$$\inf_{(z,\tilde{z}) \in C \times C} \bar{P}^{n_0}((z, \tilde{z}), \mathcal{D}) \geq \epsilon. \tag{4.3}$$

Next we will assume that the marginal kernel P admits a polynomial drift condition and a small set C; we will later consider that small set to be the same set C as in Assumption 16. Intuitively, the polynomial drift condition on C will ensure regular entries of the pair of chains in the set $C \times C$, from which the diagonal can be hit in n_0 steps under Assumption 16.

Assumption 17. *There exist $\epsilon_0 > 0$, a probability measure ν on $\mathcal{Z}$ and a set $C \subset \mathcal{Z}$ such that*

$$\inf_{z \in C} P(z, \cdot) \geq \epsilon_0 \nu(\cdot). \tag{4.4}$$

In addition, there exist a measurable function $V : \mathcal{Z} \to [1, \infty)$, constants $b_V, c_V > 0$, $\epsilon_b \in (0, 1)$, and a value $\alpha \in (0, 1)$, such that, defining $\phi(x) := dx^\alpha$ for a constant $d > 0$ and all $x \in [1, \infty)$, then for any $z \in \mathcal{Z}$,

$$PV(z) \leq V(z) - \phi \circ V(z) + b_V \mathbb{1}_C(z), \tag{4.5}$$

$$\sup_{z \in C} V(z) \leq c_V, \tag{4.6}$$

$$\inf_{z \notin C} \phi \circ V(z) \geq b_V (1 - \epsilon_b)^{-1}. \tag{4.7}$$

The following result states that Assumptions 16 and 17 guarantee that the tail probabilities of the meeting time are polynomially bounded. The proof is provided in Appendix 4.6.2.

Theorem 18. *Suppose that Assumptions 16 and 17 hold for the same set $C \subset \mathcal{Z}$, and that π_0 admits a density with respect to π and is supported on a compact set. Then we have that for all $n \geq 1$ and some constant $K > 0$,*

$$\mathbb{P}(\tau \geq n) \leq K n^{-\kappa},$$

where $\kappa = 1/(1 - \alpha)$, with α defined as in Assumption 17.

We note the direct relation between the exponent α in the polynomial drift condition and the exponent κ in the bound on the survival probabilities $\mathbb{P}(\tau \geq n)$. In turn this relates to the existence of finite moments for τ, as discussed after Assumption 13. In particular, if we can take large values of η in Assumption 12, then we require in Assumption 13 that κ is just above 2, which is implied by $\alpha > 1/2$ according to Theorem 18. However, if we consider $\eta = 1$ in Assumption 12, for instance, then we require in Assumption 13 that κ is just above 6, which is implied by $\alpha > 5/6$ according to Theorem 18. The condition $\alpha > 5/6$ will appear again in the next section.

4.1.5 Efficiency under polynomial tails

In removing the bias from MCMC estimators, we expect that $H_{k:m}$ will have an increased variance compared to an MCMC estimator with equivalent cost. In this section we study the overall efficiency of $H_{k:m}$ in comparison to standard MCMC estimators. This mirrors Proposition 3.3 in Jacob et al. (2017) in the case of geometrically ergodic chains.

We can define the inefficiency of the estimator $H_{k:m}$ as the product of its variance and of its expected computational cost via $\mathrm{IF}[H_{k:m}] := \mathbb{E}[T_m]\mathbb{V}[H_{k:m}]$, with T_m denoting the computational cost. This quantity appears in the study of estimators with random computing costs, since seminal works such as Glynn and Heidelberger (1990) and Glynn and Whitt (1992). The inefficiency can be understood as the asymptotic variance of the proposed estimator as the computing budget goes to infinity. The following provides a precise comparison between this inefficiency and the inefficiency of the standard "serial" algorithm. Since the cost T_m is measured in units equal to the cost of sampling from P, the cost of obtaining a serial MCMC estimator based on m iterations is equal to m such units. The mean squared error associated with an MCMC estimator based on $(Z_n)_{n\geq 0}$ is denoted by $\mathrm{MSE}_{b:m} := \mathbb{E}\left[(\mathrm{MCMC}_{b:m} - \pi(h))^2\right]$, where $\mathrm{MCMC}_{b:m} := (m-b+1)^{-1}\sum_{l=b}^{m} h(Z_l)$ and where $b-1$ denotes the number of discarded iterations. We are particularly interested in the comparison between $\mathrm{IF}[H_{k:m}]$, the inefficiency of the proposed estimator

with parameters k, m, and $\lim_{m\to\infty} m \times \mathrm{MSE}_{b:m}$, the asymptotic inefficiency of the serial MCMC algorithm. Both correspond to asymptotic variances when the computing budget goes to infinity.

We first express the estimator $H_{k:m}$, for $m \geq k \geq 0$ as $\mathrm{MCMC}_{k:m} + \mathrm{BC}_{k:m}$, where the bias correction term is

$$\mathrm{BC}_{k:m} := \sum_{n=k+1}^{\tau-1} \min\left(1, \frac{n-k}{m-k+1}\right)\left(h(Z_n) - h(\tilde{Z}_{n-1})\right). \tag{4.8}$$

Then Cauchy-Schwarz provides a relationship between the variance of $H_{k:m}$, the MCMC mean squared error, and the second moment of the bias-correction term:

$$\mathbb{V}[H_{k:m}] \leq \mathrm{MSE}_{k:m} + 2\sqrt{\mathrm{MSE}_{k:m}\mathbb{E}\left[\mathrm{BC}_{k:m}^2\right]} + \mathbb{E}\left[\mathrm{BC}_{k:m}^2\right]. \tag{4.9}$$

This relationship motivates the study of the second moment of $\mathrm{BC}_{k:m}$. The following result shows that if the Markov chains are mixing well enough, in the sense of the exponent α in the polynomial drift condition of Assumption 17 being close enough to one, then we can obtain a bound on $\mathbb{E}\left[\mathrm{BC}_{k:m}^2\right]$ which is explicit in k and m. The proof can be found in Appendix 4.6.3.

Proposition 19. *Suppose that the marginal chain evolving according to P is ψ-irreducible and that the assumptions of Theorem 18 hold for $5/6 < \alpha \leq 1$ and some measurable function $V : \mathcal{Z} \to [1, \infty)$, such that $S_V := \{z : V(z) < \infty\} \neq \emptyset$. In addition assume that there exists a $\gamma \in (1-\alpha, 1)$ such that $\pi(V^{4\gamma}) < \infty$. Then for any measurable function $h : \mathcal{Z} \to \mathbb{R}$ such that $\sup_{z\in\mathcal{Z}} V(z)^{-\alpha-\gamma+1}|h(z)| < \infty$, and any integers $m \geq k \geq 0$ we have that, for $\kappa := 1/(1-\alpha)$, and a constant $B < \infty$,*

$$\mathbb{E}\left[\mathrm{BC}_{k:m}^2\right] \leq B\left[\frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-k+1)^2}\frac{1}{k^{\kappa/2-3}}\right]. \tag{4.10}$$

The fact that a restriction on the exponent α has to be specified to control the second moment of $\mathrm{BC}_{k:m}$ is to be expected: we have already seen in the previous section that such a restriction is also necessary to apply Theorem 18 to verify Assumption 13 with an adequate exponent κ, which, in turn, leads to a finite variance for $H_{k:m}$ through Theorem 15. The specific condition $5/6 < \alpha \leq 1$ could perhaps be relaxed with a more refined technical analysis, thus we interpret the

condition qualitatively: the chains are allowed to satisfy only a polynomial drift condition but it needs to be "close" enough to a geometric drift condition. With pseudo-marginal MCMC algorithms this can be achieved by tuning the number of Monte Carlo samples per iteration.

It follows from (4.9) and (4.10) that under the assumptions of Proposition 19, we have

$$\mathbb{V}[H_{k:m}] \leq \mathrm{MSE}_{k:m} + 2\sqrt{B\mathrm{MSE}_{k:m}}\sqrt{\frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-k+1)^2}\frac{1}{k^{\kappa/2-3}}} + B\left[\frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-k+1)^2}\frac{1}{k^{\kappa/2-3}}\right]. \tag{4.11}$$

The variance of $H_{k:m}$ is thus bounded by the mean squared error of an MCMC estimator, and additive terms that vanish polynomially when k, $m-k$ and m increase. To compare the efficiency of $H_{k:m}$ to that of MCMC estimators, we add simplifying assumptions as in Jacob et al. (2017). As k increases and for $m \geq k$, we expect $(m-k+1)\mathrm{MSE}_{k:m}$ to converge to $\mathbb{V}[(m-k+1)^{-1/2}\sum_{t=k}^{m} h(Z_t)] := V_{k,m}$ as $m \to \infty$, where $Z_k \sim \pi$. We will make the simplifying assumption that $\mathrm{MSE}_{k:m} \approx V_{k,m}/(m-k+1)$ for k large enough. As the condition $5/6 < \alpha$ is equivalent to $\kappa > 6$, $\mathbb{E}\left[\mathrm{BC}_{k:m}^2\right]$ will be negligible compared to the two other terms appearing on the right hand side of (4.11), so we obtain the approximate inequality

$$\mathbb{E}[2(\tau-1)+\max(1,m-\tau+1)]\mathbb{V}[H_{k:m}] \lessapprox \frac{m}{m-k+1}V_{k,m} + 2m\sqrt{BV_{k,m}}\sqrt{\frac{1}{(m-k+1)\,m^{\kappa/2-1}} + \frac{1}{(m-k+1)^3}\frac{1}{k^{\kappa/2-3}}},$$

where the cost of $H_{k:m}$ is approximated by the cost of m calls to P. For the left-hand side to be comparable to $V_{k,m}$, we can select m as a large multiple of k such that $m/(m-k+1)$ is close to one. The second term on the right-hand side is then negligible as k increases, and we see that the polynomial index determining the rate of decay is monotonic in κ.

4.2 Unbiased pseudo-marginal MCMC

4.2.1 Pseudo-marginal Metropolis–Hastings

The pseudo-marginal approach (Lin et al. 2000, Beaumont 2003, Andrieu and Roberts 2009) generates Markov chains that target a distribution of interest, while using only non-negative unbiased estimators of target density evaluations. For concreteness we focus on target distributions that are posterior distributions in a standard Bayesian framework. The likelihood function associated to data $y \in \mathcal{Y}$ is denoted by $\theta \mapsto p(y|\theta)$, and a prior density $\theta \mapsto p(\theta)$ w.r.t. the Lebesgue measure is assigned to an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^D$. We assume that we can compute a non-negative unbiased estimator of $p(y|\theta)$, for all θ, denoted by $\hat{p}(y|\theta, U)$ where $U \in \mathcal{U} \subset \mathbb{R}^M$ are random variables such that $U \sim m_\theta(du)$, where for any $\theta \in \Theta$, m_θ denotes a Borel probability measure on $\mathcal{U}$. We assume that $m_\theta(du)$ admits a density with respect to the Lebesgue measure denoted by $u \mapsto m_\theta(u)$. The random variables U represent variables required in the construction of the unbiased estimator of $p(y|\theta)$. The pseudo-marginal algorithm targets a distribution with density

$$(\theta, u) \mapsto \pi(\theta, u) = p(\theta \mid y) \frac{\widehat{p}(y \mid \theta, u)}{p(y \mid \theta)} m_\theta(u). \tag{4.12}$$

The generated Markov chain $(Z_n)_{n \geq 0}$ takes values in $\mathcal{Z} = \Theta \times \mathcal{U}$. Since $\int \widehat{p}(y \mid \theta, u) m_\theta(u)\, du = p(y|\theta)$ for all θ, marginally $\pi(\theta) = \int \pi(\theta, u) du = p(\theta \mid y)$, corresponding to the target of interest for the θ component of $(Z_n)_{n \geq 0}$. Sampling from $\pi(d\theta, du)$ is achieved with an MH scheme, with proposal $q(d\theta'|\theta)\, m_{\theta'}(du')$. This results in an acceptance probability that simplifies to

$$\alpha_{\mathrm{PM}}\left\{(\theta, \widehat{p}(y \mid \theta, u)), (\theta', \widehat{p}(y \mid \theta', u'))\right\} := \min\left\{1, \frac{\widehat{p}(y \mid \theta', u') p(\theta') q(\theta|\theta')}{\widehat{p}(y \mid \theta, u) p(\theta) q(\theta'|\theta)}\right\}, \tag{4.13}$$

which does not involve any evaluation of $u \mapsto m_\theta(u)$. Thus the algorithm proceeds exactly as a standard MH algorithm with proposal density $q(\theta'|\theta)$, with the difference that likelihood evaluations $p(y|\theta)$ are replaced by estimators $\hat{p}(y|\theta, U)$ with $U \sim m_\theta(\cdot)$. The performance of the pseudo-marginal algorithm depends on the likelihood estimator: lower variance estimators typically yield ergodic averages

with lower asymptotic variance, but the cost of producing lower variance estimators tends to be higher which leads to a trade-off analyzed in detail in Doucet et al. (2015), Schmon et al. (2018).

In the following we will generically denote by g_θ the distribution of $\widehat{p}(y \mid \theta, U)$ when $U \sim m_\theta(\cdot)$, and for notational simplicity, we might write $\widehat{p}(y \mid \theta)$ instead of $\widehat{p}(y \mid \theta, U)$. The above description defines a Markov kernel P and we next proceed to defining a coupled kernel $\bar{P}$, to be used for unbiased estimation as in Algorithm 7.

4.2.2 Coupled pseudo-marginal Metropolis–Hastings

To define a kernel $\bar{P}$ that is marginally identical to P but jointly allows the chains to meet, we proceed as follows, mimicking the coupled MH kernel in Algorithm 8. First, the proposed parameters are sampled from a maximal coupling of the two proposal distributions. If the two proposed parameters θ' and $\tilde{\theta}'$ are identical, we sample a unique likelihood estimator $\widehat{p}(y \mid \theta') \sim g_{\theta'}(\cdot)$ and we use it in the acceptance step of both chains. Otherwise, we sample two estimators, $\widehat{p}(y \mid \theta') \sim g_{\theta'}(\cdot)$ and $\widehat{p}(y \mid \tilde{\theta}') \sim g_{\tilde{\theta}'}(\cdot)$. Denoting the two states of the chains at step $n \geq 1$ by $(\theta_n, \widehat{p}(y \mid \theta_n))$ and $(\tilde{\theta}_{n-1}, \widehat{p}(y \mid \tilde{\theta}_{n-1}))$, Algorithm 9 describes how to obtain $(\theta_{n+1}, \widehat{p}(y \mid \theta_{n+1}))$ and $(\tilde{\theta}_n, \widehat{p}(y \mid \tilde{\theta}_n))$; thereby describing a kernel $\bar{P}$.

Algorithm 9 Sampling from the coupled pseudo-marginal MH kernel given $\left\{(\theta_n, \widehat{p}(y \mid \theta_n)), (\tilde{\theta}_{n-1}, \widehat{p}(y \mid \tilde{\theta}_{n-1}))\right\}$.

1. Sample θ' and $\tilde{\theta}'$ from a maximal coupling of $q(\cdot|\theta_n)$ and $q(\cdot|\tilde{\theta}_{n-1})$.
2. If $\theta' = \tilde{\theta}'$, then sample $\widehat{p}(y \mid \theta') \sim g_{\theta'}(\cdot)$ and set $\widehat{p}(y \mid \tilde{\theta}') = \widehat{p}(y \mid \theta')$.
   Otherwise sample $\widehat{p}(y \mid \theta') \sim g_{\theta'}(\cdot)$ and $\widehat{p}(y \mid \tilde{\theta}') \sim g_{\tilde{\theta}'}(\cdot)$.
3. Sample $\mathfrak{u} \sim \mathcal{U}[0, 1]$.
4. If $\mathfrak{u} < \alpha_{\mathrm{PM}}\left\{(\theta_n, \widehat{p}(y \mid \theta_n)), (\theta', \widehat{p}(y \mid \theta'))\right\}$ then set $(\theta_{n+1}, \widehat{p}(y \mid \theta_{n+1})) = (\theta', \widehat{p}(y \mid \theta'))$.
   Otherwise, set $(\theta_{n+1}, \widehat{p}(y \mid \theta_{n+1})) = (\theta_n, \widehat{p}(y \mid \theta_n))$.
5. If $\mathfrak{u} < \alpha_{\mathrm{PM}}\left\{(\tilde{\theta}_{n-1}, \widehat{p}(y \mid \tilde{\theta}_{n-1}), (\tilde{\theta}', \widehat{p}(y \mid \tilde{\theta}'))\right\}$ then set $(\tilde{\theta}_n, \widehat{p}(y \mid \tilde{\theta}_n)) = (\tilde{\theta}', \widehat{p}(y \mid \tilde{\theta}'))$.
   Otherwise, set $(\tilde{\theta}_n, \widehat{p}(y \mid \tilde{\theta}_n)) = (\tilde{\theta}_{n-1}, \widehat{p}(y \mid \tilde{\theta}_{n-1}))$.
6. Return $\left\{(\theta_{n+1}, \widehat{p}(y \mid \theta_{n+1})), (\tilde{\theta}_n, \widehat{p}(y \mid \tilde{\theta}_n))\right\}$.

In step 2. of Algorithm 9 the two likelihood estimators $\widehat{p}(y \mid \theta')$ and $\widehat{p}(y \mid \tilde{\theta}')$ can be generated independently, as we will do below for simplicity. They can

also be sampled together in a way that induces positive correlations, for instance using common random numbers and other methods described in Deligiannidis et al. (2015), Jacob et al. (2018). We leave the exploration of possible gains in correlating likelihood estimators in that step as a future avenue of research. An appealing aspect of Algorithm 9, particularly when using independent estimators in step 2., is that existing implementation of likelihood estimators can be readily used. In Section 4.3.2 we will exploit this by demonstrating the use of controlled sequential Monte Carlo (Heng et al. 2017) in the proposed framework. Likewise, one could explore the use of other advanced particle filters such as sequential quasi Monte Carlo (Gerber and Chopin 2015). To summarize, given an existing implementation of a pseudo-marginal kernel, Algorithm 9 involves only small modifications and the extra implementation of a maximal coupling which itself is relatively simple following, for example, Jacob et al. (2017).

4.2.3 Theoretical guarantees

We provide sufficient conditions to ensure that the coupled pseudo-marginal algorithm returns unbiased estimators with finite variance and finite expected computation time, i.e. sufficient conditions to satisfy the requirements of Theorem 18 are provided. By introducing the parameterization $w = \hat{p}(y|\theta, u)/p(y|\theta)$ and using the notation $w \sim \bar{g}_\theta(\cdot)$ when $u \sim m_\theta(\cdot)$, we can rewrite the pseudo-marginal kernel

$$P\left((\theta, w), (d\theta', dw')\right) = q(\theta, \theta')\bar{g}_{\theta'}(w')\,\alpha_{\mathrm{PM}}\left\{(\theta, w), (\theta', w')\right\} d\theta' dw' + \varrho_{\mathrm{PM}}(\theta, w)\,\delta_{(\theta,w)}(d\theta', dw'),$$

where, in this parameterization, we write

$$\alpha_{\mathrm{PM}}\left\{(\theta, w), (\theta', w')\right\} = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{q(\theta', \theta)}{q(\theta, \theta')}\frac{w'}{w}\right\},$$

and $\varrho_{\mathrm{PM}}(\theta, w)$ is the corresponding rejection probability. We first make assumptions about the target and proposal densities.

Assumption 20. *The target posterior density $\theta \mapsto \pi(\theta)$ is strictly positive everywhere and continuously differentiable. Its tails are super-exponentially decaying and have regular contours, that is,*

$$\lim_{|\theta|\to\infty} \frac{\theta}{|\theta|}.\nabla \log \pi(\theta) = -\infty, \qquad \limsup_{|\theta|\to\infty} \frac{\theta}{|\theta|}.\frac{\nabla\pi(\theta)}{|\nabla\pi(\theta)|} < 0,$$

where $|\theta|$ denotes the Euclidean norm of θ. Moreover, the proposal distribution satisfies $q(\theta, A) = \int_A q(\theta' - x)\, d\theta'$ with a bounded, symmetric density q that is bounded away from zero on all compact sets.

We then make assumptions about the moments of the noise.

Assumption 21. *There exist constants $a' > 0$ and $b' > 1$ such that*

$$M_W := ess\, sup_{\theta\in\Theta} \int_{\mathbb{R}^+} \max\left(w^{-a'}, w^{b'}\right) \bar{g}_\theta(dw) < \infty,$$

where the essential supremum is taken with respect to the Lebesgue measure. Additionally the family of distributions defined by the densities $\bar{g}_\theta$ is continuous with respect to θ in the topology of weak convergence.

Both assumptions are used in (Andrieu and Vihola 2015) to establish a drift condition for the pseudo-marginal algorithm. Assumption 20 can be understood as a condition on the 'ideal' algorithm, i.e. if the likelihood could be evaluated exactly, and Assumption 21 ensures the likelihood estimate has neither too much mass around zero nor in the tails. The following proposition follows from establishing minorization conditions for both the pseudo-marginal and coupled pseudo-marginal kernels along with (Andrieu and Vihola 2015; Theorem 38).

Proposition 22. *Under Assumptions 20 and 21 then Equations* (4.5)*,* (4.6) *and* (4.7) *hold for any $\chi \in (0, \min(1, a'))$, $a \in (\chi, a']$ and $b \in (0, b' - \chi)$ for the drift function defined as*

$$V(\theta, w) := \left\{\sup_{\theta\in\mathbb{R}} \pi(\theta)\right\}^{\chi} \pi^{-\chi}(\theta) \max(w^{-a}, w^{b}),$$

where $\alpha = 1 - 1/b$ and $C = \{(\theta, w) \in \Theta \times \mathbb{R}^+ : |\theta| \leq M, w \in [\underline{w}, \overline{w}]\}$ for some constants $M \geq 1$, $\underline{w} \in (0, 1]$ and $\overline{w} > \underline{w}$. Additionally the minorization conditions

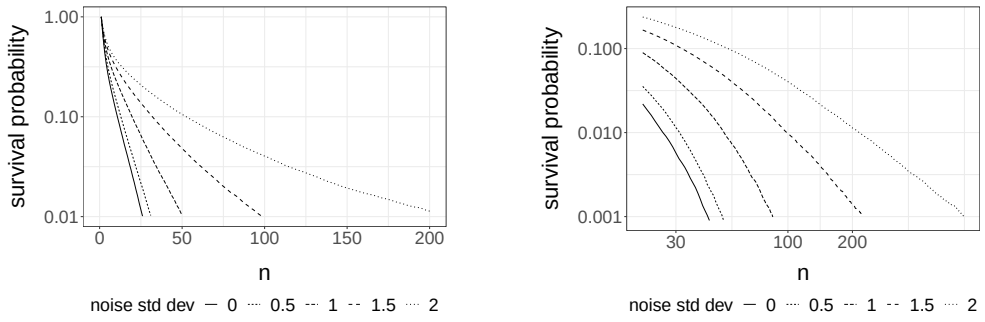
(4.3) *and* (4.4) *hold for the same* C *and* $n_0 = 1$.*Finally if* $\varrho_{\mathrm{PM}}(\theta, w) < 1$ *for all* θ, w *and if for some* $\theta \in B(0, M)$ *we have* $\int_{\underline{w}}^{\overline{w}} \bar{g}_\theta(w) w dw > 0$ *then Assumptions 16 and 17 hold with the same* C *for the kernel* $\bar{P}$ *induced by Algorithm 9.*

If the assumptions of Proposition 22 are satisfied for a', b' such that $b' - \min(1, a') > 6$ then, by application of Theorem 18, the coupling times exhibit the required tail bounds of Assumption 13 with $\alpha > 5/6$ - provided also π_0 admits a density with respect to π and is supported on a compact set. We note that the uniform moments bounds of Assumption 21 might not be satisfied in many non-compact parameter spaces. A weaker assumption allowing to satisfy the polynomial drift condition is provided in (Andrieu and Vihola 2015; Condition 44) and could be alternatively used here.

4.2.4 Tails of meeting times in a toy experiment

We provide numerical experiments on the tails of the meeting time τ in a toy example, to illustrate the transition from geometric to polynomial tails. The target π is a bivariate Normal distribution $\mathcal{N}(\mu, I)$ with $\mu = (1, 2) \in \mathbb{R}^2$, and the initial distribution π_0 is uniform over the unit square. To emulate the pseudo-marginal setting, we pretend that we cannot evaluate $\theta \mapsto \pi(\theta)$; instead we have access to $\hat{\pi}(\theta, W)$ for all θ, where $\hat{\pi}(\theta, W) = \pi(\theta) \times W$ and W is a log-Normal variable: $\log W \sim \mathcal{N}(-\sigma^2/2, \sigma^2)$, and σ calibrates the amount of noise of the unbiased estimator $\hat{\pi}(\theta, W)$ of $\pi(\theta)$. We consider a pseudo-marginal Metropolis–Hastings algorithm with proposal distribution $q(d\theta'|\theta) = \mathcal{N}(d\theta'; \theta, I)$, and a coupled version following Algorithm 9. Indeed, in this setting we are able to verify Assumptions 20 and 21 directly. We note that in the case $\sigma = 0$, we recover the standard MCMC setting.

We draw $R = 10^5$ independent realizations of the meeting time for σ in a grid of values $\{0, 0.5, 1, 1.5, 2\}$. We then approximate survival probabilities $\mathbb{P}(\tau > n)$ by empirical counterparts, for n between 1 and the 99.9% quantile of the meeting times for each σ. The resulting estimates of $\mathbb{P}(\tau > n)$ are plotted against n in Figure 4.1a, where the y-axis is in log-scale. First note that in the case $\sigma = 0$, $\log \mathbb{P}(\tau > n)$ seems



(a) **(b)**

Figure 4.1: Survival probabilities of the meeting time $\mathbb{P}(\tau > n)$ along n, approximated with $10,000$ copies of the meeting times in the pseudo-marginal toy example of Section 4.2.4. Left: y-axis in log-scale and x-axis in natural scale. Right: log-scale for both axes, and restriction to $n \geq 20$, in order to focus on the tails. Each line corresponds to a different value of σ, which calibrates the amount of noise in the estimators of target density evaluations.

to be bounded by a linear function of n, which would correspond to $\mathbb{P}(\tau > n) \leq K\rho^n$ for some constants $K < \infty$ and $\rho \in (0, 1)$. This is indeed the expected behavior in the case of geometrically ergodic Markov chains (Jacob et al. 2017).

As σ increases, $\mathbb{P}(\tau > n)$ decreases less rapidly as a function of n. To verify whether $\mathbb{P}(\tau > n)$ might be bounded by $Kn^{-\kappa}$ (as our theoretical considerations suggest), we plot $\mathbb{P}(\tau > n)$ against n with both axes in log-scale in Figure 4.1b, with a focus on the tails, with $n \geq 20$. The figure confirms that the log survival probability might indeed by upper bounded by $\kappa \log n$, up to a constant offset, for large enough values of n. The figure suggests also that in this case κ decreases with σ.

4.3 Experiments in state space models

State space models are a popular class of time series models. These latent variable models are defined by an unobserved Markov process $(X_t)_{t\geq 0}$ and an observation process $(Y_t)_{t\geq 1}$ where the observations are conditionally independent given $(X_t)_{t\geq 0}$ with

$$X_0 \sim \mu_\theta(\cdot), \qquad X_t|\{X_{t-1} = x\} \sim f_\theta(\cdot|x), \qquad Y_t|\{X_t = x\} \sim g_\theta(\cdot|x), \tag{4.14}$$

where θ parameterizes the distributions μ_θ, f_θ and g_θ (termed the 'initial', 'transition' and 'observation' distribution respectively). Given a realization of the observations

$Y_{1:T} = y_{1:T}$, we are interested in performing Bayesian inference on the parameter θ to which we assign a prior density $p(\theta)$. The posterior density of interest is thus $\pi(\theta) \propto p(\theta)p(y_{1:T}|\theta)$ where the likelihood $p(y_{1:T}|\theta) = \int \mu_\theta(dx_0) \prod_{t=1}^{T} f_\theta(dx_t|x_{t-1})g_\theta(y_t|x_t)$ is usually intractable. It is possible to obtain a non-negative unbiased estimator $\widehat{p}(y \mid \theta, u)$ of $p(y|\theta)$ using particle filtering where here u represents all the random variables simulated during the run of a particle filter. The resulting pseudo-marginal algorithm is known as the particle marginal MH algorithm (PMMH) (Andrieu et al. 2010). This algorithm can also be easily modified to perform unbiased smoothing for state inference and is an alternative to existing methods in Jacob et al. (2018). Guidelines on the selection of the number of particle in this context are provided in Middleton et al. (2019). For state-space models, it is unfortunately extremely difficult to check that Assumptions 20 and 21 are verified.

4.3.1 Linear Gaussian state space model

The following experiments explore the proposed unbiased estimators in a linear Gaussian state space model where the likelihood can be evaluated exactly. This allows a comparison between the pseudo-marginal kernels, that use bootstrap particle filters (Gordon et al. 1993) with N particles to estimate the likelihood, and the ideal kernels that use exact likelihood evaluations obtained with Kalman filters. We assume $X_0 \sim \mathcal{N}(0, 1)$, $X_t|\{X_{t-1} = x\} \sim \mathcal{N}(ax, \sigma_X^2)$ and $Y_t|\{X_t = x\} \sim \mathcal{N}(x, 1)$ where a and σ_X are assigned prior distributions, $a \sim \mathcal{U}[0, 1]$ and $\sigma_X \sim \Gamma(2, 2)$.

4.3.1.1 Effect of the number of particles

A dataset of $T = 100$ observations was generated from the model with parameters $a = 0.5$ and $\sigma_X = 1$. We study how the meeting times and the efficiency vary as a function of N, the number of particles. We set the initial distribution to $\mathcal{U}[0, 1]$ over a and $\mathcal{U}[0, 5]$ over σ_X, and the proposal covariance of the Normal random walk proposals to $0.2^2 I$. In the following we consider a grid of values for the number of particles, varying N between 50 and 250.

We estimate large quantiles of the distribution of the meeting time over 20,000 repetitions of coupled PMMH, with the results shown in Figure 4.2a. As expected, increasing N generally reduces the meeting time at the cost of more computation per iteration.

We examine $\mathrm{IF}[H_{k:m}]$ for the proposed unbiased estimators with $h : x \mapsto x_1 + x_2 + x_1^2 + x_2^2$, for each of these values of N and consider three cases for k and m, in particular

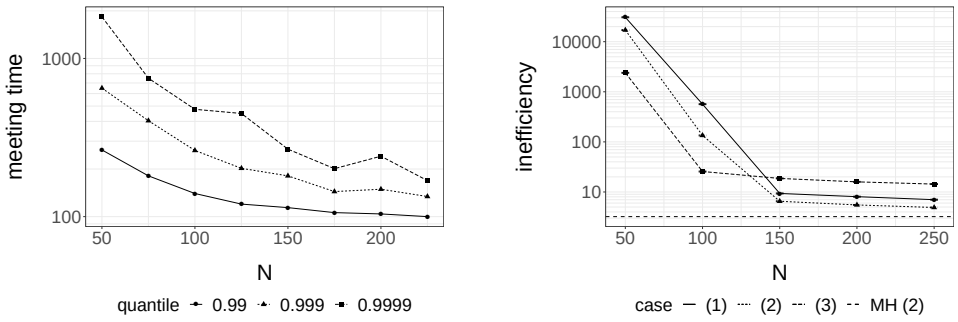
$$(k, m) \in \{(250, 500), (250, 1000), (750, 1000)\}$$

corresponding to the following: (1) a smaller value of $m - k$, (2) a larger value of $m - k$ and (3) a smaller value of $m - k$ with a more conservative choice of k. Estimates of $\mathrm{IF}[H_{k:m}]$ were obtained using 20,000 repetitions of coupled PMMH where for each value of (k, m) estimators were obtained using a single realisation of the largest value of $m = 1,000$ using 30 cores of an Intel Xeon CPU E5-4657L 2.40GHz, taking approximately 60 hours in total.

The results are plotted in Figure 4.2b where we plot also the inefficiency of estimators obtained using coupled Metropolis-Hastings (horizontal line) for (k, m) as in case (2). We see first of all that the inefficiency is reduced by increasing N in all cases, and that the inefficiency of estimators obtained using coupled PMMH asymptotes over this range of N towards the inefficiency of estimators obtained using coupled Metropolis-Hastings for N increasing. We also see that for case (3) that the larger value of k can ameliorate the efficiency of the estimators for small numbers of particles.

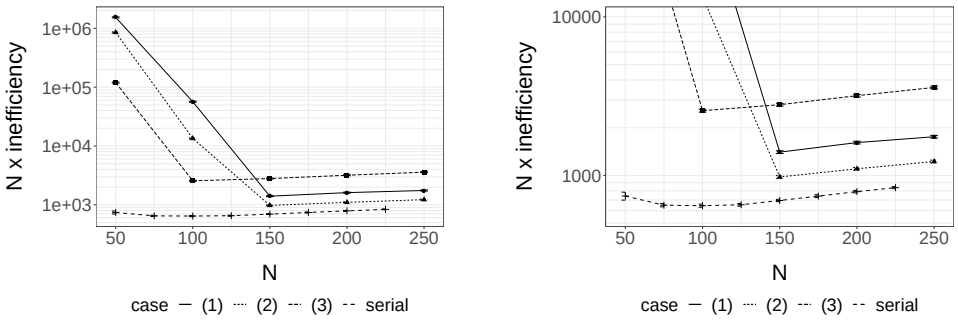
We also examine the inefficiency weighted by the cost of obtaining each estimator, i.e. $N\mathrm{IF}[H_{k:m}]$, and compare this to the inefficiency of the serial algorithm using NV_{as}, with the notation of Section 4.1. Here, V_{as} was estimated using the `spectrum0.ar` function in R's `CODA` package (Plummer et al. 2006), averaging over 10 runs of 500,000 iterations and discarding the first 10% as burn-in. Figure 4.3 shows the results of this procedure, showing ± 2 sample standard errors for the inefficiency estimates. Figure 4.3a demonstrates that despite the lower cost of

obtaining unbiased estimators for lower values of N, the initial decline in inefficiency is still significant. In Figure 4.3b we show the same results though with a focus around the optimum inefficiency. Here, we see that the optimum is attained at $N = 100$ with value $NV_{\text{as}} = 640$ for the serial algorithm and at $N = 150$ with $N\text{IF}[H_{k:m}] = 980$ for case (2). Therefore, we see that the increase in inefficiency is estimated to be under 55% relative to a well-tuned serial algorithm for the values considered. Indeed, for this particular batch of $N = 150$ and $m = 1,000$, the parallel execution time on the stated machine was under 14 hours, which we compare to approximately 12 days of serial execution time (the mean time to obtain an estimator was 53 seconds) or 8 days after accounting for the increase in inefficiency of 55%.



(a) **(b)**

Figure 4.2: Coupled PMMH meeting times and inefficiency of estimators for a linear Gaussian state space model with $T = 100$ observations and over a range of particles, N. Left: estimates of the quantiles of the meeting times. Right: inefficiencies for serial PMMH as a function of N, compared to the inefficiency of unbiased estimators obtained using coupled MH.



(a) **(b)**

Figure 4.3: Inefficiencies weighted by N for a linear Gaussian state space model, comparing directly the inefficiency of estimators obtained using the serial algorithm to those obtained using coupled PMMH. Left: inefficiencies weighted by N. Right: inefficiencies weighted by N close to their optima.

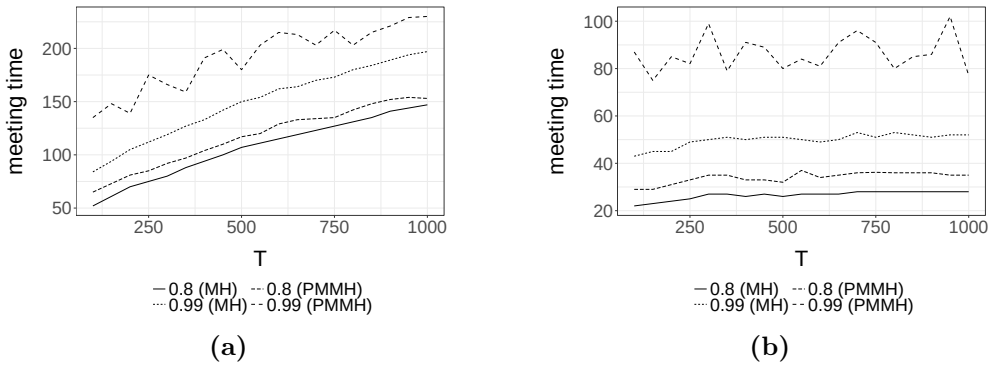
4.3.1.2 Effect of the time horizon

We investigate the distribution of meeting times as a function of T, with N scaling linearly with T. Such a scaling is motivated through the guarantee that the variance of the log-likelihood estimates obtained at each iteration are asymptotically constant (Bérard et al. 2014, Deligiannidis et al. 2015, Schmon et al. 2018). For the model as before, we consider a grid of $T \in \{100, ..., 1000\}$, using a single realisation of the data. Throughout the following, we fix the proposal covariance to be $\frac{2^2}{T}I$, coinciding with the proposal covariance in 4.3.1.1 for $T = 100$.

We consider two cases. Firstly, we examine how the distribution of meeting time changes for a fixed initial distribution (the distribution used previously of $\mathcal{U}[0, 1]$ over a and $\mathcal{U}[0, 5]$ over σ_X); we refer to this as Scaling 1. Secondly, for Scaling 2, we examine how the distribution of meeting times changes if we also scale the initial distribution by setting $\pi_0 = \mathcal{N}(\mu^*, \frac{50}{T}I)$, truncated to ensure it is dominated by the prior and where μ^* denotes the true parameter values.

In both cases we compare the distribution of meeting times for $N = T$ with the distribution of meeting times for the exact algorithm (i.e. $\bar{P}$ as in Algorithm 8) with likelihood evaluations performed using the Kalman filter. Figure 4.4a and 4.4b show estimates of the 80^{th} and 99^{th} percentile over 1,000 repetitions for Scaling 1 and Scaling 2 respectively. Firstly, it can be seen that in all cases the

meeting times for coupled PMMH are higher than the meeting times for coupled MH. Furthermore the smaller difference between the 80^{th} percentiles, compared to the difference between the 99^{th} percentiles, reflects a heavier tail of the distribution of the meeting time in the case of PMMH. Finally, it can be seen that out of the two scalings Scaling 2 appears to stabilise for larger values of T whereas Scaling 1 exhibits an increase with T.



(a) **(b)**

Figure 4.4: Scaling quantiles of meeting times with T over 1,000 repetitions. Left: fixing the initial distribution and scaling the proposals (Scaling 1). Right: scaling both the proposals and the initial distribution (Scaling 2).

4.3.2 Neuroscience experiment

We apply the proposed methodology to a neuroscience experiment described in Temereanca et al. (2008). The same data and model were used to illustrate the controlled Sequential Monte Carlo (cSMC) algorithm in Heng et al. (2017).

4.3.2.1 Model, data and target distribution

The model aims at capturing the activation of neurons of rats as their whiskers are being moved with a periodic stimulus. The experiment involves $M = 50$ repeated experiments, and $T = 3000$ measurements (one per millisecond) during each experiment. The activation of a neuron is recorded as a binary variable for each time and each experiment. These activation variables are then aggregated by summing over the M experiments at each time step, yielding a series of variables Y_t taking values between 0 and M; see Zhang et al. (2018) for an alternative analysis that avoids aggregating over experiments. Letting $\text{Bin}(\cdot; n, p)$ denote the binomial

distribution for n trials with success probability p, the model for neuron activation is given by $X_0 \sim \mathcal{N}(0,1)$ and, for $t \geq 1$,

$$X_t|\{X_{t-1} = x\} \sim \mathcal{N}(\cdot; ax, \sigma_X^2), \quad Y_t|\{X_t = x\} \sim \mathrm{Bin}(\cdot; M, s(x))$$

where $s(x) := (1 + \exp(-s))^{-1}$. We focus on the task of estimating (a, σ_X^2) from the data using the proposed method. Following Heng et al. (2017) we specify a uniform prior on $[0,1]$ for a and an inverse-Gamma prior on σ_X^2 with parameters $(1, 0.1)$, where the probability density function of an inverse-Gamma with parameters (a, b) is $x \mapsto \Gamma(a)^{-1} b^a x^{-a-1} \exp(-b/x)$. The PMMH kernels employed below use a Gaussian random walk proposal. The likelihood is estimated with cSMC with $N = 128$ particles and 3 iterations, where the exact specification is taken from the appendix of Heng et al. (2017). Such cSMC runs take approximately one second, on a 2015 desktop computer and a simple R implementation. Figure 4.5 presents the time series of observations (4.5a) and the estimated log-posterior density (4.5b), obtained on a 500×500 grid of parameter values, and one cSMC likelihood estimate per parameter value. In Figure 4.5b, the upper right corner presents small black circles, generated by the contour plot function, which indicate high variance in the likelihood estimators for these parameters. Thus we expect PMMH chains to have a lower acceptance rate in that part of the space. On the other hand, the maximum likelihood estimate (MLE) is indicated by a black dot on the bottom right corner. The variance of the log-likelihood estimators is of the order of 0.2 around the MLE, so that PMMH chains are expected to perform well there, as was observed in Heng et al. (2017) where the chains were initialized close to the MLE.

4.3.2.2 Standard deviation of the proposal

Here, we initialize the chains from a uniform distribution on $[0,1]^2$, and we investigate two choices of standard deviation for the random walk proposals: the one used in Heng et al. (2017), that is 0.002 for a and 0.01 for σ_X^2, and another choice equal to 0.01 for a and 0.05 for σ_X^2, i.e. five times larger. For each choice, we can run pairs of chains until they meet and record the meeting time; we can do so on P

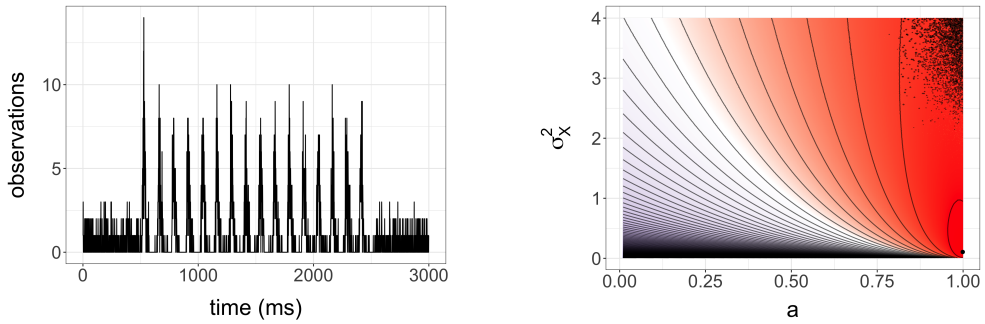
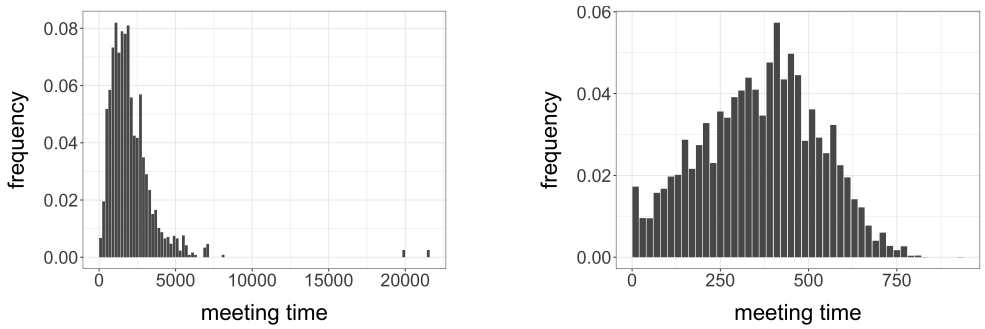


Figure 4.5: Left: counts of neuron activation in 50 experiments, over a duration of three seconds. Right: estimated log-posterior density in the neuroscience experiment of Section 4.3.2.

processors in parallel (e.g. hundreds), and for a certain duration (e.g. a few hours). Thus the number of meeting times produced by each processor is a random variable. Following Glynn and Heidelberger (1990), if no meeting time was produced by a processor within the time budget, the computation continues until one meeting time is produced, otherwise on-going calculations are interrupted when the budget is reached. This allows unbiased estimation of functions of the meeting time on each processor via Corollary 7 of Glynn and Heidelberger (1990), and then we can average across processors. In particular we use this strategy to produce all histograms in the present section, as in Figure 4.6.

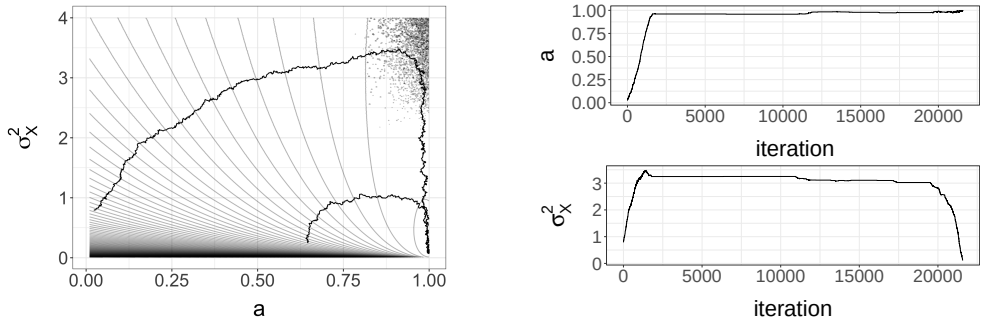
We observe that the meeting times are significatively larger when using the smaller standard deviation (4.6a), with a maximum value of $21,570$ over 1565 realizations. With the larger choice of standard deviation (4.6b), we observe shorter meeting times, with a maximum of 928 over 5572 realizations. This suggests that the values of k and m should be chosen very differently in both cases.

To explain this difference we investigate the realization of the coupled chains that led to the largest meeting time of $21,570$, in Figure 4.7. Figure 4.7a presents the trajectories of two chains overlaid with contours of the target density function. The chains seem to follow approximately the gradient of the density. Given the shape of this density, it means that small starting values for component a result in the chains going to the region of high variance of the likelihood estimator, in the top



(a) **(b)**

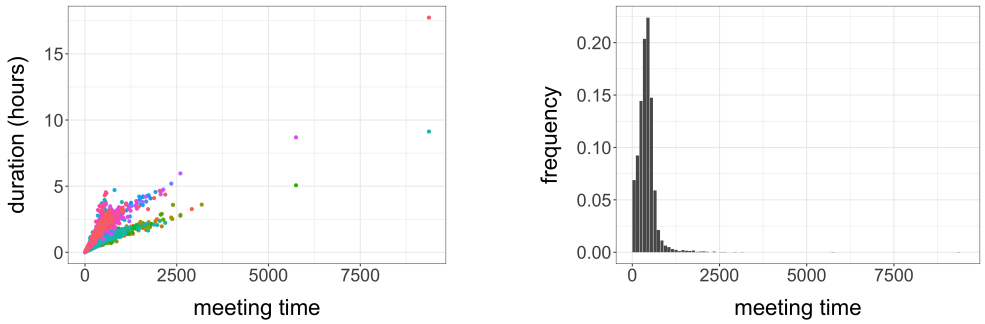
Figure 4.6: Histograms of meeting times associated with coupled PMMH chains, obtained with a standard deviation of the random walk proposal of 0.002 for a and 0.01 for σ_X^2 on the left, and with a larger standard deviation (0.01 on a and 0.05 on σ_X^2) on the right. In both cases, the likelihood was estimated with controlled SMC, with $N = 128$ particles, $I = 3$ iterations, in the neuroscience model of Section 4.3.2.



(a) **(b)**

Figure 4.7: Traces of the chains corresponding to the largest observed meeting time $(21,570)$ obtained with a small standard deviation of the random walk proposal (0.002 for a and 0.01 for σ_X^2), in form of a two-dimensional trajectory on the left, and trace plots of one of the two chains on the right. The likelihood is estimated with controlled SMC, with $N = 128$ particles, $I = 3$ iterations, in the neuroscience experiment of Section 4.3.2.

right corner of the plot. The marginal trace plots of one of the two chains are shown in 4.7b. From the trace plots we see that most of the $21,570$ iterations have been spent in that top right corner, where the chain got stuck, approximately between iterations $2,000$ and $20,000$. The overall acceptance rate is of 6% for that chain, compared to 39% for the other chain shown in 4.7a Therefore the use of a larger proposal standard deviation seems to have a very noticeable effect here on the ability of the Markov chain to escape a region of high variance of the likelihood estimator.



(a) (b)

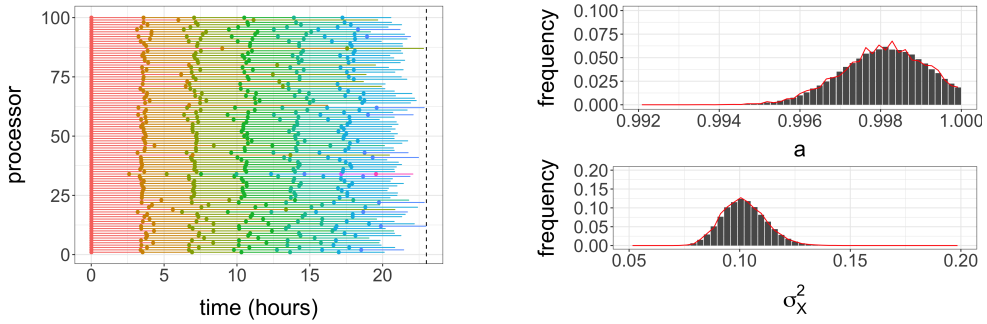
Figure 4.8: Left: duration (in hours) versus meeting times, using BPF with $N = 4,096$ particles. Each color corresponds to a different processor. Right: estimated histogram of the meeting times, in the neuroscience experiment of Section 4.3.2.

4.3.2.3 Comparison with PMMH using bootstrap particle filters

We use the larger choice of standard deviation (0.01 on a and 0.05 on σ_X^2) hereafter, and compare meeting times obtained with cSMC with those obtained with bootstrap particle filters, with $N = 4,096$ particles. This number is chosen so that the compute times are comparable. Over 23 hours of compute time, the number of meeting times obtained per processor varied between 4 and 35, and a total of $7,776$ meeting times were obtained from 400 processors. The meeting times are plotted against the duration it took to produce them in Figure 4.8a. The compute time associated with meeting times is not only proportional to the meeting times themselves, but also varies across processors. This is partly due to hardware heterogeneity across processors, and to concurrent tasks being executed on the cluster during our experiments. The histogram in Figure 4.8b shows that meeting times are larger, and heavier tailed, than when using cSMC (see Figure 4.6b). The maximum observed value is $9,371$. From these plots, we see that to produce unbiased estimators $H_{k:m}$ using BPF with a similar variance as when using cSMC, we would have to choose larger values of k and m, and thus the cost per estimator would likely be higher.

4.3.2.4 Efficiency compared to the serial algorithm

Using cSMC and the larger choice of standard deviation for the proposal, we produce unbiased estimators $H_{k:m}$ with $k = 1,000$ and $m = 10,000$. We run 100



(a) (b)

Figure 4.9: Left: start and end time for the calculation of unbiased estimators on 100 parallel processors, for a budget of 23 hours (dashed line), with cSMC, $N = 128$ particles, $I = 3$ iterations and $k = 1,000$, $m = 10,000$. Right: estimated histograms of both parameters a and σ_X^2; the red lines correspond to estimates obtained with $250,000$ iterations of PMMH and discarding the first $10,000$ as burn-in; this is for the neuroscience experiment of Section 4.3.2.

processors for a time budget of 23 hours, and each processor produced between 2 and 9 estimators, for a total of 578 estimators. The generation of samples for each processor is represented chronologically in Figure 4.9a. The variation among durations is due to the randomness of meeting times and also to external factors such as concurrent tasks being executed on the cluster. We produce histograms of the posterior marginals in Figure 4.9b, with the result from a long run of PMMH with cSMC ($250,000$ iterations) overlaid in red lines.

We compute the loss of efficiency incurred by debiasing the PMMH chain with the proposed estimators. We consider the test function $h : x \mapsto x_1 + x_2 + x_1^2 + x_2^2$. Along the PMMH chain of length $n_{\text{mcmc}} = 250,000$, after a burn-in of $n_{\text{burnin}} = 10,000$ steps, and using the `spectrum0` function of the `CODA` package, we find the asymptotic variance associated with h to be $V_{\text{as}} = 7.53 \cdot 10^{-3}$. If we measure computing cost in terms of MCMC iterations, we obtain an inefficiency of $n_{\text{mcmc}} \times V_{\text{as}}/(n_{\text{mcmc}} - n_{\text{burnin}}) \approx 7.84 \cdot 10^{-3}$. With the unbiased estimators $H_{k:m}$, if the cost of an estimator is $2(\tau - 1) + \max(1, m + 1 - \tau)$, then the average cost per processor is $59,860$. The empirical variance of the unbiased estimators obtained per processor is equal to $1.4 \cdot 10^{-7}$, thus we obtain an inefficiency of $59860 \times 1.4 \cdot 10^{-7} \approx 8.4 \cdot 10^{-3}$. This inefficiency is slightly above $7.84 \cdot 10^{-3}$.

Next, we parameterize cost in terms of time (in seconds) instead of number of MCMC steps. This accounts for the fact that running jobs on a cluster involve heterogeneous hardware and concurrent tasks. The serial PMMH algorithm was run on a desktop computer for $169,952$ seconds and thus the inefficiency might be measured as $169,952 \times V_{\text{as}}/(n_{\text{mcmc}} - n_{\text{burnin}}) \approx 5.3 \cdot 10^{-3}$. Note that each iteration took less than a second on average, because parameter values proposed outside of the support of the prior were rejected before running a particle filter; on the other hand the cost of a cSMC run is above one second on average. For the proposed estimators, the budget was set to 23 hours and we obtained a variance across processors of $1.4 \cdot 10^{-7}$; thus we can compute the inefficiency as $1.16 \cdot 10^{-2}$, which is approximately twice the inefficiency of the serial algorithm.

4.4 Methodological extensions

The following provides two further examples of coupled MCMC algorithms to perform inference when the likelihood function is intractable. The associated estimators are not covered by our theoretical results.

4.4.1 Block pseudo-marginal method

Block pseudo-marginal methods have demonstrated significant computational savings for Bayesian inference for random effects models over standard pseudo-marginal methods (Tran et al. 2016). Such methods proceed through introducing strong positive correlation between the current likelihood estimate $\hat{p}(y|\theta)$ and the likelihood estimate of the proposed parameter $\hat{p}(y|\theta')$ through only modifying a subset of the auxiliary variables used to obtain the likelihood estimate at each iteration. We here demonstrate the computational benefits of such a scheme in obtaining unbiased estimators of posterior expectations.

We focus here on random effects models, i.e. we have for $t = 1, ..., T$

$$X_t \overset{i.i.d.}{\sim} f_\theta(\cdot), \qquad Y_t|\{X_t = x\} \sim g_\theta(\cdot|x). \tag{4.15}$$

In this context, the likelihood of data $y = (y_1, ..., y_T)$ is of the form $p(y|\theta) = \prod_{t=1}^{T} p(y_t|\theta)$ where $p(y_t|\theta) = \int f_\theta(dx) g_\theta(y_t|x)$ and the likelihood estimate is given by $\hat{p}(y|\theta, U) = \prod_{t=1}^{T} \hat{p}(y_t|\theta, U_t)$, where $\{\hat{p}(y_t|\theta, U_t)\}_{t=1,...,T}$ are T independent non-negative unbiased likelihood estimates of $\{p(y_t|\theta)\}_{t=1,...,T}$ when $U_t \sim m_t(\cdot)$. In the following, we provide a minor modification of the blocking strategy proposed in Tran et al. (2016), where instead of jointly proposing a new parameter and a single block of auxiliary random variables, a parameter update is performed, followed by sequentially iterating through the auxiliary random variables used to construct the likelihood estimate of observation t. For each data t, new values are proposed according to $U_t' \sim m_t(\cdot)$ and accepted with probability

$$\alpha_{\mathrm{BPM},t}\left\{\widehat{p}(y_t \mid \theta, U_t), \hat{p}(y_t|\theta, U_t')\right\} = \min\left\{1, \frac{\hat{p}(y_t|\theta, U_t')}{\hat{p}(y_t|\theta, U_t)}\right\}. \tag{4.16}$$

As remarked in Tran et al. (2016), such blocking strategies are generally not applicable to particle filter inference in state space models, whereby likelihood estimates for observation t typically depend on all auxiliary random variables generated up to and including t. We provide pseudo-code for the proposed blocking strategy in Algorithm 10. We denote by $U_{t,n}$ the set of auxiliary variables U_t at iteration n,

Algorithm 10 Sampling from the block pseudo-marginal kernel given $\left(\theta_{n-1}, (U_{t,n-1})_{t\geq 1}\right)$

1. Sample $\theta' \sim q(\cdot|\theta_{n-1})$ and compute $\widehat{p}(y_t \mid \theta', U_{t,n-1})$ for $t = 1, ..., T$.
2. With probability $\alpha_{\mathrm{PM}}\left\{(\theta_{n-1}, \prod_{t=1}^{T} \widehat{p}(y_t \mid \theta_{n-1}, U_{t,n-1})), (\theta', \prod_{t=1}^{T} \widehat{p}(y_t \mid \theta', U_{t,n-1}))\right\}$, set $\theta_n = \theta'$. Otherwise, set $\theta_n = \theta_{n-1}$.
3. For $t = 1, ..., T$
   (a) Sample $U_t' \sim m_t(\cdot)$.
   (b) With probability $\alpha_{\mathrm{BPM,t}}\left\{\widehat{p}(y_t \mid \theta_n, U_{t,n-1}), \widehat{p}(y_t \mid \theta_n, U_t')\right\}$, set $U_{t,n} = U_t'$. Otherwise, set $U_{t,n} = U_{t,n-1}$.

4.4.1.1 Coupled block pseudo-marginal method

An algorithm to couple two block pseudo-marginal algorithms to construct unbiased estimators $H_{k:m}$ is provided in Algorithm 11. Denoting the two states of the chains at step $n \geq 1$ by $(\theta_n, (U_{t,n})_{t\geq 1})$ and $(\tilde{\theta}_{n-1}, (\widetilde{U}_{t,n-1})_{t\geq 1})$, Algorithm 11 describes how to obtain $(\theta_{n+1}, (U_{t,n+1})_{t\geq 1})$ and $(\tilde{\theta}_n, (\widetilde{U}_{t,n})_{t\geq 1})$; thus it describes a kernel $\bar{P}$.

Algorithm 11 Sampling from the coupled block pseudo-marginal kernel given $(\theta_n, (U_{t,n})_{t\geq 1}, \tilde{\theta}_{n-1}, (\widetilde{U}_{t,n-1})_{t\geq 1})$

1. Sample $(\theta', \tilde{\theta}')$ from the maximal coupling of $q(\cdot|\theta_n)$ and $q(\cdot|\tilde{\theta}_{n-1})$.
2. Compute $\widehat{p}(y_t \mid \theta', U_{t,n})$ and $\widehat{p}(y_t \mid \tilde{\theta}', \widetilde{U}_{t,n-1})$ for $t = 1, ..., T$.
3. Sample $\mathfrak{u} \sim \mathcal{U}[0,1]$.
4. If $\mathfrak{u} < \alpha_{\mathrm{PM}} \left\{ \left(\theta_n, \prod_{t=1}^T \widehat{p}(y_t \mid \theta_n, U_{t,n})\right), (\theta', \prod_{t=1}^T \widehat{p}(y_t \mid \theta', U_{t,n})) \right\}$ then set $\theta_{n+1} = \theta'$. Otherwise, set $\theta_{n+1} = \theta_n$.
5. If $\mathfrak{u} < \alpha_{\mathrm{PM}} \left\{ (\tilde{\theta}_{n-1}, \prod_{t=1}^T \widehat{p}(y_t \mid \tilde{\theta}_{n-1}, \widetilde{U}_{t,n-1})), (\tilde{\theta}', \prod_{t=1}^T \hat{p}(y_t|\tilde{\theta}', \widetilde{U}_{t,n-1})) \right\}$ then set $\tilde{\theta}_n = \tilde{\theta}'$. Otherwise, set $\tilde{\theta}_n = \tilde{\theta}_{n-1}$.
6. For $t = 1, ..., T$
   (a) Sample $U'_t \sim m_t(\cdot)$.
   (b) Sample $\mathfrak{u} \sim \mathcal{U}[0,1]$.
   (c) If $\mathfrak{u} < \alpha_{\mathrm{BPM,t}} \{\widehat{p}(y_t \mid \theta_{n+1}, U_{t,n}), \widehat{p}(y_t \mid \theta_{n+1}, U'_t)\}$ then set $U_{t,n+1} = U'_t$. Otherwise, set $U_{t,n+1} = U_{t,n}$.
   (d) If $\mathfrak{u} < \alpha_{\mathrm{BPM,t}} \left\{ \widehat{p}(y_t \mid \tilde{\theta}_n, \widetilde{U}_{t,n-1}), \widehat{p}(y_t \mid \tilde{\theta}_n, U'_t) \right\}$ then set $\widetilde{U}_{t,n} = U'_t$. Otherwise, set $\widetilde{U}_{t,n} = \widetilde{U}_{t,n-1}$.

4.4.1.2 Bayesian multivariate probit regression

The following demonstrates the proposed algorithm for a latent variable model applied to polling data and explores the possible gains when compared to the unbiased estimators obtained using the coupled pseudo-marginal algorithm. The data consists of polling data collected between February 2014 and June 2017 as part of the British Election Study (Fieldhouse et al. 2018). We use a multivariate probit model, which for $i \in \{1, .., T\}$ and $j \in \{1, 2, 3\}$ can be expressed as $X_{ij} = \beta' \zeta_{ij} + \epsilon_{ij}$ and $Y_{ij} = \mathbb{1}[X_{ij} > 0]$ for observed binary response Y_{ij}, latent state X_{ij} and where i indexes the i^{th} participant, j indexes the j^{th} wave of questions, β is a vector of regression coefficients (including an intercept) and ζ_{ij} is a vector of independent variables.

We use a random sample of $T = 2,000$ participants over three waves (one a year) in the run up to the United Kingdom's European Union membership referendum on 23^{rd} June 2016, regressing the binary outcome asking participants how they would vote in an EU referendum against how they perceive the general economic situation in the UK has changed over the previous 12 months (graded 1-5, with 1='Got a lot worse', 5='Got a lot better'). A detailed description of

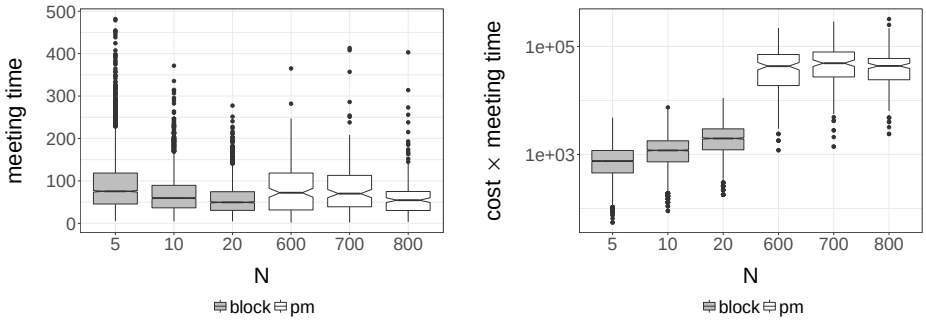
the data is provided in Appendix 4.6.5.

We allow for correlations between waves through modelling the perturbations $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}) \sim \mathcal{N}(0, \Sigma_\rho)$ with a generic correlation matrix Σ_ρ. In total, we have five unknown parameters $\theta = (\beta_1, \beta_2, \rho_{2,1}, \rho_{3,1}, \rho_{3,2})$, with β_1 denoting a regressor coefficient, β_2 a constant offset and $\rho_{s,t}$ element (s,t) of Σ_ρ. We place independent priors on each parameter with $\beta_1, \beta_2 \overset{i.i.d.}{\sim} \mathcal{N}(0, 10^2)$ and $\rho_{s,t} \overset{i.i.d.}{\sim} \mathcal{U}[-1, 1]$, where we additionally truncate the prior on Σ_ρ to ensure support only on the manifold of positive definite matrices.

Inference For each observation $y_i := (y_{i1}, y_{i2}, y_{i3})$, we obtain unbiased estimates of the likelihood of θ using the sequential importance sampling algorithm of Geweke, Hajivassiliou and Keane; see, e.g., Train (2009; 5.6.3) and references therein. We set the initial distribution $\pi_0 = \mathcal{N}(\hat{\mu}, 0.01^2 I)$, supported only on areas of positive mass under the prior (we employ a simple rejection sampling algorithm to sample Σ_ρ initially) and use a Normal random walk proposal with covariance set to $\frac{2.38^2}{5}\hat{\Sigma}$, where $\hat{\mu}$ and $\hat{\Sigma}$ are an empirical estimate of the posterior mean and covariance on a preliminary run of 10,000 iterations of block pseudo-marginal with $N = 40$ discarding the first 10% as burn-in.

We compare coupled block pseudo-marginal with coupled pseudo-marginal. We examine values of N for the latter that are close to the optimal value of N for the serial algorithm, estimated through ensuring the variance of the log-likelihood estimates is between 1 and 2, as per the guidance in Doucet et al. (2015). In this case we consider $N \in \{600, 700, 800\}$, providing a corresponding variance of the log-likelihood estimates given by $\{1.67, 1.40, 1.25\}$ (estimated using 10,000 likelihood estimates at $\hat{\mu}$). For the block pseudo-marginal algorithm we consider $N \in \{5, 10, 20\}$.

Meeting times Both algorithms were run continuously until coupling for half an hour each on a 48 CPU Intel Xeon 2.4Ghz E5-4657L server, with the number of estimators produced for block pseudo-marginal varying between 2,000 and 4,000 for the values of N considered and between 160 and 180 for the pseudo-marginal.



(a) **(b)**

Figure 4.10: Meeting times for coupled block pseudo-marginal and coupled pseudo-marginal algorithms. Left: raw meeting times for the two algorithms. Right: meeting times weighted by cost for the two algorithms, i.e. τN for coupled pseudo-marginal and $2\tau N$ for coupled block pseudo-marginal.

The meeting times are plotted in Figure 4.10a, where it can be seen that despite the lower cost of the block pseudo-marginal algorithm the absolute values of the meeting times are comparable across algorithms.

Accordingly, we also plot the distribution of meeting times accounting for the cost of running each algorithm, i.e. $N\tau$ for the pseudo-marginal algorithm and $2N\tau$ for the block pseudo-marginal algorithm. The additional factor of 2 for the latter can be seen as an upper bound on the additional computational cost of the block pseudo-marginal algorithm, assuming twice the density evaluations per complete iteration and less than twice the number of pseudo-random numbers generated. Figure 4.10b shows the results of this additional cost-weighting where it can be seen that meeting times are between 1 and 2 orders of magnitude larger for the pseudo-marginal over the block pseudo-marginal algorithm.

Variance of estimators We estimate the increase in inefficiency of the coupled over the serial algorithm for $N = 10$, $k = 500$ and $m = 5,000$; the choice of k is guided by the meeting times in Figure 4.10a. Running coupled block pseudo-marginal 200 times, we estimate the variance using the test function $h : x \mapsto \sum_i (x_i + x_i^2)$ to be $1.05 \cdot 10^{-5}$. Estimating the cost of $2(\tau - 1) + \max(1, m + 1 - \tau)$ to be 5121, implies an inefficiency of $5.36 \cdot 10^{-2}$. In comparison, we estimate the inefficiency of the serial algorithm using `spectrum0.ar` as before on runs of length 125,000 (discarding 10%

as burn-in and averaging over 20 estimators) to be $n_{\text{mcmc}} \times V_{as}/(n_{\text{mcmc}} - n_{\text{burnin}}) = 4.82 \cdot 10^{-2}$ suggesting an increase in inefficiency of 11% for the unbiased estimators.

Finally, we compare the inefficiency of unbiased estimators generated with coupled block pseudo-marginal kernels with those produced using standard coupled pseudo-marginal kernels with $N_{\text{PM}} = 700$ particles. For coupled pseudo-marginal, the variance of the unbiased estimator was estimated to be $1.53 \cdot 10^{-5}$ and the expected cost was estimated to be 5147, implying an inefficiency of $7.86 \cdot 10^{-2}$. As a result we estimate the improvement of inefficiency for the coupled block pseudo-marginal by $\frac{N_{\text{PM}}}{2N} \times \frac{7.86 \cdot 10^{-2}}{5.36 \cdot 10^{-2}}$ to be approximately 51 times. Estimation of the asymptotic variance of the serial pseudo-marginal algorithm was computationally infeasible for this many particles, with a single iteration taking on average six seconds on the aforementioned server, hence the choice of N motivated by the guidance in Doucet et al. (2015) instead.

4.4.2 Exchange algorithm

Problems where the likelihood function is only known only up to a constant of proportionality occur frequently across Bayesian statistics; see, e.g., Park and Haran (2018) for a recent account of current methodology and applications. In this case, posterior distributions $\pi(\theta) \propto p(y|\theta)p(\theta)$ are given by

$$p(y|\theta) = \frac{f(y|\theta)}{\mathcal{Z}(\theta)}, \quad \mathcal{Z}(\theta) := \int f(y|\theta)dy,$$

where $f(y|\theta)$ can be evaluated pointwise but its parameter-dependent normalizing constant $\mathcal{Z}(\theta)$ is intractable. This is a scenario common for undirected graphical models and spatial point processes (Møller et al. 2006, Murray et al. 2006). The exchange method detailed in Algorithm 12 is an MCMC scheme proposed by Murray et al. (2006) to sample such distributions under the assumption that, although $\mathcal{Z}(\theta)$ cannot be evaluated, it is possible to simulate exactly artificial observations from $p(y|\theta)$. This is indeed possible for a large class of spatial point processes as well as the Ising and Potts models using perfect simulation procedures.

Algorithm 12 Sampling from the Exchange kernel given θ_{n-1}

1. Sample $\theta' \sim q\left(\cdot|\theta_{n-1}\right)$ and $Y' \sim p(\cdot|\theta')$.
2. With probability

$$\alpha_{\mathrm{EX}}\left(\theta_{n-1}, \theta', Y'\right) := \min\left\{1, \frac{f(y|\theta')p(\theta')f(Y'|\theta_{n-1})q\left(\theta_{n-1}|\theta'\right)}{f(y|\theta_{n-1})p(\theta_{n-1})f(Y'|\theta')q(\theta'|\theta_{n-1})}\right\}, \quad (4.17)$$

set $\theta_n = \theta'$. Otherwise, set $\theta_n = \theta_{n-1}$.

4.4.2.1 Coupled exchange algorithm

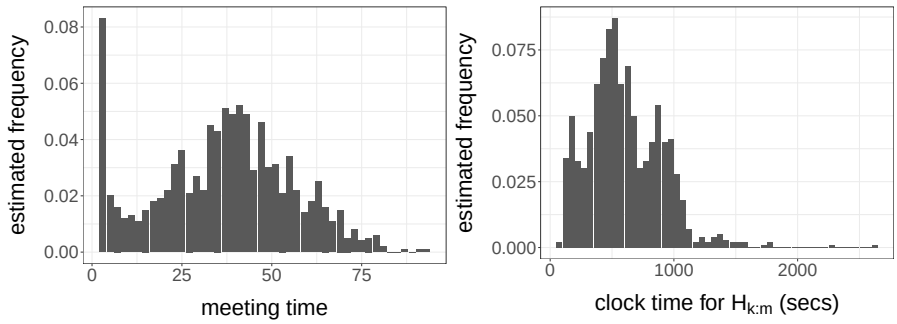
An algorithm to couple two block pseudo-marginal algorithms to construct unbiased estimators $H_{k:m}$ is provided in Algorithm 13. Denoting the two states of the chains at step $n \geq 1$ by θ_n and $\tilde{\theta}_{n-1}$, Algorithm 13 describes how to obtain θ_{n+1} and $\tilde{\theta}_n$; thus it describes a kernel $\bar{P}$.

Algorithm 13 Sampling from the coupled Exchange kernel given $(\theta_n, \tilde{\theta}_{n-1})$

1. Sample θ' and $\tilde{\theta}'$ from the maximal coupling of $q\left(\cdot|\theta_n\right)$ and $q(\cdot|\tilde{\theta}_{n-1})$.
2. If the proposals couple, i.e. if $\theta' = \tilde{\theta}'$, then sample $Y' \sim p(\cdot|\theta')$ and set $\tilde{Y}' = Y'$.
3. If the proposals do not couple, sample $Y' \sim p(\cdot|\theta')$ and $\tilde{Y}' \sim p(\cdot|\tilde{\theta}')$.
4. Sample $\mathfrak{u} \sim \mathcal{U}\left[0,1\right]$.
5. If $\mathfrak{u} < \alpha_{\mathrm{EX}}\left(\theta_n, \theta', Y'\right)$ then set $\theta_{n+1} = \theta'$. Otherwise, set $\theta_{n+1} = \theta_n$.
6. If $\mathfrak{u} < \alpha_{\mathrm{EX}}(\tilde{\theta}_{n-1}, \tilde{\theta}', \tilde{Y}')$ then set $\tilde{\theta}_n = \tilde{\theta}'$. Otherwise, set $\tilde{\theta}_n = \tilde{\theta}_{n-1}$.

4.4.2.2 High temperature Ising model

We examine the proposed algorithm for inference in a planar lattice Ising model without an external field. The model comprises observations $y_i \in \{-1, +1\}$ on a $L \times L$ square lattice such that $p(y|\theta) \propto \exp\left(\beta \sum_{i\sim j} y_i y_j\right)$ where $i \sim j$ denotes the neighbours j of node i and $\theta = \beta$ denotes the inverse temperature. We restrict interest to high temperature models specifying a prior distribution $\beta \sim \mathcal{U}[0, \beta_c]$, with $\beta_c = \frac{1}{2}\log(1+\sqrt{2})$ denoting the critical temperature of the Ising model on the infinite lattice (Ullrich 2013, Onsager 1944). Here, perfect simulation can be performed using coupling from the past techniques with simple heat bath dynamics developed by Propp and Wilson (1996). We generate observations for $L = 80$ and set the proposal covariance to $10^{-4}I$, initialising the chains from the prior.



(a) **(b)**

Figure 4.11: Coupled exchange algorithm for unbiased Bayesian inference with an 80×80 Ising model. Left: distribution of meeting times (1,000 runs). Right: clock time to obtain 1,000 unbiased estimators.

We obtain estimates of the distribution of meeting times using 1,000 repetitions of coupled exchange, with the results shown in Figure 4.11a. Based on this, we obtain unbiased estimates of the expectation of β under the posterior distribution using $k = 100$ and $m = 10k$ over 1,000 repetitions. It is noted that the clock time to obtain a single estimator (on the same machine) varies significantly due to the variable computational cost of performing coupling from the past, depending on θ. We plot a histogram of the clock times to obtain each $H_{k:m}$ in Figure 4.11b.

Based on the heterogeneity of times to produce a single unbiased estimator, we compare the serial inefficiency with the inefficiency of coupled exchange based on the clock time to obtain a certain variance, with the test function $h : x \mapsto x$. We estimate the asymptotic variance with $n_{\text{mcmc}} = 200,000$ iterations of the original algorithm (discarding the first 10% as burn-in, and using `spectrum0.ar` as before) to be $V_{as} \approx 4.22 \cdot 10^{-4}$, and the algorithm taking in total $41,095$ seconds. As a result, we estimate the serial inefficiency in terms of clock-time to be $41,095 \times V_{\text{as}}/(n_{\text{mcmc}} - n_{\text{burnin}}) \approx 9.6 \cdot 10^{-5}$. Comparatively, the mean time to return a single estimator $H_{k:m}$ was estimated to be 546 seconds, with the variance of a single $H_{k:m}$ estimated to be $4.78 \cdot 10^{-7}$ providing an estimated inefficiency of $2.6 \cdot 10^{-4}$, implying a three-fold increase in inefficiency.

4.5 Conclusion

Markov chain Monte Carlo algorithms designed for scenarios where the target density function is intractable can be coupled and utilized in the framework of Glynn and Rhee (2014), Jacob et al. (2017). The validity of the resulting unbiased estimators can be related to polynomial drift conditions on the underlying Markov kernels. These estimators open new ways of using parallel computing hardware to perform numerical integration in such scenarios.

In the context of state space models, in addition to parameter estimation, the proposed coupling strategy for PMMH would additionally provide unbiased estimators with respect to the joint distribution over state and parameters. This would enable unbiased smoothing under parameter uncertainty, instead of fixing the parameters as in Jacob et al. (2018), Lee et al. (2018) and Middleton et al. (2019).

Acknowledgement

The authors are grateful to Jeremy Heng for very helpful discussions. The data of Section 4.3.2 was kindly shared by Demba Ba. The experiments of that section were performed on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Pierre E. Jacob acknowledges support from the National Science Foundation through grant DMS-1712872.

4.6 Appendix

In the rest of the paper we will often use the symbol c to denote a generic positive constant whose value may vary from line to line.

4.6.1 Proof of Theorem 15

The following provides a slight relaxation on Assumption 2.2 in Jacob et al. (2017), where geometric conditions were imposed on the tails of the distribution of the meeting time. The following proof considers $H_0(Z, \tilde{Z})$ instead of $H_{k:m}$; one can first perform the same reasoning for $H_k(Z, \tilde{Z})$ for all $k \geq 0$, and then consider the finite average $(m-k+1)^{-1} \sum_{\ell=k}^{m} H_\ell(Z, \tilde{Z})$ to obtain the result for $H_{k:m}$.

By Assumption 13, it follows that $\mathbb{E}[\tau] < \infty$. This implies that the estimator H_0 can be computed in expected finite time. To show that $H_0(Z, \tilde{Z})$ admits a finite variance, we proceed by following Jacob et al. (2017; Proposition 3.1), adapting the proof under the proposed weaker assumptions. We denote the complete space of random variables with finite second moment by L_2. We then construct a Cauchy sequence of random variables $H^n(Z, \tilde{Z})$ in L_2 converging to $H_0(Z, \tilde{Z})$, where $H^n(Z, \tilde{Z}) := \sum_{t=0}^{n} \Delta_t$ with $\Delta_t = h(Z_t) - h(\tilde{Z}_{t-1})$ if $t > 0$ and $\Delta_t = h(Z_t)$ for $t = 0$. As $\mathbb{E}[\tau] < \infty$, we have $\mathbb{P}[\tau < \infty] = 1$ and $Z_t = \tilde{Z}_{t-1}$ for $t > \tau$. This implies that $H^n(Z, \tilde{Z}) \to H_0(Z, \tilde{Z})$ almost surely. For positive integers n, n' we have

$$\begin{aligned}
\mathbb{E}\left[\left(H^n(Z, \tilde{Z}) - H^{n'}(Z, \tilde{Z})\right)^2\right] &= \sum_{s=n+1}^{n'} \sum_{t=n+1}^{n'} \mathbb{E}[\Delta_s \Delta_t] \\
&\leq \sum_{s=n+1}^{n'} \sum_{t=n+1}^{n'} \mathbb{E}[\Delta_s^2]^{1/2} \mathbb{E}[\Delta_t^2]^{1/2} \\
&= \left(\sum_{t=n+1}^{n'} \mathbb{E}[\Delta_t^2]^{1/2} \right)^2 .
\end{aligned}$$

We note that $\mathbb{E}[\Delta_t^2] = \mathbb{E}[\Delta_t^2 1_{\tau > t}]$. Thus by Minkowski's inequality we obtain

$$\begin{aligned}
\mathbb{E}[\Delta_t^2] &\leq \mathbb{E}\left[|\Delta_t|^{2+\eta}\right]^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{E}[1_{\tau > t}]^{\frac{\eta}{2+\eta}} \\
&\leq c^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{P}[\tau > t]^{\frac{\eta}{2+\eta}},
\end{aligned}$$

where $\mathbb{E}[|\Delta_t|^{2+\eta}] < c$ for all t as $\mathbb{E}[h(Z_t)^{2+\eta}] < c$ by Assumption 12. Consequently we have

$$\mathbb{E}\left[\left(H^n(Z,\tilde{Z}) - H^{n'}(Z,\tilde{Z})\right)^2\right] \leq \left(\sum_{t=n+1}^{n'} \left(c^{\frac{1}{1+\frac{\eta}{2}}} \mathbb{P}[\tau > t]^{\frac{\eta}{2+\eta}}\right)^{\frac{1}{2}}\right)^2$$
$$= c^{\frac{1}{1+\frac{\eta}{2}}} \left(\sum_{t=n+1}^{n'} \mathbb{P}[\tau > t]^{\frac{1}{2}\frac{\eta}{2+\eta}}\right)^2.$$

With $\gamma = \frac{1}{2}\frac{\eta}{2+\eta}$, it follows from Assumption 13 that $\mathbb{P}[\tau > t] \leq Kt^{-\kappa}$ for $\kappa > 1/\gamma$ which yields

$$\sum_{t=n+1}^{\infty} \mathbb{P}[\tau > t]^{\gamma} \leq K \sum_{t=n+1}^{\infty} \frac{1}{t^{\gamma\kappa}} \leq K \int_n^{\infty} \frac{dt}{t^{\gamma\kappa}} < \infty.$$

We obtain $\lim_{n\to\infty} \sum_{t=n+1}^{\infty} \mathbb{P}[\tau > t]^{\gamma} = 0$. This proves that $H_n(Z,\tilde{Z})$ is a Cauchy sequence in L_2. We can thus conclude that the variance of $H_0(Z,\tilde{Z})$ is finite and that its expectation is $\lim_{n\to\infty} \mathbb{E}[H^n(Z,\tilde{Z})] = \lim_{n\to\infty} \mathbb{E}[h(Z_n)] = \pi(h)$.

4.6.2 Proof of Theorem 18

The following establishes a bivariate drift condition that we will later use to bound moments of the hitting time to the diagonal set $\mathcal{D}$. A similar statement is provided in Andrieu et al. (2015; Lemma 1).

Lemma 23. *Let $\bar{P}$ be a coupling of the Markov kernel P with itself, and V be as in Assumption 17. Then the function $\bar{V}(z,\tilde{z}) := V(z) + V(\tilde{z}) - 1$ satisfies*

$$\bar{P}\bar{V}(z,\tilde{z}) \leq \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) + \bar{b}\mathbb{1}_{\bar{C}}(z,\tilde{z}), \tag{4.18}$$

for all $(z,\tilde{z}) \in \mathcal{Z} \times \mathcal{Z}$, where $\bar{b} := 2b_V + \epsilon_b\phi(1)$ and $\bar{C} = C \times C$.

Proof. For $(z,\tilde{z}) \notin \bar{C}$ we have

$$\begin{aligned}
\bar{P}\bar{V}(z,\tilde{z}) &= PV(z) + PV(\tilde{z}) - 1 \\
&\leq V(z) + V(\tilde{z}) - 1 - \phi \circ V(z) - \phi \circ V(\tilde{z}) + b_V\left(\mathbb{1}_C(z) + \mathbb{1}_C(\tilde{z})\right) \\
&\leq V(z) + V(\tilde{z}) - 1 - \phi \circ V(z) - \phi \circ V(\tilde{z}) + b_V \\
&= V(z) + V(\tilde{z}) - 1 - \epsilon_b\left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right] \\
&\quad - (1-\epsilon_b)\left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right] + b_V.
\end{aligned}$$

Since $(z, \tilde{z}) \notin \bar{C}$ then at least one of $z, \tilde{z}$ is not in C, and $\phi \circ V \geq 0$, so

$$\begin{aligned}
&\leq V(z) + V(\tilde{z}) - 1 - \epsilon_b \left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right] - (1 - \epsilon_b) \inf_{z \notin C} \phi \circ V(z) + b_V \\
&\leq V(z) + V(\tilde{z}) - 1 - \epsilon_b \left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right] - b_V + b_V \\
&= \bar{V}(z, \tilde{z}) - \epsilon_b \left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right],
\end{aligned}$$

where we used (4.7) in Assumption 17. By two applications of the mean value theorem, we have that for any $t \geq s \geq 1$ there exist $r \in [t, t + s - 1]$ and $r^* \in [1, s]$ such that

$$\phi(t + s - 1) - \phi(t) = \phi'(r)\,(s - 1)\,, \qquad \phi(s) - \phi(1) = \phi'(r^*)(s - 1).$$

By concavity, since $t \geq s$ implies that $r \geq r^*$, it follows that $\phi'(r) \leq \phi'(r^*)$ and thus

$$\phi(t + s - 1) - \phi(t) \leq \phi(s) - \phi(1),$$

or equivalently

$$\phi(t + s - 1) + \phi(1) \leq \phi(t) + \phi(s).$$

Therefore, with $t = \max\{V(z), V(\tilde{z})\}$ and $s = \min\{V(z), V(\tilde{z})\}$ we get

$$\phi \circ \bar{V}(z, \tilde{z}) + \phi(1) \leq \phi \circ V(z) + \phi \circ V(\tilde{z}), \tag{4.19}$$

whence

$$\begin{aligned}
\bar{P}\bar{V}(z, \tilde{z}) &\leq \bar{V}(z, \tilde{z}) - \epsilon_b \left[\phi \circ V(z) + \phi \circ V(\tilde{z})\right] \\
&\leq \bar{V}(z, \tilde{z}) - \epsilon_b \left[\phi\left(\bar{V}(z, \tilde{z})\right) + \phi(1)\right] \\
&\leq \bar{V}(z, \tilde{z}) - \epsilon_b \phi \circ \bar{V}(z, \tilde{z}).
\end{aligned} \tag{4.20}$$

For $(z, \tilde{z}) \in \bar{C}$ we get by Assumption 17,

$$\begin{aligned}
\bar{P}\bar{V}(z, \tilde{z}) &= PV(z) + PV(\tilde{z}) - 1 \\
&\leq V(z) - \phi \circ V(z) + b_V + V(\tilde{z}) - \phi \circ V(\tilde{z}) + b_V - 1 \\
&= \bar{V}(z, \tilde{z}) - \phi \circ V(z) - \phi \circ V(\tilde{z}) + 2b_V.
\end{aligned} \tag{4.21}$$

Combining (4.20) and (4.21), (4.19) and the fact that $\phi \geq 0$, we have for any $(z, \tilde{z})$

$$\begin{aligned}
\bar{P}\bar{V}(z,\tilde{z}) &\leq \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) \mathbb{1}_{\bar{C}^{\mathsf{c}}}(z,\tilde{z}) \\
&\quad - \left[\phi \circ V(z) + \phi \circ V(\tilde{z}) - 2b_V\right] \mathbb{1}_{\bar{C}}(z,\tilde{z}) \\
&\leq \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) \mathbb{1}_{\bar{C}^{\mathsf{c}}}(z,\tilde{z}) \\
&\quad - \left[\phi \circ \bar{V}(z,\tilde{z}) + \phi(1) - 2b_V\right] \mathbb{1}_{\bar{C}}(z,\tilde{z}) \\
&= \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) \mathbb{1}_{\bar{C}^{\mathsf{c}}}(z,\tilde{z}) \\
&\quad - \left[\epsilon_b \phi \circ \bar{V}(z,\tilde{z}) + (1-\epsilon_b)\phi \circ \bar{V}(z,\tilde{z}) + \phi(1) - 2b_V\right] \mathbb{1}_{\bar{C}}(z,\tilde{z}) \\
&\leq \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) + \left[2b_V - \phi(1)\right] \mathbb{1}_{\bar{C}}(z,\tilde{z}) \\
&\leq \bar{V}(z,\tilde{z}) - \epsilon_b \phi \circ \bar{V}(z,\tilde{z}) + \left[2b_V + \phi(1)\right] \mathbb{1}_{\bar{C}}(z,\tilde{z}).
\end{aligned}$$

□

The proof of Theorem 18 then follows through making use of Douc et al. (2004; Proposition 2.1), which we provide below for the reader's convenience, noting that the exact statement is taken from Andrieu et al. (2015; Proposition 4). We borrow the following definitions from Andrieu et al. (2015). For any non-decreasing concave function $\psi : [1, \infty) \to (0, \infty)$, let

$$H_\psi(v) := \int_1^v \frac{dx}{\psi(x)}, \tag{4.22}$$

Let $H_\psi^{-1} : [0, \infty) \to [1, \infty)$ be its inverse. For $k \in \mathbb{N}$, $n \geq 0$, $v \geq 1$, let

$$\begin{aligned}
r_\psi(n) &:= \frac{\psi \circ H_\psi^{-1}(n)}{\psi(1)} \\
H_k(v) &:= H_\psi^{-1}\left(H_\psi(v) + k\right) - H_\psi^{-1}(k).
\end{aligned} \tag{4.23}$$

Proposition 24. *(Proposition 2.1 from Douc et al. (2004)). Assume that P is a Markov kernel such that for some function $V \geq 1$ we have*

$$PV(z) \leq V(z) - \psi \circ V(z) + b\mathbb{1}_C(z),$$

where $\psi : [1, \infty) \mapsto (0, \infty)$ is a nondecreasing concave function. Let r_ψ and H_ψ be defined as in (4.23). Then we have for $V_k := H_k \circ V$

$$PV_{k+1}(z) \leq V_k(z) - \psi(1) r_\psi(k) + b r_\psi(k+1) \mathbb{1}_C(z), \qquad k \geq 0.$$

Equipped with the above results we proceed to the proof of Theorem 18. Applying Proposition 24 with $\bar{P}$, $\bar{V}$, $\psi = \epsilon_b \phi$ and and $b = \bar{b}$, then letting

$$r(n) := \frac{\phi \circ H_\phi^{-1}(\epsilon_b n)}{\phi(1)}$$

we have the sequence of drift conditions

$$\bar{P}\bar{V}_{k+1}(z,\tilde{z}) \leq \bar{V}_k(z,\tilde{z}) - \epsilon_b \phi(1) r(k) + \bar{b} r(k+1) \mathbb{1}_{\bar{C}}(z,\tilde{z}), \quad k \geq 0,$$

where $\bar{V}_k := H_k \circ \bar{V}$. Letting $\tilde{V}_k = \bar{V}_k + 1 \geq 1$ we obtain

$$\bar{P}\tilde{V}_{k+1}(z,\tilde{z}) \leq \tilde{V}_k(z,\tilde{z}) - \epsilon_b \phi(1) r(k) + \bar{b} r(k+1) \mathbb{1}_{\bar{C}}(z,\tilde{z}), \quad k \geq 0. \tag{4.24}$$

To proceed we follow the proof of Douc et al. (2004; Proposition 2.5), specifically the steps leading up to Douc et al. (2004; Equation (2.6)). Notice that by Assumption 16 the diagonal $\mathcal{D}$ is an accessible set, since clearly $\pi_{\mathcal{D}}(\mathcal{D}) = 1 > 0$. Therefore by Dynkin's formula we have

$$\epsilon_b \phi(1) \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k) \right] \leq \tilde{V}_0(z,\tilde{z}) + \bar{b} \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k+1) \mathbb{1}_{\bar{C}}(\Xi_k) \right],$$

where in the above, $\mathbb{E}_{z,\tilde{z}}$ denotes expectation with respect to the probability measure under which the joint chain $\Xi_n := \left(Z_n, \tilde{Z}_{n-1}\right)$ is initialized at (z,z') and evolves according to the transition kernel $\bar{P}$, $\tau_{\mathcal{D}} := \inf\{n \geq 1 : \Xi_n \in \mathcal{D}\}$, c_1, c_2 are positive constants depending on the set B and the various constants in the drift condition, but not on $(z,\tilde{z})$. Notice that by Assumption 16 we have that for all $(z,\tilde{z}) \in \bar{C}$, and $\rho \in (0,1)$

$$K_\rho\left((z,\tilde{z}),\mathcal{D}\right) := \sum_{i=0}^{\infty} \rho^i \bar{P}^i\left((z,\tilde{z}),\mathcal{D}\right) \geq \rho^{n_0}\epsilon.$$

In particular it easily follows that

$$\mathbb{1}_{\bar{C}}\left((z,\tilde{z})\right) \leq (\rho^{n_0}\epsilon)^{-1} K_\rho\left((z,\tilde{z}),\mathcal{D}\right),$$

and therefore continuing from above

$$\begin{aligned}
\epsilon_b \phi(1) \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k) \right] &\leq \tilde{V}_0(z,\tilde{z}) + \frac{\bar{b}}{\rho^{n_0}\epsilon} \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k+1) K_\rho(\Xi_k,\mathcal{D}) \right] \\
&= \tilde{V}_0(z,\tilde{z}) + \frac{\bar{b}}{\rho^{n_0}\epsilon} \sum_{i=0}^{\infty} \rho^i \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k+1) \bar{P}^i(\Xi_k,\mathcal{D}) \right] \\
&= \tilde{V}_0(z,\tilde{z}) + \frac{\bar{b}}{\rho^{n_0}\epsilon} \sum_{i=0}^{\infty} \rho^i \sum_{k=0}^{\infty} \mathbb{E}_{z,\tilde{z}} \Big[\mathbb{1}\{k \leq \tau_{\mathcal{D}} - 1\} r(k+1) \mathbb{1}_{\mathcal{D}}(\Xi_{k+i}) \Big].
\end{aligned}$$

A careful look above reveals that the integrand will be non-zero only for k such that $\tau_{\mathcal{D}} \leq k+i$ and $k \leq \tau_{\mathcal{D}} - 1$. There are at most i such values of k, and since $r(\cdot)$ is non-decreasing for each one of these values we will have $r(k+1) \leq r(\tau_{\mathcal{D}})$. Therefore

$$\epsilon_b \phi(1) \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k) \right] \leq \tilde{V}_0(z,\tilde{z}) + \frac{\bar{b}}{\rho^{n_0}\epsilon} \sum_{i=0}^{\infty} \rho^i i \times \mathbb{E}_{z,\tilde{z}} \Big[r(\tau_{\mathcal{D}}) \Big].$$

Similarly to the proof of Douc et al. (2004; Proposition 2.5), using the fact that $r(\cdot)$ grows sub-geometrically we can find for any $\delta > 0$ a constant $c(\delta) > 0$ such that

$$r(k) \leq \delta \sum_{j=0}^{k-1} r(j) + c(\delta),$$

and therefore conclude that for some constants c_1, c_2, independent of $(z, \tilde{z})$, we have

$$\mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k) \right] \leq \frac{\tilde{V}_0(z,\tilde{z}) + c_1}{c_2}.$$

From the definition of $\phi(y)$ we have that

$$r(n) = [d(1-\alpha)\epsilon_b n + 1]^{\alpha/(1-\alpha)} \geq c n^{\alpha/(1-\alpha)},$$

where recall that c denotes a generic constant whose value may change from line to line. Thus for any N

$$\sum_{k=0}^{N} r(k) \geq c \sum_{k=0}^{N} k^{\alpha/(1-\alpha)} \geq c \int_{x=0}^{N} x^{\alpha/(1-\alpha)} dx = c N^{1/(1-\alpha)},$$

hence we obtain

$$\mathbb{E}_{z,\tilde{z}} \left[\tau_{\mathcal{D}}^{1/(1-\alpha)} \right] \leq c \mathbb{E}_{z,\tilde{z}} \left[\sum_{k=0}^{\tau_{\mathcal{D}}-1} r(k) \right] \leq c \frac{\tilde{V}_0(z,\tilde{z}) + c_1}{c_2}.$$

We have that the chain $\left(Z_n, \tilde{Z}_{n-1}\right)$ is initialised at $n = 1$ under $\pi_0 P \otimes \pi_0$. Recalling the definition of $\tilde{V}_0$ we have that $\tilde{V}_0(z,\tilde{z}) \leq V(z) + V(\tilde{z})$ and as π_0 is compactly supported, $\pi_0(V) < \infty$. Similarly by Assumption 17 we have that $\pi_0 P(V) < \infty$, in which case it follows that $\mathbb{E}_{\pi_0 P \otimes \pi_0} \left[\tau_{\mathcal{D}}^{1/(1-\alpha)} \right] < \infty$. An application of Markov's inequality completes the proof

$$\mathbb{P}_{\pi_0 P \otimes \pi_0} \left[\tau_{\mathcal{D}} \geq t \right] \leq \frac{\mathbb{E}_{\pi_0 P \otimes \pi_0} [\tau_{\mathcal{D}}^{1/(1-\alpha)}]}{t^{1/(1-\alpha)}} \leq \frac{c}{t^{1/(1-\alpha)}}.$$

4.6.3 Proof of Proposition 19

To fix notation, we have that for any measurable functions $W : \mathcal{Z} \to [1, \infty)$, $g : \mathcal{Z} \to \mathbb{R}$, and a finite signed measure μ on $\mathcal{X}$, we write

$$|g|_W := \sup_{z \in \mathcal{Z}} \frac{|g(z)|}{W(z)}, \qquad \|\mu\|_W := \sup_{f : \|f\|_W \leq 1} |\mu(f)|.$$

Our starting point is Assumption 17 which we restate here

$$PV(z) \leq V(z) - dV^{\alpha}(z) + b_V \mathbb{1}_C(z), \tag{4.25}$$

for some function $V : \mathcal{Z} \to [1, \infty)$, some $\alpha \in (0, 1)$, constants $b_V, d > 0$ and a small set C. As before we assume that $(Z_n, \tilde{Z}_{n-1})$ evolves according to $\bar{P}$, and that marginally the components Z_n and $\tilde{Z}_n$ evolve according to P. Notice that we write $\mathbb{E}$ for the measure with the chains started from π_0 and $\mathbb{E}_\pi$ for the measure with the chains initialized at π.

By Jarner and Roberts (2002; Lemma 3.5) for any $\eta \in (0, 1)$ there exist $b', d' > 0$ such that

$$PV^{\gamma}(z) \leq V^{\gamma}(z) - d'V^{\alpha+\gamma-1}(z) + b'\mathbb{1}_C(z). \tag{4.26}$$

With $\gamma \in (1 - \alpha, 1)$ as in the statement of Proposition 19, we have that $\alpha + \gamma - 1 \in (0, 1)$. Under this assumption, from (4.26), Meyn and Tweedie (2009; Theorem 14.0.1) applied with $f = V^{\alpha+\gamma-1}$ and the fact that π is a maximal irreducibility measure (see Meyn and Tweedie (2009; Proposition 10.1.2)), it follows that $\pi(S_V) = 1$, with S_V as defined in the statement of Proposition 19. From this we conclude that V is π-a.e. finite. Also from Meyn and Tweedie (2009; Theorem 14.0.1), since $\pi(V^{\gamma}) \leq \pi(V^{4\gamma})^{1/4} < \infty$ by assumption, we have that for all π-a.e. $z \in \mathcal{Z}$ there exists a finite constant c such that

$$\sum_{n=0}^{\infty} \|P^n(z, \cdot) - \pi\|_{V^{\alpha+\gamma-1}} \leq c(1 + V^{\gamma}(z)). \tag{4.27}$$

Since by assumption $|h|_{V^{\alpha+\gamma-1}} < \infty$, we have

$$\begin{aligned} \sum_{n=0}^{\infty} |P^n[h - \pi(h)](z)| &\leq \|h\|_{V^{\alpha+\gamma-1}} \sum_{n=0}^{\infty} \|P^n(z, \cdot) - \pi\|_{V^{\alpha+\gamma-1}} \\ &\leq c\|h\|_{V^{\alpha+\gamma-1}}(1 + V^{\gamma}(z)) < \infty, \end{aligned}$$

for π-almost all z. Therefore the function

$$g(z) := \sum_{j=0}^{\infty} P^j \left[h - \pi(h)\right](z)$$

is well-defined and satisfies $|g|_{V^\gamma} < \infty$, $\pi(g^2) < \infty$, where the second property follows from $\pi(V^{4\gamma}) < \infty$. In particular it follows that $g - Pg = h - \pi(h)$, and therefore g is the solution to the Poisson equation with respect to P and h. We continue with the calculation in the proof of Jacob et al. (2017; Proposition 3.3). Let

$$S_j^{(N)} := \mathbb{1}\{\tau_{\mathcal{D}} > j\} \sum_{t=j}^{N \wedge \tau_{\mathcal{D}} - 1} b_t \left[h(Z_t) - h(\tilde{Z}_{t-1})\right],$$

where $(b_t)_{t \geq 0}$ is an arbitrary bounded sequence. Writing $\mathbf{Z}_t := (Z_t, \tilde{Z}_{t-1})$, $\bar{g}(x, y) = g(x) - g(y)$ and $\bar{P}$ for the transition kernel of $\mathbf{Z}_t$ we then have

$$\begin{aligned}
h(Z_t) - h(\tilde{Z}_{t-1}) &= \left[h(Z_t) - \pi(h)\right] - \left[h(\tilde{Z}_{t-1}) - \pi(h)\right] \\
&= \left[g(Z_t) - Pg(Z_t)\right] - \left[g(\tilde{Z}_{t-1}) - Pg(\tilde{Z}_{t-1})\right] \\
&= \left[g(Z_t) - g(\tilde{Z}_{t-1})\right] - \left[Pg(Z_t) - Pg(\tilde{Z}_{t-1})\right] \\
&= \bar{g}(\mathbf{Z}_t) - \bar{P}\bar{g}(\mathbf{Z}_t),
\end{aligned}$$

where we used the fact that, by construction of $\bar{P}$, we have $\bar{P}\bar{g}(z, \tilde{z}) = Pg(z) - Pg(\tilde{z})$.

Then from Jacob et al. (2017; Equation (A.3)) we have

$$\begin{aligned}
\mathbb{E}\left\{\left[S_j^{(N)}\right]^2\right\} \leq 4 \sum_{t=j}^{N-1} b_t^2 \mathbb{E}\left\{\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]^2 \mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
+ 4 b_j^2 \mathbb{E}\left[\bar{g}^2(\mathbf{Z}_j)\mathbb{1}\{\tau_{\mathcal{D}} > j\}\right] + 4 b_N^2 \mathbb{E}\left[\bar{g}^2(\mathbf{Z}_N)\mathbb{1}\{\tau_{\mathcal{D}} > N\}\right] \\
+ 4 \left\{\sum_{t=j}^{N-1} |b_{t+1} - b_t| \mathbb{E}^{1/2}\left[\bar{g}^2(\mathbf{Z}_{t+1})\mathbb{1}\{\tau_{\mathcal{D}} > t+1\}\right]\right\}^2,
\end{aligned}$$

and we proceed to bound these terms. Letting $\mathcal{F}_t := \sigma\left(\boldsymbol{Z_s}; 0 \leq s \leq t\right)$, notice that

$$
\begin{aligned}
&\mathbb{E}\left\{\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]^2 \mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&\quad - \mathbb{E}\left\{\bar{P}\bar{g}(\mathbf{Z}_t)\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&\quad - \mathbb{E}\left\{\mathbb{E}\left[\bar{P}\bar{g}(\mathbf{Z}_t)\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\middle|\mathcal{F}_t\right]\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&\quad - \mathbb{E}\left\{\bar{P}\bar{g}(\mathbf{Z}_t)\mathbb{1}\{\tau_{\mathcal{D}} > t\}\mathbb{E}\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\middle|\mathcal{F}_t\right]\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} - \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})\bar{P}\bar{g}(\mathbf{Z}_t)\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} - \mathbb{E}\left\{\mathbb{E}\left[\bar{g}(\mathbf{Z}_{t+1})\middle|\mathcal{F}_t\right]\bar{P}\bar{g}(\mathbf{Z}_t)\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&= \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} - \mathbb{E}\left\{\bar{P}\bar{g}(\mathbf{Z}_t)^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&\leq \mathbb{E}\left\{\bar{g}(\mathbf{Z}_{t+1})^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \leq |g|^2_{V^\gamma}\mathbb{E}\left\{\left[V^\gamma(Z_t) + V^\gamma(\tilde{Z}_{t-1})\right]^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\}.
\end{aligned}
$$

We next bound the last quantity using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ and the Cauchy-Schwarz inequality

$$
\begin{aligned}
&\mathbb{E}\left\{\left[V^\gamma(Z_t) + V^\gamma(\tilde{Z}_{t-1})\right]^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \leq 2\mathbb{E}\left\{\left[V^{2\gamma}(Z_t) + V^{2\gamma}(\tilde{Z}_{t-1})\right]\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \\
&\quad \leq c\left[\mathbb{E}\left\{V^{4\gamma}(Z_t)\right\} + \mathbb{E}\left\{V^{4\gamma}(\tilde{Z}_{t-1})\right\}\right]^{1/2}\mathbb{P}\left(\tau_{\mathcal{D}} > t\right)^{1/2}.
\end{aligned}
$$

Finally notice that since V is non-negative

$$
\begin{aligned}
\mathbb{E}\left\{V^{4\gamma}(Z_t)\right\} &\leq \left\|\frac{d\pi_0}{d\pi}\right\|_\infty \mathbb{E}_\pi\left\{V^{4\gamma}(Z_t)\right\} \\
&\leq \left\|\frac{d\pi_0}{d\pi}\right\|_\infty \mathbb{E}_\pi\left\{V^{4\gamma}(Z_0)\right\} = c\pi(V^{4\gamma}) < \infty,
\end{aligned}
$$

where we used the fact that when started from π and evolved through $\bar{P}$, the Markov chain $\{Z_t\}_{t\geq 0}$ is stationary. From the above and Theorem 18 we conclude that there exists a positive constant $c < \infty$ such that

$$
\mathbb{E}\left\{\left[\bar{g}(\mathbf{Z}_{t+1}) - \bar{P}\bar{g}(\mathbf{Z}_t)\right]^2\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \leq \frac{c}{t^{\kappa/2}}.
$$

On the other hand for terms of the form $\mathbb{E}[\bar{g}^2(\mathbf{Z}_t)\mathbb{1}\{\tau_{\mathcal{D}} > t\}]$, using the same techniques we have

$$\mathbb{E}\left[\bar{g}^2(\mathbf{Z}_t)\mathbb{1}\{\tau_{\mathcal{D}} > t\}\right] \leq |g|^2_{V^\gamma}\mathbb{E}\left\{\left[V^\gamma(Z_t) + V^\gamma(\tilde{Z}_{t-1})\right]^2 \mathbb{1}\{\tau_{\mathcal{D}} > t\}\right\} \leq \frac{c}{t^{\kappa/2}}.$$

Overall we thus have that

$$\mathbb{E}\left\{\left[S_j^{(N)}\right]^2\right\} \leq c\left[\frac{b_j^2}{j^{\kappa/2}} + \frac{b_N^2}{N^{\kappa/2}} + \sum_{t=j}^{N-1}\frac{b_t^2}{t^{\kappa/2}} + \left(\sum_{t=j}^{N-1}\frac{|b_{t+1}-b_t|}{t^{\kappa/2}}\right)^2\right],$$

$$\mathbb{E}\left\{S_j^2\right\} \leq c\left[\frac{b_j^2}{j^{\kappa/2}} + \sum_{t\geq j}\frac{b_t^2}{t^{\kappa/2}} + \left(\sum_{t=j}^{\infty}\frac{|b_{t+1}-b_t|}{t^{\kappa/2}}\right)^2\right],$$

where $S_j := \lim_{N\to\infty} S_j^{(N)}$ is the limit in the L^2 sense as in Jacob et al. (2017; Proposition 3.1). Setting $b_j = 0$, $b_t := (t-j)/(m-j+1)$ for $j < t < m+1$ and $b_t := 1$ for $t > m+1$ we then obtain

$$\mathbb{E}\left[S_j^2\right] \leq c\left[\sum_{t=j+1}^{m}\frac{(t-j)^2}{(m-j+1)^2 t^{\kappa/2}} + \sum_{t=m+1}^{\infty}\frac{1}{t^{\kappa/2}} + \left(\sum_{t=j}^{m+1}\frac{1}{(m-j+1)t^{\kappa/2}}\right)^2\right].$$

For the first term notice that, after changing variables $r = t-j$ and writing $M = m-j+1$ we have

$$\begin{aligned}
\sum_{t=j+1}^{m+1}\frac{(t-j)^2}{(m-j+1)^2 t^{\kappa/2}} &= \frac{1}{M^2}\sum_{r=1}^{M}\frac{r^2}{(r+j)^{\kappa/2}} \\
&\leq \frac{c}{M^2}\int_{x=1}^{M}\frac{x^2}{(x+j)^{\kappa/2}}dx \\
&= \frac{c}{M^2 j^{\kappa/2}}\int_{x=1}^{M}\frac{x^2}{(x/j+1)^{\kappa/2}}dx \qquad \text{(changing } z = x/j\text{)} \\
&\leq \frac{c}{M^2 j^{\kappa/2}}\int_{z=1/j}^{M/j}\frac{j^3 z^2}{(z+1)^{\kappa/2}}dz \\
&= \frac{c}{M^2 j^{\kappa/2-3}}\int_{z=1/j}^{M/j}\frac{z^2}{(z+1)^{\kappa/2}}dz \leq \frac{c}{M^2 j^{\kappa/2-3}},
\end{aligned}$$

since by assumption $\kappa = 1/(1-\alpha) > 6$. Finally we get

$$\begin{aligned}
\mathbb{E}\left[S_j^2\right] &\leq c\left[\frac{1}{(m-j+1)^2 j^{\kappa/2-3}} + \frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-j+1)^2 j^{\kappa-2}}\right] \\
&= c\left[\frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-j+1)^2}\left(\frac{1}{j^{\kappa/2-3}} + \frac{1}{j^{\kappa-2}}\right)\right]. \\
&\leq c\left[\frac{1}{m^{\kappa/2-1}} + \frac{1}{(m-j+1)^2}\frac{1}{j^{\kappa/2-3}}\right]
\end{aligned}$$

as $\kappa/2-3 \leq \kappa-2$. With our choice of sequence $(b_t)_{t\geq 0}$, S_j coincides with $\mathrm{BC}_{j:m}$ in the notation of the statement of the proposition which thus follows from the above.

4.6.4 Proof of Proposition 22

First we want to prove the minorization condition (4.4) for the set $C = B(0,M) \times [\underline{w}, \overline{w}]$, where $M, \underline{w}, \overline{w} > 0$ are given and fixed. That is, we want to establish that there exist $\epsilon_0 > 0$ and a probability measure ν such that

$$P\left((\theta,w), d\theta', dw'\right) \geq \epsilon_0 \nu\left(d\theta', dw'\right)$$

for all $(\theta, w) \in C$. We have

$$\begin{aligned} P\left((\theta,w), d\theta', dw'\right) \geq & \overline{g}_{\theta'}(w') \min\left\{q(\theta,\theta'), \frac{\pi(\theta')}{\pi(\theta)} q(\theta',\theta)\right\} \min\left\{1, \frac{w'}{w}\right\} d\theta' dw' \\ \geq & \varepsilon_\pi \mathbb{I}\left(\theta' \in B(0,M)\right) \min\left\{q(\theta,\theta'), q(\theta',\theta)\right\} \\ & \min\left\{\overline{g}_{\theta'}(w'), \overline{g}_{\theta'}(w') \frac{w'}{\overline{w}}\right\} d\theta' dw', \end{aligned}$$

where

$$\varepsilon_\pi := \frac{\inf_{\theta:|\theta|\leq M} \pi(\theta')}{\sup_{\theta:|\theta|\leq M} \pi(\theta)} > 0,$$

by the assumption that π is bounded from above, and bounded away from zero on all compact sets. Since the proposal q is bounded away from zero on compact sets we also have that $\min\left\{q(\theta,\theta'), q(\theta',\theta)\right\} \geq \varepsilon_q$ for $|\theta'-\theta| < 2M$ which ensures that

$$P\left((\theta,w), d\theta', dw'\right) \geq \varepsilon_q \varepsilon_\pi \min\left\{\overline{g}_{\theta'}(w'), \overline{g}_{\theta'}(w') \frac{w'}{\overline{w}}\right\} d\theta' dw'.$$

This can be rewritten as

$$P\left((\theta,w), d\theta', dw'\right) \geq \varepsilon_q \varepsilon_\pi \mathbb{I}\left(\theta' \in B(0,M)\right) Z(\theta') \widetilde{g}_{\theta'}(w) d\theta' dw'$$

with

$$Z(\theta) := \int \overline{g}_\theta(w) \min\left\{1, \frac{w}{\overline{w}}\right\} dw \leq 1,$$

and

$$\widetilde{g}_\theta(w) = Z^{-1}(\theta) \overline{g}_\theta(w) \min\left\{1, \frac{w}{\overline{w}}\right\}.$$

Suppose now that for fixed $M, \overline{w}$ we have

$$\inf_{\theta:|\theta|\leq M} Z(\theta) = 0,$$

which implies that there is a sequence $\theta_n \in B(0,M)$ such that $\lim_{n\to\infty} Z(\theta_n) = 0$. Since $B(0,M)$ is compact we can extract a convergent subsequence $\theta_{n_k} \to \bar{\theta} \in B(0,M)$ such that $\lim_{k\to\infty} Z(\theta_{n_k}) = 0$. By weak convergence, since $w \mapsto \min\{1, w/\overline{w}\}$ is bounded and continuous, we also have that

$$0 = \lim_{k\to\infty} Z(\theta_{n_k}) = \lim_{k\to\infty} \int \overline{g}_{\theta_{n_k}}(w) \min\left\{1, \frac{w}{\overline{w}}\right\} dw = \int \overline{g}_{\bar{\theta}}(w) \min\left\{1, \frac{w}{\overline{w}}\right\} dw.$$

Since $w \mapsto \min\{1, w/\overline{w}\}$ is strictly positive for $w > 0$, this implies that the support of $\overline{g}_{\bar{\theta}}$ is $\{0\}$ which is a contradiction, since in that case necessarily $\int \overline{g}_{\bar{\theta}}(w)\, w\, dw = 0 \neq 1$. Therefore we conclude that for all finite $M, \overline{w} > 0$, there exists $\varepsilon_Z(M, \overline{w}) > 0$ such that $Z(\theta) > \varepsilon_Z(M, \overline{w})$, for all $\theta \in B(0,M)$.

Therefore we obtain

$$P((\theta, w), d\theta', dw') \geq \varepsilon_Z \varepsilon_q \varepsilon_\pi \mathbb{I}(\theta' \in B(0,M))\, \widetilde{g}_{\theta'}(w)\, d\theta' dw',$$

which proves the result for $\epsilon_0 = \varepsilon_Z \varepsilon_q \varepsilon_\pi \mathrm{vol}\{B(0,M)\}$ and with minorising measure $\nu(d\theta', dw') = \mathcal{U}(\theta' \in B(0,M))\, \widetilde{g}_{\theta'}(w)$.

Next we establish that the minorization condition (4.3) holds for $\bar{P}$, the coupled transition kernel defined by Algorithm 9, and C as defined above. Let the current states be $z := (\theta, w), \tilde{z} := \left(\tilde{\theta}, \tilde{w}\right) \in C$ respectively. According to Algorithm 9 the next parameter states $\theta', \tilde{\theta}'$ will be sampled from $\mathfrak{Q}\left(\left(\theta, \tilde{\theta}\right), \mathrm{d}\theta', \mathrm{d}\tilde{\theta}'\right)$, the γ-coupling of $q(\cdot|\theta)$ and $q\left(\cdot\middle|\tilde{\theta}\right)$. This is the maximal coupling generated by the rejection sampler described in Jacob et al. (2017). If the coupling is successful, that is $\theta' = \tilde{\theta}'$, then the algorithm samples $w' \sim \bar{g}_{\theta'}(\cdot)$, sets $\tilde{w}' = w'$ in which case we know by definition that $((\theta', w'), (\widetilde{\theta}', \widetilde{w}')) \in \mathcal{D}$ if the proposal (θ', w') is accepted since the same uniform is used in both acceptance steps. Therefore under the coupled transition kernel $\bar{P}$, writing $z := (\theta, w), \tilde{z} := \left(\tilde{\theta}, \tilde{w}\right)$ and letting

$\mathcal{D}_\theta := \left\{\left(\theta, \tilde{\theta}\right) : \theta = \tilde{\theta}\right\}$ be the diagonal of $\Theta \times \Theta$, we have for $(z, z') \in C \times C$ that

$$\begin{aligned}
\bar{P}\left((z,z'),\mathcal{D}\right) &\geq \bar{P}\left((z,z'),\mathcal{D}\cap\left(B(0,M)\times\mathbb{R}^+\right)^2\right) \\
&= \iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right)\int_{\mathbb{R}^+}\bar{g}_{\theta'}\left(w'\right) \\
&\qquad \int_{u=0}^1 \mathbb{I}\left[u \leq \min\left\{1, \frac{\pi\left(\theta'\right)}{\pi\left(\theta\right)}\frac{w'}{w}\right\}\right]\mathbb{I}\left[u \leq \min\left\{1, \frac{\pi\left(\theta'\right)}{\pi\left(\tilde{\theta}\right)}\frac{w'}{\tilde{w}}\right\}\right]\mathrm{d}u\mathrm{d}w' \\
&= \iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\
&\qquad \int_{\mathbb{R}^+}\bar{g}_{\theta'}\left(w'\right)\int_{u=0}^1 \mathbb{I}\left[u \leq \min\left\{1, \frac{\pi\left(\theta'\right)}{\pi\left(\theta\right)}\frac{w'}{w}, \frac{\pi\left(\theta'\right)}{\pi\left(\tilde{\theta}\right)}\frac{w'}{\tilde{w}}\right\}\right]\mathrm{d}u\mathrm{d}w' \\
&= \iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\
&\qquad \int_{\mathbb{R}^+}\bar{g}_{\theta'}\left(w'\right)\min\left\{1, \frac{\pi\left(\theta'\right)}{\pi\left(\theta\right)}\frac{w'}{w}, \frac{\pi\left(\theta'\right)}{\pi\left(\tilde{\theta}\right)}\frac{w'}{\tilde{w}}\right\}\mathrm{d}w',
\end{aligned}$$

where we also used the fact that the proposal is symmetric by assumption. Continuing from the above inequality, letting $\varepsilon_\pi, \varepsilon_q$ and ε_Z be as above, we have that

$$\begin{aligned}
\bar{P}\left((z,z'),\mathcal{D}\right) &\geq \iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\
&\qquad \int_{\mathbb{R}^+}\bar{g}_{\theta'}\left(w'\right)\min\left\{1, \varepsilon_\pi\frac{w'}{\overline{w}}, \varepsilon_\pi\frac{w'}{\overline{w}}\right\}\mathrm{d}w' \\
&\geq \iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\
&\qquad \int_{\mathbb{R}^+}\bar{g}_{\theta'}\left(w'\right)\min\left\{1,\varepsilon_\pi\right\}\min\left\{1, \frac{w'}{\overline{w}}\right\}\mathrm{d}w' \\
&= \min\left\{1,\varepsilon_\pi\right\}\iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right)Z\left(\theta'\right)\int_{\mathbb{R}^+}\widetilde{g}_{\theta'}\left(w'\right)\mathrm{d}w' \\
&\geq \varepsilon_Z\varepsilon_\pi\iint_{\mathcal{D}_\theta\cap B(0,M)^2} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\
&\geq \varepsilon_Z\varepsilon_\pi\int_{B(0,M)}\min\left\{q\left(\theta'\left|\theta\right.\right), q\left(\theta'\left|\tilde{\theta}\right.\right)\right\}\mathrm{d}\theta' \\
&\geq \varepsilon_Z\varepsilon_\pi\int_{B(0,M)}\varepsilon_q\mathrm{d}\theta' \\
&= \varepsilon_Z\varepsilon_\pi\varepsilon_q\mathrm{vol}\left(B\left(0,M\right)\right) > 0,
\end{aligned}$$

where we used the fact that in the $\gamma-$coupling, conditionally on the coupling succeeding, the variables are sampled from a density proportional to the minimum of their respective densities. This establishes that condition (4.3) holds with $C = B\left(0,M\right)\times\left[\underline{w},\overline{w}\right]$ for any $M, \underline{w}, \overline{w}$.

Next we establish that $\bar{P}$ is $\pi_{\mathcal{D}}-$irreducible. Let $A \subset \mathcal{D}$ such that $\pi_{\mathcal{D}}(A) > 0$. For sets $A \subset \mathcal{D}$ we will write $A^{(1)}$ for the projection onto its first coordinate, that is if $A \subset \mathcal{D}$ then $A = A^{(1)} \times A^{(1)}$. We need to show that for any $z, \tilde{z} \in \mathcal{Z}$ there exists $n \geq 1$ such that $\bar{P}^n((z,\tilde{z}),A) > 0$. Notice that by construction if $(z,\tilde{z}) \in \mathcal{D}$ then $\bar{P}((z,\tilde{z}),\mathrm{d}z',\mathrm{d}\tilde{z}') = P(z,\mathrm{d}z')\,\delta_{z'}(\mathrm{d}\tilde{z}')$, that is the chain couples automatically from the diagonal and proceeds as the pseudo-marginal kernel P. Letting $z, \tilde{z} \in \mathcal{Z}$ and $n \geq 1$ we have

$$\begin{aligned}\bar{P}^{n+1}((z,\tilde{z}),A) &\geq \iint_{\mathcal{D}} \bar{P}((z,\tilde{z}),\mathrm{d}z',\mathrm{d}\tilde{z}')\,\bar{P}^n((z',\tilde{z}'),A) \\ &= \iint_{\mathcal{D}_\theta} \mathfrak{Q}\left(\left(\theta,\tilde{\theta}\right),\mathrm{d}\theta',\mathrm{d}\tilde{\theta}'\right) \\ &\qquad \int_{\mathbb{R}^+} \bar{g}_{\theta'}(w') \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{w}, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{\tilde{w}}\right\}\mathrm{d}w' \int P^n((\theta',w'),A),\end{aligned}$$

where we have provided a lower bound by considering the event where the joint chain couples in the first step and then moves to the set A in n steps. Continuing we have

$$\begin{aligned}\bar{P}^{n+1}((z,\tilde{z}),A) &\geq \int_\Theta \min\left\{q(\theta,\theta'), q\left(\tilde{\theta},\theta'\right)\right\}\mathrm{d}\theta' \int_\Theta \frac{\min\left\{q(\theta,\theta'), q\left(\tilde{\theta},\theta'\right)\right\}}{\int_\Theta \min\left\{q(\theta,\theta'), q\left(\tilde{\theta},\theta'\right)\right\}\mathrm{d}\theta'} \\ &\qquad \int_{\mathbb{R}^+} \bar{g}_{\theta'}(w') \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{w}, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{\tilde{w}}\right\}\mathrm{d}w' \int P^n\left((\theta',w'),A^{(1)}\right)\mathrm{d}\theta' \\ &= \int_\Theta \min\left\{q(\theta,\theta'), q\left(\tilde{\theta},\theta'\right)\right\} \\ &\qquad \int_{\mathbb{R}^+} \bar{g}_{\theta'}(w') \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{w}, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{\tilde{w}}\right\}\mathrm{d}w' \int P^n\left((\theta',w'),A^{(1)}\right)\mathrm{d}\theta'.\end{aligned}$$

Therefore we have that

$$\begin{aligned}\sum_{n=0}^{\infty} 2^{-(n+1)}\bar{P}^{n+1}((z,\tilde{z}),A) &\geq \int_\Theta \min\left\{q(\theta,\theta'), q\left(\tilde{\theta},\theta'\right)\right\} \\ &\qquad \int_{\mathbb{R}^+} \bar{g}_{\theta'}(w') \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{w}, \frac{\pi(\theta')}{\pi(\theta)}\frac{w'}{\tilde{w}}\right\}\mathrm{d}w' \\ &\qquad \sum_{n=0}^{\infty} 2^{-(n+1)} \int P^n\left((\theta',w'),A^{(1)}\right)\mathrm{d}\theta'.\end{aligned}$$

By Assumption 20 and Roberts and Tweedie (Theorem 2.2 1996) it easily follows that the exact algorithm is $\pi-$irreducible and aperiodic. Since by assumption we have $\varrho_{\mathrm{PM}}(\theta,w) < 1$, we deduce from Andrieu and Roberts (Theorem 1 2009) that

the kernel P is irreducible, hence $\pi-$irreducible. This further implies that

$$\sum_{n=0}^{\infty} 2^{-(n+1)} \int P^n \left((\theta', w'), B \right) > 0,$$

for all (θ', w') and sets B such that $\pi(B) > 0$ by Meyn and Tweedie (Proposition 4.2.1 2009). Since by assumption $\pi\left(A^{(1)}\right) = \pi_{\mathcal{D}}(A) > 0$ the integrand above will be strictly positive on a set of non-vanishing Lebesgue measure whence $\bar{P}$ is $\pi_{\mathcal{D}}-$irreducible. Finally to establish aperiodicity first notice that by assumption and continuity of the measures defined by the densities $\bar{g}_\theta(\cdot)$, we have that $\pi(C) > 0$. Letting $\mathcal{D}_C := \{(z, \tilde{z}) \in \mathcal{D} : z \in C\}$ and following the steps proving Equation (16) we can establish that for some $\epsilon' > 0$

$$\inf_{z \in \mathcal{D}_{\mathcal{C}}} \bar{P}\left(z, \mathcal{D}\right) \geq \epsilon',$$

and since $\pi(\mathcal{D}_{\mathcal{C}}) > 0$ this proves the aperiodicity of $\bar{P}$.

4.6.5 Description of referendum survey data

We use data from the 13 wave internet survey study (as of June 2018) (Fieldhouse et al. 2018) comprising 68,625 respondents in total, with the number of respondents varying between waves. We first subset the data into those in the four annual waves 1, 4, 7 and 11 occurring between February and May of 2014-2017. Of these 7,729 answered either 'Stay/remain in the EU' or 'Leave the EU' to the question 'If you do vote in the referendum on Britain's membership of the European Union, how do you think you will vote?' in each wave. We filter out those that answered 'Don't know' to the question 'How do you think the general economic situation in this country has changed over the last 12 months?' reducing the sample by 5 respondents. Finally, we perform inference only on waves 1, 4, and 7, the waves prior to the EU referendum on 23^{rd} June 2016. For simplicity, we do not take into account respondent weighting.

5 Conclusion

We here provide a brief overview of some of the main contributions of the thesis, outline additional research avenues and suggest any peripheral findings.

5.1 Bayesian inference of structural breaks in factor copula models

Methodology has been developed working at the intersection of modelling and inference to obtain an efficient means by which to detect structural changes in high-dimensional data. A novel particle Gibbs inference routine was developed to infer latent changes in the data. The routine exploits bespoke inference methods for Bayesian inference in multiple changepoint problems. The resulting algorithm performs well on both high dimensional data, in particular owing to an idiosyncrasy of the model we are able to exploit computationally a lack of identifiability of the model, while still performing inference on the identified one.

5.2 Unbiased Smoothing using Particle Independent Metropolis–Hastings

Motivated by the vast computational gains afforded by parallel computation for unbiased Monte Carlo estimators, we have successfully introduced a means by

which to obtain unbiased estimates for generic sequential Monte Carlo samplers. We have demonstrated the methodology on complex, intractable models arising in mathematical biology and econometrics. We have established validity of both a naive (but algorithmically intuitive) estimator and an estimator obtained when re-using all the particles. In particular, such conditions are verifiable in practise. Furthermore, owing to a precise initialisation of the algorithm we are also able to characterise its running time precisely, which is itself random variable. We have investigated tuning the number of particles for the naive estimator based on this expression. It would be instructive to apply similar arguments to help tune the more advanced estimator. This is left for further research.

5.3 Unbiased Markov chain Monte Carlo for intractable target distributions

We have applied similar ideas to de-biasing estimators obtained using MCMC to particle MCMC methods. This enables the vast parallelisation of otherwise expensive routines. Validity and efficiency of these techniques relies on establishing polynomial ergodic mixing properties of the underlying chain. Such a requirement has involved extending previous results of polynomial mixing of the Markov chain, implied by a Lyapunov drift condition. The resulting algorithm has successfully been applied to a toy model and a model from neuroscience. The toy model demonstrates desirable properties in so far as the meeting times can be made as close to the exact algorithm by increasing the number of particles. It also demonstrates - consistent with other related asymptotics in the literature - that scaling the number of particles with the number of observations ensures that the meeting times are approximately stable with the number of observations. The example from neuroscience illustrates both the dependency of the meeting time on the noise introduced through estimating the likelihood using particle methods as well as how the distribution of meeting times varies with the size of the proposals. Further extensions of the algorithm were examined, including block pseudo-marginal where the methodology was successfully applied to polling data. Finally, a variant

of randomised MCMC was considered, known as the exchange algorithm, whereby the methodology was applied on a simple Ising model.

5.4 A look to the future

Methodology developed to obtain unbiased estimates of stationary distributions of MCMC kernels has proven instrumental in developing new algorithms capable of leveraging the full scale of parallel computation. There are close parallels with exact sampling provided by Coupling From the Past (CFTP) Propp and Wilson (1996), providing exact (not just unbiased) samples of Markov chain stationary distributions. However, the latter (in its canonical form) relies extensively on stochastically monotone chains, i.e. those chains for which there exists a coupling under which a partial ordering may be preserved almost surely. Additionally, in the absence of a dominating process Kendall and Møller (2000), CFTP relies also on the existence of extremal elements under the same partial ordering, which at best may be hard to compute (requiring the solution of a further computational problem in finding these extrema) and at worst may not even exist for kernels on arbitrary spaces. In contrast, unbiased estimation considerably relaxes these requirements, demanding only that the chains meet almost surely irrespective of any ordering relation imposed on the underlying space. Experimental evidence suggests this is aided by chains that tend to mix more quickly, although there may be examples of chains that mix quickly but for which no straightforward coupling exists. Further formulation and understanding of these cases could help to bring about improvements to existing algorithms, provide new analytical insights and potentially elucidate any limitations of existing algorithms. Finally, it would be exciting and instructive to more precisely formalise any deeper connections between perfect sampling and unbiased estimation, especially if they can be seen as particular examples of a broader algorithmic framework in which they squarely fit.

Bibliography

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *The Annals of Applied Probability*, 25(2):1030–1077, 2015.

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

C. Andrieu, G. Fort, and M. Vihola. Quantitative convergence rates for subgeometric Markov chains. *Journal of Applied Probability*, 52(2):391–404, 2015.

C. Andrieu, A. Doucet, S. Yıldırım, and N. Chopin. On the utility of Metropolis-Hastings with asymmetric acceptance ratio. *arXiv preprint arXiv:1803.09527*, 2018a.

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

J. Bai and P. Wang. Econometric analysis of large factor models. *Annual Review of Economics*, 8:53–80, 2016.

O. E. Barndorff-Nielsen and N. Shephard. Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

O. E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280, 2002.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

M. Beaumont. Estimation of population growth of decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

J. Bérard, P. Del Moral, and A. Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *Electronic Journal of Probability*, 19, 2014.

J. Breitung and S. Eickmeier. Testing for structural breaks in dynamic factor models. *Journal of Econometrics*, 163(1):71–84, 2011.

A. E. Brockwell and J. B. Kadane. Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14(2):436–458, 2005.

J. Carpenter, P. Clifford, and P. Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7, 1999.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

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

N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, 2013.

J. N. Corcoran and R. L. Tweedie. Perfect sampling from independent Metropolis-Hastings chains. *Journal of Statistical Planning and Inference*, 104(2):297–314, 2002.

D. D. Creal and R. S. Tsay. High dimensional dynamic stochastic copula models. *Journal of Econometrics*, 189(2):335–345, 2015.

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

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

P. Del Moral, A. Doucet, and G. W. Peters. Sharp propagation of chaos estimates for Feynman–Kac particle models. *Theory of Probability & Its Applications*, 51 (3):459–485, 2007.

P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278, 2012.

G. Deligiannidis, A. Doucet, and M. K. Pitt. The correlated pseudo-marginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) to appear. arXiv preprint arXiv:1511.04992*, 2015.

R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14(3): 1353–1377, 2004.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

C. Du, C.-L. M. Kao, and S. Kou. Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111(513):314–330, 2016.

P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 589–605, 2007.

E. Fieldhouse, J. Green, G. Evans, H. Schmitt, C. V. D. Eijk, J. Mellon, and C. Prosser. British Election Study Internet Panel Waves 1-13, 2018.

M. Gerber and N. Chopin. Sequential quasi Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579, 2015.

R. Gerlach, C. Carter, and R. Kohn. Diagnostics for time series analysis. *Journal of Time Series Analysis*, 20(3):309–330, 1999.

A. Ghalanos. Introduction to the rugarch package.(version 1.3-8). 2017.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

P. W. Glynn and P. Heidelberger. Bias properties of budget constrained simulations. *Operations Research*, 38(5):801–814, 1990.

P. W. Glynn and P. Heidelberger. Analysis of parallel replicated simulations under a completion time constraint. *ACM Transactions on Modeling and Computer Simulations*, 1(1):3–23, 1991.

P. W. Glynn and C.-h. Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

P. W. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520, 1992.

S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for nonlinear time series. *Journal of the american statistical association*, 99(465), 2004.

A. Golightly and T. Kypraios. Efficient SMC2 schemes for stochastic kinetic models. *Statistics and Computing*, 28(6):1215–1230, 2018.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

J. Heng and P. E. Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *arXiv preprint arXiv:1709.00404*, 2017.

J. Heng, A. N. Bishop, G. Deligiannidis, and A. Doucet. Controlled sequential Monte Carlo. *arXiv preprint arXiv:1708.08396*, 2017.

P. E. Jacob, L. M. Murray, and S. Rubenthaler. Path storage in the particle filter. *Statistics and Computing*, 25(2):487–496, 2015.

P. E. Jacob, J. O'Leary, and Y. F. Atchadé. Unbiased Markov chain Monte Carlo with couplings. *arXiv preprint arXiv:1708.03625*, 2017.

P. E. Jacob, F. Lindsten, and T. B. Schön. Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, (just-accepted): 1–32, 2018.

S. F. Jarner and G. O. Roberts. Polynomial convergence rates of Markov chains. *Annals of Applied Probability*, pages 224–247, 2002.

A. M. Johansen, N. Whiteley, and A. Doucet. Exact approximation of Rao-Blackwellised particle filters. *IFAC Proceedings Volumes*, 45(16):488–493, 2012.

V. E. Johnson. Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91 (433):154–166, 1996.

V. E. Johnson. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248, 1998.

N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30 (3):328–351, 2015.

W. S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in applied probability*, 32(3):844–865, 2000.

P. Kingsley. Financial crisis: timeline. *The Guardian*, Aug 2012. URL `https://www.theguardian.com/business/2012/aug/07/credit-crunch-boom-bust-timeline`.

A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.

A. Lee, S. S. Singh, and M. Vihola. Coupled conditional backward sampling particle filter. *arXiv preprint arXiv:1806.05852*, 2018.

A. Lee, L. Murray, and A. M. Johansen. Resampling in conditional SMC algorithms. *Technical Report*, 2019.

L. Lin, K. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Physical Review D*, 61:074505, 2000.

F. Lindsten, M. I. Jordan, and T. B. Schön. Particle gibbs with ancestor sampling. *The Journal of Machine Learning Research*, 15(1):2145–2184, 2014.

H. Manner and O. Reznikova. A survey on time-varying copulas: Specification, simulations, and application. *Econometric Reviews*, 31(6):654–687, 2012.

H. Manner, F. Stark, and D. Wied. Testing for structural breaks in factor copula models. *Journal of Econometrics*, 2018.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased Markov chain Monte Carlo for intractable target distributions. *arXiv preprint arXiv:1807.08691*, 2018.

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased smoothing using particle independent Metropolis–Hastings. *arXiv preprint arXiv:1902.01781*, 2019.

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006.

J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

P. Mykland, L. Tierney, and B. Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241, 1995.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2): 125–139, 2001.

R. M. Neal. Circularly-coupled Markov chain sampling. *arXiv preprint arXiv:1711.04399*, 2017.

G. K. Nicholls, C. Fox, and A. M. Watt. Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*, 2012.

D. H. Oh and A. J. Patton. Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, 35(1):139–154, 2017.

L. Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117, 1944.

J. Park and M. Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, (just-accepted), 2018.

A. J. Patton. Modelling asymmetric exchange rate dependence. *International economic review*, 47(2):527–556, 2006.

M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

M. Plummer, N. Best, K. Cowles, and K. Vines. Coda: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006.

J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random structures and Algorithms*, 9(1-2): 223–252, 1996.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1): 95–110, 1996.

J. S. Rosenthal. Parallel computing and Monte Carlo algorithms. *Far East journal of Theoretical Statistics*, 4(2):207–236, 2000.

S. M. Schmon, G. Deligiannidis, A. Doucet, and M. K. Pitt. Large sample asymptotics of the pseudo-marginal method. *arXiv preprint arXiv:1806.10060*, 2018.

S. Temereanca, E. N. Brown, and D. J. Simons. Rapid changes in thalamic firing synchrony during repetitive whisker stimulation. *Journal of Neuroscience*, 28(44): 11153–11164, 2008.

H. Thorisson. *Coupling, Stationarity, and Regeneration*. Springer: New York, 2000.

K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

M.-N. Tran, R. Kohn, M. Quiroz, and M. Villani. Block-wise pseudo-marginal Metropolis-Hastings. *arXiv preprint arXiv:1603.02485*, 2016.

M. Ullrich. Exact sampling for the Ising model at all temperatures. In *Monte Carlo Methods and Applications: Proceedings of the 8th IMACS Seminar on Monte Carlo Methods, August 29–September 2, 2011, Borovets, Bulgaria*, volume 223. Walter de Gruyter, 2013.

N. Whiteley, C. Andrieu, and A. Doucet. Bayesian computational methods for inference in multiple change-points models, 2011.

X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine learning*, pages 1055–1062. ACM, 2007.

H. Yu, C. Li, and J. Dauwels. Network inference and change point detection for piecewise-stationary time series. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE, 2014.

Y. Zhang, N. Malem-Shinitski, S. A. Allsop, K. M. Tye, and D. Ba. Estimating a separably Markov random field from binary observations. *Neural Computation*, 30(4):1046–1079, 2018.

Y. Zhou, A. M. Johansen, and J. A. Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.